Mid-board optics as an essential building block for future data center switches

Gonzalo Guelbenzu[1], Nicola Calabretta[1], Oded Raz[1]

[1] COBRA Research Institute, Eindhoven University of Technology, Eindhoven, The Netherlands

Data center switches with front-panel optical transceivers are limited by the front-panel size while switching ASICs continue scaling in bandwidth. This work investigates the impact of mid-board optical transceivers on the design of data center switches. We conclude that mid-board optics overcomes the front-panel bottleneck of data center switches, and reduces the size of the switches. The increased port and bandwidth density, and the reduced power consumption, enable further up scaling of data centers requiring more computational power and communication bandwidth. As a proof of concept, a 1.28 Tbps data center switch in a 200mm x 200mm board is implemented.

Introduction

Data center switch designers keep up scaling the number of ports and bandwidth of their devices to satisfy the increasing demand of data center services [1]. This is needed to interconnect the growing number of servers in mega data centers, in the order of hundreds of thousands and beyond. Currently, state of the art switches provide up to 128-ports at 25Gbps [2], and require one rack unit (RU) of the rack height. However, switch designers face important challenges, like the packaging bottleneck and the front-panel bottleneck [3], which are limiting further up scaling of the number of ports and the bandwidth density. The packaging bottleneck means that devices use a high portion of their pins for power, because of the large amount of consumed power, and that limits the number of ports that can be provided. The front-panel bottleneck means that the front-panel of the switch is completely filled with optical transceivers.

There are several approaches to solve these problems. One of them keeps using front-panel transceivers. With this approach, the only option to continue up scaling the rack space and bandwidth density is by the development of more compact front-panel transceivers. For instance, technology has evolved from SFP+ (10G) transceivers, to QFSP (40G) transceivers, and now to CDFP (400G) transceivers. However, this approach eventually leads again to the mentioned front-panel and packaging bottlenecks. The relatively long traces interconnecting ASIC and transceivers cause high channel requirements for those devices, resulting in more complex, bigger, and more power hungry devices. Another approach looks for more integrated electronic-optical devices: e.g. by stacking the optics on top of the electronics [4], or by fully integrated electronic-optical devices [5]. This approach has clear advantages, like great size and power reduction, but the technology is not mature yet.

This work investigates the use of mid-board optics (MBO) transceivers for data center switches [6]. We implement a 128-ports 1.28 Tbps data center switch in a 200mm x 200mm board. We find that MBO is the natural transition technology from current front-panel transceivers to more integrated electronic-optical devices. It avoids the front-panel bottleneck, improves ports and bandwidth scaling of the rack space, and may help to solve the packaging bottleneck. For instance, up to six of those switches could be packaged into one OpenU case, or 180 switches in one OpenRack [7].

The benefits of using Mid-board optics in data center switches

Figure 1 (left) shows the layout of our proof-of-concept design: a 128-ports 1.28Tbps data center switch with mid-board optics [8,9]. The MBO transceivers are placed now on the board instead of in the front-panel, surrounding the switch ASIC. This approach gives several advantages. First, the use of MBO modules, which are smaller than front-panel transceivers and are placed closer to the ASIC, allows reducing the size of the system. Our proof-of-concept design, with a size of 200mm x 200mm, is roughly four to six times smaller than traditional data center switches with similar performance. Therefore bandwidth and port density in a rack level is improved. Second, this approach avoids the front-panel bottleneck. Now, the front-panel is free from optical transceivers, and more (smaller) MPO connectors can be placed in the front-panel to provide access to the ports. Using the Open Rack standard, six of those switches can be placed in one OpenU, and 66 MPO connectors can provide access to the ports of the 6 1.28Tbps switches inside the Open Rack case. Assuming 30 OpenU filled with switches in one rack, a total of 180 of those switches could be integrated in a single Open Rack, as shown in Figure 1 (right). Third, the electronic requirements of the channel connecting ASIC and transceivers decrease because shorter high-speed traces are required to interconnect them. This may lead to substantial power savings, since the devices may be adjusted or better dimensioned to lower channel requirements. In that case, the packaging bottleneck may be solved, and some pins of the devices can be changed from power to useful ports for data.

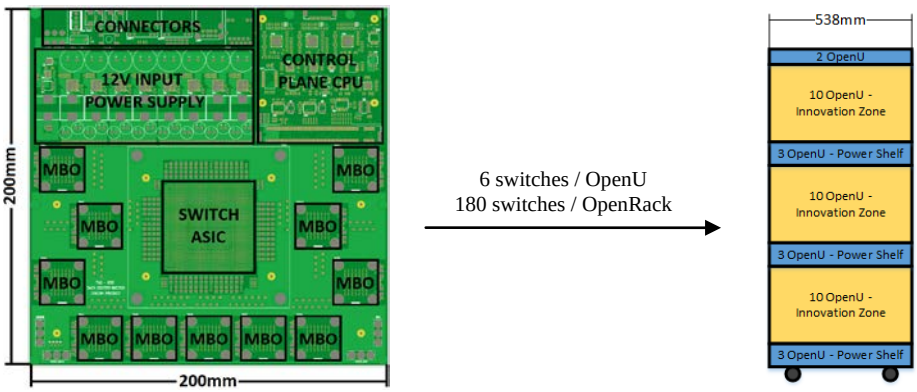

Fig. 1 (left): Component layout of implemented data center switch with mid-board optics; (right): Open-Rack

Our proof-of-concept design leaves space for further scaling of the rack space. The bandwidth can be up scaled by using, for instance, the 25G version of ASIC and transceivers [2,10]. It could be also improved in size, but this would require a multi-board design and a new generation of mid-board optics transceivers with increased port density. In that case, it should be relatively straightforward to place up to 12 switches in one RU box of an Open Rack.

Mid-board optics roadmap

Currently, data center switches follow the front-panel transceiver approach. The main advantage of this method is that transceivers are easily replaceable, without removing the switch from the rack. However, as already mentioned, this approach leads to the front-panel and packaging bottlenecks. Moreover, this path does not help in the transition to more integrated electronic-optical devices.

In contrast, the mid-board optics approach avoids the front-panel bottleneck, enables further scaling of current devices, and helps in the transition to top-of-the-package interconnects. Figure 2 shows the possible evolution of the mid-board optics solution. Figure 2 (left) shows the hypothetical layout of a multi-board data center switch integrating next generation of mid-board optics transceivers with increased port-density. Thanks to this increased port density, the mid-board optics transceivers can be placed closer to the ASIC, and the size of the system can be reduced even further. This would increase the bandwidth and port density of the rack space. It would also allow better adjustment or dimensioning of the devices to lower channel requirements, which may lead to a solution to the packaging bottleneck. Furthermore, it will help in the transition to top-of-the-package interconnects. When switches with top-of-the-package interconnects become available, it would be possible to assemble the developed technology for Figure 2 (left) directly on top of the switch ASIC, as shown on Figure 2 (right).

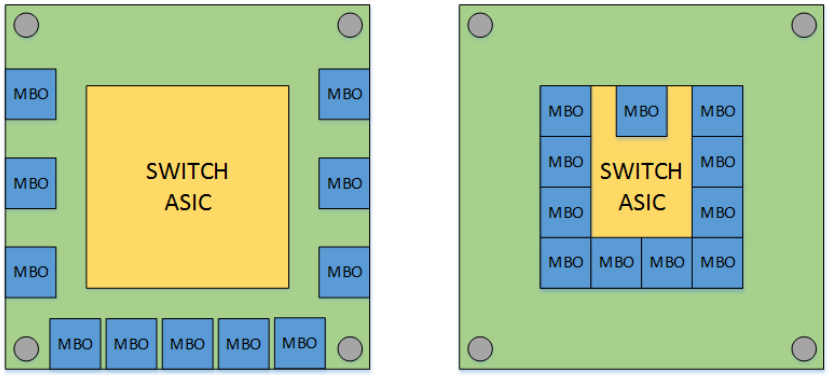

Fig. 2: (left): Data center switch with next generation mid-board optics transceivers;
(right): Data center switch with next generation mid-board optics transceivers, and top-of-the-package interconnects

Discussion

The use of mid-board optics for data center switches brings another advantage: a single kind of switch can be used for the whole data center. The front-panel transceivers approach lacks this feature, leading to different kind of switches in the market: leaf/edge switches with a high number of thin ports (e.g. SFP+), and spine/core switches with a low number of fat ports (e.g. QFSP). In contrast, by using mid-board optics transceivers, and a smart use of the flexibility provided by the ASIC manufacturers to distribute the bandwidth available in their devices, it is possible to configure the exact amount of ports and bandwidth per port of the switches. Therefore, a single kind of switch can be used for the entire data center. The same switch can be configured and modified by software, for instance, to have 128-ports at 10G, 32-ports at 40G, or a combination of them. This enables to take full advantage of the economy of scale, and also, it provides the flexibility to implement and evolve the data center.

Another important remark is that current ASICs and transceivers provide relatively limited flexibility to configure the input/output power of their ports. This limits the configuration capabilities of each system to adapt the power consumption of the integrating devices to the channel requirements of each system. In order to take full benefit of the mid-board optics approach, and achieve maximum savings in power, a more flexible configurability would be desirable in those devices.

Conclusions

This work explores the use of mid-board optics transceivers in data center switches, because this approach naturally avoids the front-panel bottleneck. The results show that mid-board optics enables significant size reduction of the switches, enabling further up scaling of data center racks in ports and bandwidth density. For instance, six units of our proof-of-concept design with a 200mm x 200mm size can be packaged in a single RU, or 180 units in an OpenRack. This technology may enable also to solve the packaging bottleneck, since shorter PCB traces lead to lower channel requirements. However, more flexible configuration capabilities of the input/output power per port would be required to achieve maximum power savings. Further up scaling is also possible by developing a new generation of mid-board optics transceivers with increased port density. This may also help in the transition to top-of-the-package interconnects.

Acknowledgments

This work is supported by the FP7-COSIGN project (No. 619572)

References

[1] M. Al-fare, et al, “A Scalable , Commodity Data Center Network Architecture,” pp. 63–74.

[2] Broadcom, “High-Density 25/100 Gigabit Ethernet StrataXGS Tomahawk Ethernet Switch Series.” [Online]. Available: https://www.broadcom.com/products/Switching/Data-Center/BCM56960-Series.

[3] A. Ghiasi, “Large data centers interconnect bottlenecks,” Opt. Express, vol. 23, no. 3, p. 2085, 2015.

[4] K. Hasharoni, et al, “A High End Routing Platform for Core and Edge Applications Based on Chip To Chip Optical Interconnect,” in Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2013, 2013, p. OTu3H.2.

[5] B. G. Lee, et al, “Monolithic Silicon Integration of Scaled Photonic Switch Fabrics, CMOS Logic, and Device Driver Circuits,” J. Light. Technol., vol. 32, no. 4, pp. 743–751, 2014.

[6] “COBO - Consortium for On-Board Optics.” [Online]. Available: http://cobo.azurewebsites.net/.

[7] “Open Rack.” [Online]. Available: http://www.opencompute.org/projects/open-rack/.

[8] Broadcom, “Broadcom BCM56850 StrataXGS Trident II Switching Technology.” [Online]. Available: http://www.broadcom.nl/collateral/pb/56850-PB03-R.pdf.

[9] Finisar, “10G Board-mount Optical Assembly.” [Online]. Available: https://www.finisar.com/optical-engines/fbotd10sl1c00.

[10] Finisar, “25G Board-mount Optical Assembly.” [Online]. Available: https://www.finisar.com/optical-engines/fbotd25fl2c00.