Co-packaged Optics for Data Center Switching

Rob Stone[1,2], Ruby Chen[1], Jeff Rahn[1], Srinivas Venkataraman[1], Xu Wang[1],
Katharine Schmidtke[1], James Stewart[1]

[1] Facebook Inc., 1 Hacker Way, Menlo Park 94025, USA. [2] robstone@fb.com

Abstract *As the bandwidth of data center switches increases, a disproportionate amount of power is becoming dedicated to the switch – optics interface. Reducing the physical separation between these two components by co-packaging enables system power savings which is essential to continued bandwidth scaling.*

Data Center Scaling Challenge

As data center (DC) network traffic demands continue to grow, operators are facing challenges ensuring the network capacity can be scaled to support the required load. In recent years, new workloads have been introduced, most notably machine learning and artificial intelligence which require large volumes of data movement[1]. In the Facebook data center network for example, machine to machine communication now represents not only the majority of the DC traffic, but also has the highest traffic growth rate[2].

DC infrastructure can be categorized as either fixed (such as buildings, installed fiber, power delivery and cooling), or replaceable (such as network hardware). The desire is to reuse the fixed infrastructure over several generations of network hardware to both amortize the infrastructure cost and minimize disruption and down-time during network upgrades. Subsequent generations of network hardware therefore have to be compatible with the constraints of the existing fixed infrastructure; size, power, and fiber plant.

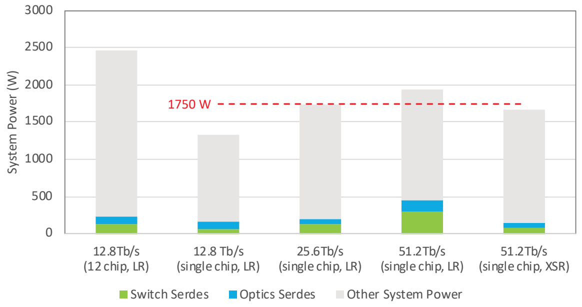

Fig. 1: Switch System Power vs Generation

Fig. 1 shows the power dissipation of five generations of Facebook optically connected network switch hardware from 12.8 Tb/s to 51.2 Tb/s total bandwidth. Although the switch power was successfully reduced from the 12.8T Backpack design to single-chip Minipack (by leveraging larger radix switch silicon)[3], the switch power is forecast to increase for subsequent generations as the total switch bandwidth is increased. The power increase is the result of many factors; increased lane speeds necessitating more complex serdes technologies and higher switch silicon bandwidths (with the associated greater amount of logic and higher clock speeds). Such generational increases in complexity are partially offset by transitioning to a more advanced CMOS node, however recent process driven power scaling has only resulted in a best case ~30% improvement, which only partially offsets the required increase in logic generation over generation of switch chip.

If the power scaling trend continues, undesirable fixed infrastructure upgrades will be required to support the higher network power demand. One architectural rationalization which addresses both the power, and overall system density scaling challenge is to reduce the power consumed in the optical module to switch interface. As shown in Fig. 1, for a conventional 51.2 Tb/s switch system based on front panel pluggable modules (FPPs), approximately 400 W of the total power budget is projected to be allocated to the electrical interface connecting the switch to optics. By placing the optics in close proximity to the switch silicon the electrical channel response is improved, enabling use of less complex, lower power serdes architectures. This is anticipated to lower the total system power by up to 300 W, which brings the switch power under the 1750 W fixed infrastructure constraint.

CPO Architecture

Figure 2 shows an example of a co-packaged

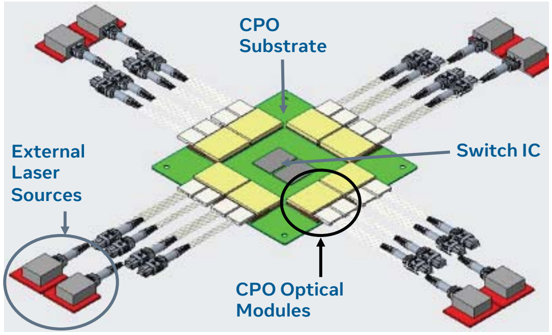

Fig. 2: CPO Switch Subassembly

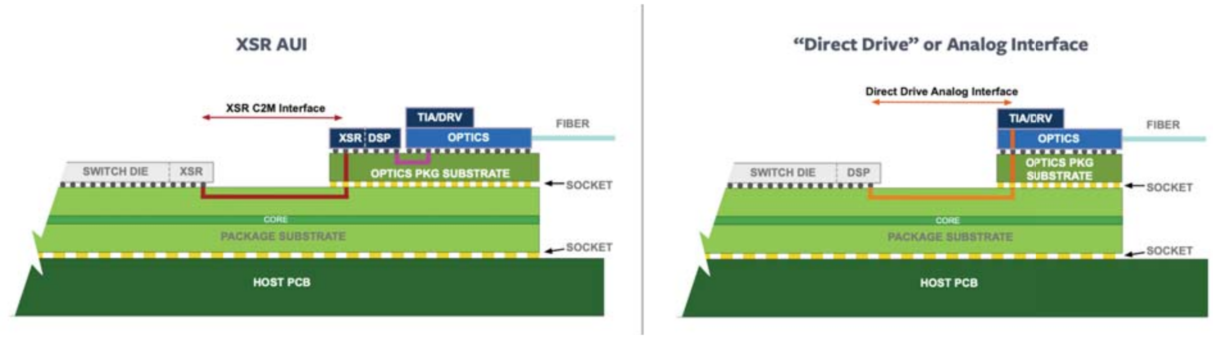

Fig. 3: CPO Architecture Options

optics (CPO) sub-assembly. Fiber-pigtailed CPO transceiver modules comprising the transmitter and receiver arrays are shown socket mounted around the perimeter of the switch package substrate. Unmodulated external laser sources (ELS) may be optionally used as the light input to the CPO modules. The decision to use an ELS as opposed to an on-chip laser source is dependent on multiple factors including anticipated operating temperature, forecast laser reliability, as well as whether the silicon photonics (SiPh) process technology includes provision for III/V material integration.

A cross sectional view of two possible switch to optics interface architecture options are shown in Figure 3. The left side of the figure shows an AUI style interface, as defined in the IEEE 802.3 standard[4], but supported by an XSR electrical interface[5] instead of the typical VSR. The optical link is driven from a DSP style serdes which is packaged within the CPO module. Such a retimed architecture follows current industry practices for 25, 50 and 100 Gb/s per lane switch to optics front panel pluggable (FPP) interfaces, with regard to functional partitioning and offers the benefits of robustness and well specified and understood interface definitions between the different sub-components, thus enabling interoperability. The right-hand side of the figure shows an alternate further rationalized approach. Here the composite electrical and optical links are directly driven "end to end" as an analog channel from a DSP located within the switch silicon. The benefits of this approach are reduced component count and lower power resulting from elimination of a dedicated switch to optics serdes interface. However, although offering an additional power saving over the retimed or "AUI" approach, the challenge for adoption of direct-drive centers around interoperability and robustness. This remains to be demonstrated at the candidate lane rates, although it is recognized that for 10 Gb/s per lane QSFP+ and SFP+ FPPs, the direct drive approach continues to enjoy broad commercial success.

Challenges for CPO

Migration to CPO from FPP based systems requires modifications to multiple aspects of both system manufacturing, as well as deployment and operation. For FPP based switch systems, hyperscale operators typically multi-source optical modules (enabled by industry standardization) directly from the optics manufacturers, installing them on-site during the system commissioning process. In contrast, for CPO systems, the optics is expected to be integrated off-site, either as part of the system manufacturing process, or as a CPO switch sub-assembly prior to system manufacturing. As indicated in Fig. 3, it is proposed that the CPO modules utilize an electrical socket to attach to the switch package. Use of a socket, rather than a solder attach process addresses two important considerations. Firstly, the CPO to switch integration process is high yield, and easily reworkable in case of a faulty or damaged component. Secondly, establishing a common electrical socket definition enables a path towards eventual multi-sourcing of the CPO modules, which due to supply considerations is a pre-requisite for wide adoption of CPO by hyperscale users.

Operationally the differences between CPO and FPP systems are somewhat self-evident. Whereas FPPs are inherently field serviceable and a defective module may be replaced without disturbing adjacent switch ports or removing the switch from the rack, servicing a CPO module requires removal of the entire switch assembly. As a result, the field failure rate of CPO modules will be required to be an order of magnitude lower than what is acceptable for FPPs to offer the same switch system level reliability. Although SiPh based transceivers using similar technologies to those being proposed for CPO have been deployed in data-center networks for over a decade with encouraging results[6], reliability at the high channel counts associated with CPO modules remains to be proven in the field.

Conclusion

Use of co-packaged optics for data center switch interconnects will enable system power savings and density improvements which are required to sustain future bandwidth growth within data center fixed infrastructure constraints. Volume CPO deployment will require changes to the operational model and supply-chain.

Acknowledgements

We wish to acknowledge contributions to this work from the Facebook Next Generation Optics Core Team, Hardware Engineering, and Data Center Network Engineering.

References

[1] V. Rao, *OCP Keynote 2019* https://www.youtube.com/watch?v=DFrCEvPgEcQ

[2] A. Andreyev, *Introducing data center fabric, the next-generation Facebook data center network*, , https://engineering.fb.com/production-engineering/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

[3] A. Andreyev, X. Wang, A. Eckert, *Reinventing Facebook's data center network*, https://engineering.fb.com/data-center-engineering/f16-minipack/

[4] See IEEE 802.3-2018 available at https://www.techstreet.com/ieee/standards/ieee-802-3-2018?product_id=1999889#full

[5] Standard in development. See https://www.oiforum.com/technical-work/current-work/#cei-xsr

[6] See for example Lightcounting "May 2019 Integrated Optical Devices" Market Report.