The Role of Standards for Cloud-Scale Data Centers

Mark Filer, Brad Booth, David Bragg
Microsoft Corp, Redmond, WA USA
mark.filer@microsoft.com

Abstract: Standards play an increasingly important role for cloud providers considering the dramatic growth that cloud services are experiencing. Distinctions are made between open consortia, multi-source agreements, and standards, and case studies with lessons learned are presented.
OCIS codes: (060.2330) Fiber optics communications; (060.4250) Networks

1. Introduction

Historically, optical component and system supplier revenues have been dominated by traditional telecom and enterprise customers and use cases. In the last several years, tremendous growth in cloud service provider (CSP) traffic (Fig. 1a) has tipped the scales toward a majority of optical supplier revenue generated instead by cloud-scale data center operators (Fig. 1b). As CSPs increasingly consume more optical products, and as the scale of data center operations continues to grow, it's becoming vital that such products be tailored to the specific needs of CSPs. In particular, CSPs have reduced requirement sets relative to telcos (e.g., relaxed specifications, shorter product life cycles, Ethernet-only), and they are much more cost sensitive due to the massive volumes of their operations.

Until recently, most of the optical networking products in the marketplace were conceived in various standards bodies and consortia by component and system suppliers, driven by telco operators to meet telecom-grade requirements. These products were used in early cloud applications until it was recognized that there might be a better way, and increasingly, that there was a market to support it. Rather than passively waiting for the market to develop solutions based on telco requirements, and sub-optimally applying them in cloud applications, CSPs began to take a more proactive role in defining these requirements in public forums (i.e., standards and other consortia), steering the direction of product developments to better suit the unique needs of cloud-scale data centers.

This paper explores the role which standards play for cloud data center operators, drawing primarily from Microsoft's experience of working through various standards and consortia to help deliver products at cloud scale. The distinctions between standards bodies and consortia (e.g., OIF, IEEE, COBO), multi-source agreements (MSA) and "open" forums (e.g., Open Compute Project) as they apply to CSPs are discussed. Case studies along with lessons learned in projects specific to Microsoft are presented.

2. Standard vs MSA vs Open, practically speaking…

As mentioned above, CSPs have different (and streamlined) requirements from traditional telcos and enterprises. Examples of this include relaxed specifications: reduced loss and distance requirements for optical modules, or more confined operating temperature ranges due to the tightly controlled data center environment. In addition, products are typically expected to have a shorter life cycle for data center applications, as it's common to perform "forklift upgrades" on rows of servers every 3 to 5 years for technology refreshes. Likewise, optical modules and transponders for CSP applications need only support Ethernet-framed data and can make do without the added complexities and overhead of OTN or other legacy protocol support. An example of this is the development of the CWDM4 [1] and PSM4 [2] MSAs for 100G applications. These came about because telco-grade IEEE 100GBASE-LR4 was overkill for typical data center campus applications with max distances of 2 km, since LR4 supports distances of 10 km, is too costly, and the primary CFP form-factor too large. The shorter-reach, IEEE 100GBASE-SR4 standards didn't fit the bill either due to reach limitations, so multi-source agreements were formed to generate specifications for an interoperable, reduced loss/distance module which fulfilled CSP requirements at much lower per-bit costs. But, why were MSAs chosen? And what are the practical differences between MSA, standards bodies, and open consortia to begin with?

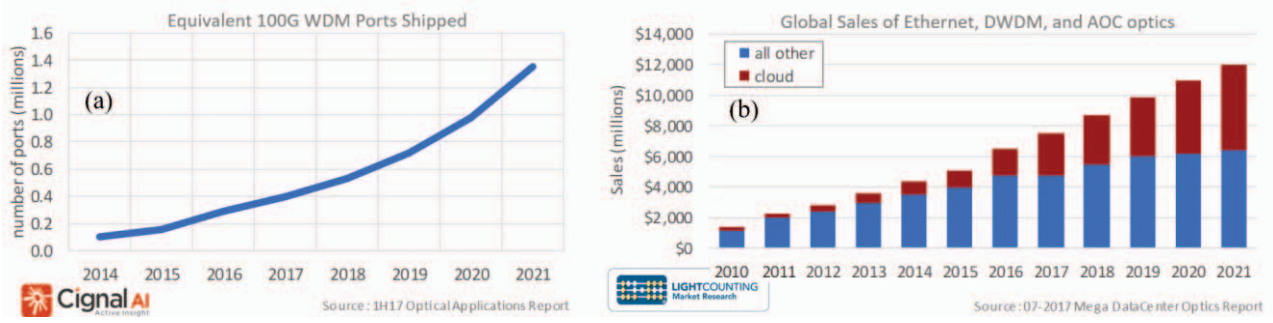

Figure 1. (a) global 100G-equivalent DWDM ports shipped (courtesy CignalAI), (b) global optical revenues (courtesy LightCounting)

Generally speaking, standards bodies, MSAs, and open forums are alike in having similar goals of delivering broadly-adoptable products to meet an identified market need. But there are some distinctions which make one or the other better suited for certain end goals. In standards bodies, designs are developed by a plurality of technical experts across multiple organizations, including component, system, and subsystem suppliers, and end users (i.e., operators). These efforts tend to be hardware- and protocol-focused. Consensus is achieved across the group of experts, and the design then becomes the de facto standard for anyone who wants to develop such a product. In principle this drives vendor-diverse ecosystems and benefits the consumers of the products by fostering competition (i.e., lower costs) and ensuring end products fully meet requirements. A drawback of this approach is it can involve longer lead times from conception until product availability, and the level of consensus may differ with each organization, which can impact the time needed to achieve consensus. Examples of this include the IEEE 802.3 working groups [3], the Optical Internetworking Forum (OIF) [4], and the Consortium for On-Board Optics (COBO) [5] where technical consensus is ≥ 75%, > 50%, or ≥ 66.67%, respectively.

On the other end of the spectrum are open forums, wherein the individual members may bring a full design forward for adoption. And it may not be the only design – others are free to bring their own as well. It's not an interoperable standard and specification; it's just an "open" design that anyone who wishes may adopt, and it's typically more software-oriented than hardware or protocol (examples are software data models and APIs). The expertise of the design is held within the member or company that brings that design forward as a contribution, and accordingly, it may benefit from short lead times from conception to product availability. Drawbacks of this approach are that the products and designs may never be adopted by anyone except the contributing party since they're not based on consensus across organizations. In addition, it places the burden of expertise on the constituent members and typically doesn't benefit from the diverse experience and viewpoints of the larger organization. Examples of open forums include the Open Compute Project (OCP) [6] and the Telecom Infra Project (TIP) [7].

Multi-source agreements typically fall somewhere in the middle as a specification development amongst agreeable parties. They are like standards bodies in that an agreeable set of technical experts come together to solve a problem and deliver a specification for an identified market need. However, this is generally a subset of the diverse groups of suppliers (and sometimes excludes end users) which would be present in a standards body. It allows the development of a solution to be "fast-tracked" relative to the time needed for extensive discussion and debate that takes place in standards, by restricting the members of the working group to a coalition of the willing. Development time can be shorter in MSAs because they can exclude those members that would vote against technical proposals which compete with their existing product lines. This has often resulted in specifications competing for the same market space, like X2 vs XPAK, 100G CWDM4 vs 100G CLR4, or QSFP-DD vs. OSFP.

In short, standards are "consensus was achieved on this specification," open is "build to my specification," and MSA is "my friends and I wrote this specification." In the case of the 100G PSM4 MSA, the industry's need for a middle-distance gray optic was identified, and none had been specified by the standards bodies. One may argue this was because CSP participation in standards at that time was low. In response, several optical component suppliers banded together to create an agreed-upon specification in an MSA to fast-track a solution for the data center market.

3. A Microsoft case study

In this section we present a case study from Microsoft demonstrating why CSP participation in standards and consortia benefits the CSPs themselves, and the marketplace as a whole. Back in 2013, in planning for the transition to the 100G ecosystem for data center networks, Microsoft recognized that to support a regionally distributed architecture (in which individual data center facilities may be up to 80 km away from the regional hubs), the current generation of coherent 100G solutions was too power consumptive, large, and expensive to support the port densities and scale necessary. To support the regional architecture, Microsoft recognized an 80 km capable DWDM technology that looked like a gray optic to the underlying switch fabric was needed. This meant that it would need to fit into the data center standard QSFP28 form-factor and consume less power than the form-factor's max of 5 Watts.

Microsoft began to engage the optical component suppliers on an individual basis to see how feasible such a solution seemed. Consensus was that it was "hard", suppliers didn't see additional demand for it, and it wouldn't be pursued until perhaps 2018 or 2019. However, one supplier, Inphi, suggested it may be possible to deliver such a solution in the desired time frame of 2016 to 2017, a full 2 years sooner than the rest of the supplier community. The solution would be based on two-wavelength 28 Gbaud PAM4, leveraging silicon photonics to meet the tight power consumption and form-factor requirements. Fast-forward a couple of years, and Inphi delivered on the technology promise in its ColorZ line of optical transceivers [8]. This transceiver technology has been deployed in the Microsoft network since 2017 and has enabled entirely new architectures at scales that were previously unavailable or unaffordable. However, in an ideal situation, this solution would be standardized and interoperable to create a robust supplier ecosystem and ensure that supply chain bottlenecks could not prevent the deployment regionally distributed data center fabrics. Fortunately, in the interim, coherent solutions optimized for the DCI market have been introduced which can stand in for the ColorZ-based deployments when absolutely needed. However, these come at a

cost of power, space, and CapEx that are between 3 to 5 times the 3-year total cost of ownership of the PAM4 solution, so they are avoided wherever possible.

Learning from the lessons of the past, Microsoft has been proactive in initiating standardized solutions for the 400G ecosystem. In the IEEE P802.3bs task force, Microsoft was a strong supporter for the development of a 400 Gb/s PSM solution for the intra-data center application. While it was possible to create another MSA, Microsoft's influence along with the 100G PSM4 MSA experience was sufficient for the task force to develop a specification for 400GBASE-DR4, as specified in IEEE Std. 802.3bs(tm)-2017. Early in the project, Microsoft spent time to educate those participating in the standards body on the requirements and cost models used for building data centers. The early participation along with the industry experience in the 100G PSM4 MSA was sufficient to help develop the 400GBASE-DR4 standard prior to market adoption.

In the OIF, Microsoft and a handful of others (including Google, Juniper Networks, and Acacia Communications) initiated a project, currently termed 400ZR, to standardize next-gen optical DCI solutions. The concept is in principle like that of the DWDM PAM4 technology described above but scaled up to support 400 Gb/s requirements. With 100G serial PAM4 already pushing the bandwidth limits of the electro-optics, it was quickly determined in the OIF that next-gen solutions for 400G would need to leverage coherent technologies. However, the drawbacks of full-blown coherent solutions (high power, complexity, and cost) dictate a new approach need be taken with the sub-120 km application space in mind. A dedicated, DCI-targeted coherent chipset with minimal features (and therefore power consumption) would need to be developed independently from the “swiss-army knife” coherent solutions currently available. Namely, a coherent DSP with limited: 1) dispersion and PMD compensation, 2) baud rate options, 3) flex-modulation support, and 4) FEC modes. Additionally, to make it viable as a switch-pluggable solution like today’s DWDM PAM4 modules, it was decided that the optics and electronics (including DSP chip) need to fit in a power envelope of 15 Watts and package sizes to achieve parity with today’s solutions (≥32 ports per 1RU). Finally, to ensure a robust supplier ecosystem, this technology must all be interoperable, including modulation, framing, pilot tones, and most importantly FEC, which is why the project is being carried out in the OIF. Work on the OIF 400GZR Implementation Agreement (IA) is underway at the time of writing [9], with several of the larger hurdles toward interoperability already passed, and good support from the supplier community in validating the assumptions and models around product feasibility.

An example of a standard development taking too long is IEEE Std. 802.3bs. Under the original project scope and timeline, a 400G 2 km objective was identified, and the task force adopted an eight-wavelength LAN-grid wavelength division multiplexing (LAN-WDM8) specification. This specification uses 25 Gbaud PAM4 per wavelength which, at the time was considered technically challenging, but also technically possible to develop. If the original project timeline had been followed, that decision may have seen broad market acceptance. The delay in the timeline and the desire by many CSPs (including Microsoft) to use a four-wavelength CWDM solution resulted in the formation of the 100G Lambda MSA [10] which was able to release its first specification within a month of the publication of IEEE Std. 802.3bs. In short, the MSA was able to build off existing the 400GBASE-DR4 specification to provide the industry a CWDM4 solution within the same time window.

There are also times when the industry needs to take a direction that is different than the normal operating mode. For Microsoft, that was the Consortium for On-Board Optics. At the time COBO was launched, the only module form factors that would support 400G was CDFP and CFP8. Neither was a great option due to their size and inability to provide sufficient port density for the 1RU 12.8T generation of switch equipment. While there was belief that the industry would provide a faceplate pluggable module that would meet these requirements – two were actually developed: QSFP-DD and OSFP – the prognosis for the next generation (beyond 400G) of faceplate pluggable didn't look promising. Microsoft and others felt it was necessary to prepare the industry for moving the optics closer to the host silicon to mitigate power and improve thermal performance; therefore, COBO was launched to permit the development of an industry specification for an embedded optical module form factor.

4. Conclusions

As the massive investment in cloud infrastructure continues, CSP participation in standards bodies and other consortia will only increase. This participation allows the CSP to streamline their technology investments by minimizing development times and targeting development resources to the cloud market. The scale of investment and the cost savings of targeted technologies make this efficiency necessary. The various types standards groups allow CSPs to tradeoff between breadth of market participation and development time. Expect to see more activity from the cloud providers in the industry associations in the coming years.

[1] http://www.cwdm4-msa.org/, retrieved 1/22/18.
[2] http://psm4.org/, retrieved 1/22/18.
[3] https://www.ieee.org/, retrieved 1/22/18.
[4] http://www.oiforum.com/, retrieved 1/22/18.
[5] http://onboardoptics.org/, retrieved 1/22/18.
[6] http://www.opencompute.org/, retrieved 1/22/18.
[7] https://telecominfraproject.com/, retrieved 1/22/18.
[8] https://www.inphi.com/products/colorz/,
[9] http://www.gazettabyte.com/home/2017/6/22/the-oifs-400zr-coherent-interface-starts-to-take-shape.html, retrieved 1/22/18.
[10] http://100glambda.com/, retrieved 1/22/18