Optical Fiber Technology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Optical Fiber Technology

journal homepage: www.elsevier.com/locate/yofte

Invited Papers

Scaling large data center interconnects: Challenges and solutions

Xiang Zhou*, Hong Liu, Ryohei Urata, Sara Zebian

Platforms, Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA, 94043, United States

A B S T R A C T

Increased demands for web and cloud-based services have been driving exponential growth of datacenter bandwidth. This paper discusses, from Google's perspective, emerging challenges and possible technical solutions for scaling intra-datacenter and intra-campus interconnection network bandwidth.

1. Introduction

Over the past decade, datacenters and their networks have become the technology enabler for a number of internet-based applications. As of today, most of the popular Internet applications, from traditional search, online interactive maps, and social networks, to video streaming and the Internet of things, are running in datacenters (DCs). The pivotal role played by the datacenter will be further heightened by wider adoption of cloud computing, in which a significant portion of compute and storage is migrated into shared DCs. This is already occurring at a rapid pace today with a number of large cloud providers leading the way. This has resulted in a dramatic increase in datacenter capabilities. As one example, the bisection bandwidth of Google's datacenter cluster networks has increased by a factor of one thousand over the past decade [1,2].

Fig. 1 provides a high-level view of Google's DC interconnection network. From distance and topology points of view, this network can be divided into four segments:

- the **intra-DC network** (i.e. the fabric cluster), which interconnects tens to hundreds of thousands of servers over a link distance from 500 m to 1 km;
- the **intra-campus DC interconnection network**, which interconnects clusters housed in different buildings but within an under-2 km campus neighborhood;
- the **point-to-point Metro edge access network**, which provides connections between our datacenters and our global backbone networks, with a link distance typically less than 80 km; and finally
- the **global backbone network**, interconnecting all of Google's DCs through long-distance transport technologies.

In this paper, we focus on first two segments of the DC interconnection network, i.e. the **intra-DC** and **intra-campus interconnects**, which are among the most cost- and power-sensitive parts of the whole network. The rest of this paper is organized as follows.

- In Section 2 we give a **brief review of critical requirements** for large-scale DCs.
- In Section 3 we present a **high-level view of the available technical design space** within which to scale interconnect bandwidth.
- Sections 4 and 5 are devoted to **emerging challenges and possible technical solutions** for scaling intra-DC and intra-campus interconnect bandwidth. Conclusions are also presented in Section 6.

2. Technology consideration criteria

For a typical intra-DC network adopting a Clos topology (see Fig. 1), a massive number of interconnection links are required to implement the large fan-out and corresponding high bisection bandwidth [1]. Thus, the primary consideration for intra-DC and intra-campus interconnection is the cost of bandwidth.

2.1. Cost of bandwidth

To minimize the total interconnection cost, different technologies are adopted at different interconnect reaches. For example, electrical interconnection technologies (PCB trace and copper cable) are typically used for switch I/O fan-out (chip to module) and intra-rack interconnection (with a reach typically less than a few meters), while fiber-based optical interconnects are used for interconnection between the top of rack (TOR) switch and the edge switch, as well as between the edge aggregation switch and the spine switch, with link distances ranging from a few meters up to 1 km. For link distances less than 100 m, i.e. the nominally SR (short reach), vertical cavity surface emitting laser (VCSEL) and multimode fiber (MMF) based technologies

* Corresponding author.
E-mail address: xiangzhou2009@gmail.com (X. Zhou).

http://dx.doi.org/10.1016/j.yofte.2017.10.002
Received 11 May 2017; Received in revised form 25 September 2017; Accepted 9 October 2017

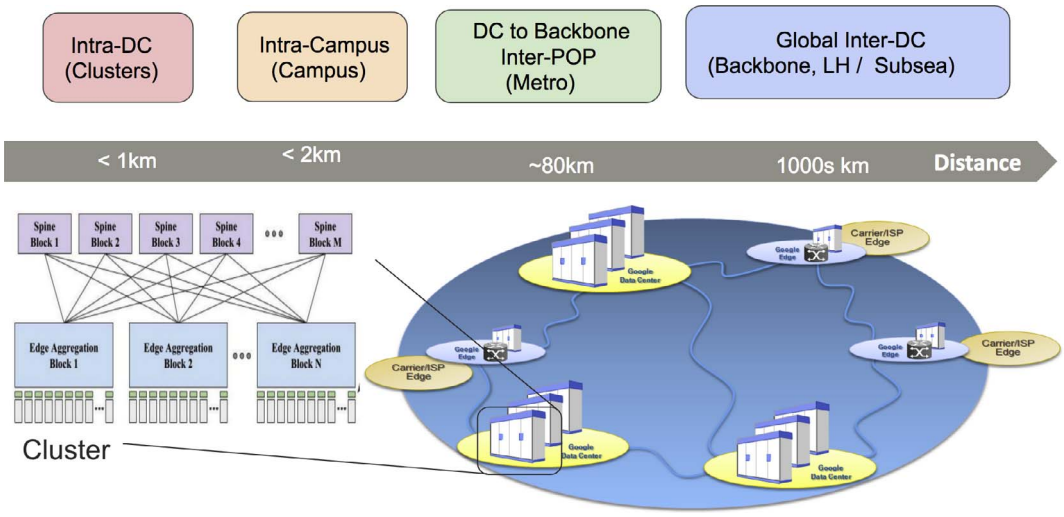

Fig. 1. A high-level view of Google's Datacenter (DC) Interconnection networks.

have proven to give the best overall link cost (transceiver cost plus fiber cost, at least up to 100 Gb/s interface rate). Beyond 100 m, i.e. the nominal LR (long reach), however, more expensive single mode fiber (SMF) transmission technologies usually must be used to achieve the required bandwidth reach.

2.2. Power consumption

The second important criterion is power consumption. From an aggregate energy consumption point of view, the power consumption of networking is only a modest portion of the total power consumed by a datacenter (on the order of less than 10% [2]). But the power efficiency of optical transceivers is essential for front panel density (the allowable transceiver size is largely determined by its power envelope). Hence, without power-efficient transceivers, there is no optimal way to take advantage of the full capacity of the switch Application-Specific Integrated Circuit (ASIC). For the SerDes used for chip-to-module interconnects, its power is limited by the allowable total-ASIC power dissipation, which is limited to around 300 W [3].

2.3. Serviceability

The third important criterion is serviceability. Since the reliability of typically-used active optical components is not very high, it is better to design the optical transceiver so it can be easily serviced or replaced. In this regard, pluggable optics is preferred over on-board optics, although on-board optics enables greater front panel density. An additional advantage of pluggable optics is that it allows us to optimize the cost of different reaches (e.g., copper for few meters, MMF for < 100 m, and SMF for > 100 m). Finally, cabling efficiency and transmission latency also need to be considered. Especially for intra-DC interconnects, low latency technologies can be critical for certain applications.

3. Technology design space

Fundamentally there are three degrees of design freedom to allow scaling of interconnect bandwidth as illustrated in Fig. 2:

- **increasing the symbol rate per lane** (i.e. the serial clock rate);
- **increasing the number of parallel lanes**, where the lanes can be in the space, polarization, or frequency domain; and
- **encoding more bits into each symbol** (i.e. higher-order modulation formats).

Each of these three orthogonal technology choices has advantages and constraints.

3.1. Symbol rate

Historically, scaling in the symbol rate axis is the most cost-effective method to increase the interface rate because it allows us to increase bandwidth while using the same amount of electrical and optical components. But the potential of this method is limited by achievable electrical and optical components' bandwidth.

3.2. Parallel lanes

Scaling bandwidth using the parallel channel axis is very effective in terms of increasing the aggregate data rate per interface, but the downside is that the required number of optical and electrical components increases linearly with the number of optical or electrical lanes. For parallel optics, the use of increased number of optical components, especially the active optical components, will impact the total yield and cost. The use of high-yield photonic integration technology (when mature) may help alleviate this problem. Tighter optoelectrical integration and/or packaging are also critical for reduction of the total power (otherwise the worst-case power increases linearly with the number of lanes). For reach beyond a few hundred meters, scaling in space (more fibers) is undesirable due to higher fiber cost and volume. For switch chip to module interconnects, the allowable number of total electrical lanes is also limited by the available chip package pins.

3.3. More bits per symbol

Finally, encoding more bits into each symbol allows us to scale the serial bit rate without imposing higher component BW requirements. However, such a lower BW requirement is achieved at the expense of the signal to noise ratio (SNR), as well as ISI (inter-symbol interference) and other channel impairments (such as various optical and electrical interferences).

For example, as compared to PAM2, PAM4 requires about 7 dB higher SNR than PAM2 (assume operations at the same baud rate with bipolar coding), and is about 9–10 dB less tolerant toward optical multipath interference [4]. To reduce the ISI penalty, more advanced digital equalization can be used. Higher coding gain FEC can also be utilized to compensate or partly compensate for the increased SNR requirement. However, effective management of optical interference can be much more challenging. To understand scaling challenges imposed by optical interferences, in Fig. 3 we show the impact of multi-path