Optical Network Technologies for Datacenter Networks (Invited Paper)

Cedric F. Lam
Google Inc., 1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA {clam@google.com}

Abstract: We review the optical communication technologies required to support data center operations and warehouse-scale computing.

OCIS codes:(060.4250) Networks, (200.4650) Optical interconnects

1. Introduction

For companies that offer ubiquitous access to user data and remote cloud computing applications through the Internet, fiber optic technology plays an important role in datacenter operations. In this invited paper, we give a generic review the optical networking technologies needed for datacenter operations.

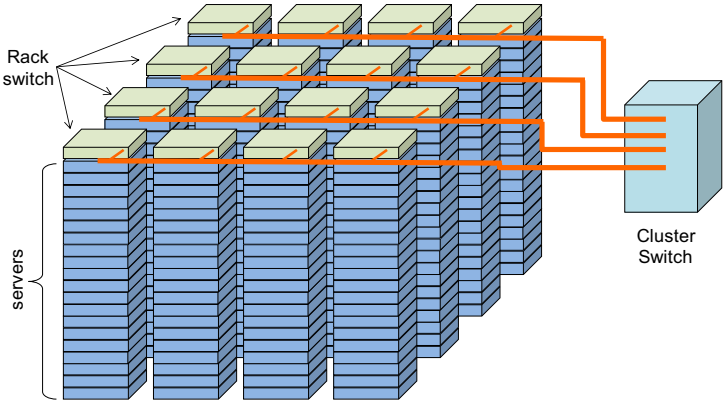

Fig. 1 Typical view of datacenter cluster

A typical datacenter [1] contains tens of thousands of identical servers arranged into one or more clusters as shown in Fig. 1. A cluster consists of multiple racks, each with 20-40 servers connected to a rack switch. Clusters are further networked through layers of cluster aggregation switches, which are then connected to datacenter routers. The datacenter routers connect to the rest of the Internet through Internet points of presence (POPs), where traffic flows from datacenter networks to end users and vice versa. Long haul WDM systems provide the connectivity between datacenters and population centers in different metro areas. Metro WDM transport links are used to interconnect datacenters and POPs located within the same metro area.

2. Optical Technologies for Intra-Datacenter Applications

Within a datacenter, optical technologies are used to interconnect vast number of identical servers and switches. Cloud computing takes the advantage of distributed computation power from massive parallel CPUs in warehouse scale-datacenters. In order to make efficient use of the distributed CPU farm, it is important to provide a flattened network with rich connectivity among all the servers so that application layer software can easily assign parallel jobs to arbitrary servers without concerning server location and communication overheads. Higher connectivity also leads to higher fault tolerance to individual link failures. Ideally, a fully meshed, completely non-blocking switch fabric with very low latency between any two servers is desirable. In reality, servers are interconnected through layers of cluster switches [2-6] comprised of massive identical switching equipment.

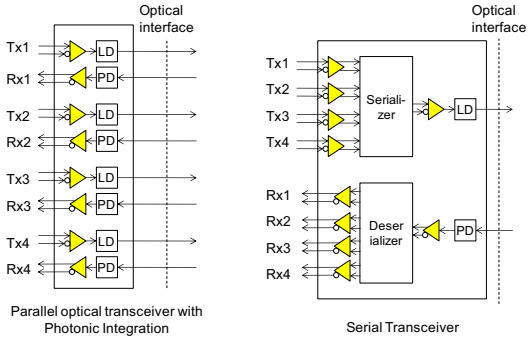

Fig. 2 Parallel vs. serial optical transceiver.

Intra-datacenter interconnects can be achieved in different topologies. Commonly seen interconnect architectures for cluster computing include torus, hypercube, fat-tree [4] and flattened butterfly [6]. Irrespective of what architecture is chosen, eventually, the port density on cluster switches will limit the number of servers that can be connected to a cluster.

Unlike in long-haul networks where fiber is scare and the figures of merit are spectral efficiency and bit-rate-distance product, the intra-datacenter environment is fiber-rich with interconnection distances spanning from a few meters to a few kilo-meters. Here the figures of merit are really power-space efficiency and bit-rate-port-count product [2]. Photonic integration of parallel transceivers using parallel ribbon fibers helps to improve transceiver

978-1-55752-884-1/10/$26.00 ©2010 IEEE

footprints and port count. However, smaller size does not directly translate to the higher port count needed by the fast-growing datacenters, unless the power footprint also scales down proportionally. According to [7], interconnect power consists more than 50% of modern CMOS chip power. Parallel optics takes the advantage of the natural parallel data lanes in a digital system to avoid power hungry serializers and deserializers (Fig. 2). Lower buad-rate also helps to improve the dispersion tolerance. As an example, a 4×10Gb/s QSFP module employing parallel multimode optics consumes roughly twice the power of a single-channel 10Gb/s SFP+ module with only slight increase of about 20% in space consumption.

In the next several years, as server output data rate increases beyond 10Gb/s, innovations will be needed to increase the speed of short-reach optical interconnects at least 4 fold to 40Gb/s, 120Gb/s and 480Gb/s, while keeping the same power-and-space profiles as those in today's SFP+, QSFP, and CXP modules. Potential technologies for achieving such goals include high-density photonic integrated circuits (PIC), integrated low-power optical vector modulators and optical demodulators for multi-level and/or phase-intensity modulations, electronic dispersion compensation and equalization techniques, etc.

3. Optical Networks for Inter-Datacenter Communications

Usually, datacenter facilities are located in remote rural areas near power stations to reduce the operation costs related to space and power. These mega datacenters are connected to the outside world and among themselves using ultra-long-haul WDM optical transport networks. Long-haul fiber resources are scarce, expensive and time consuming to acquire or construct. To maximize the utilization of precious long-haul transmission fibers, such networks need to provide high spectral efficiency.

Compared to traditional teleco networks which need frequent intermediate add/drops, mega datacenters are relatively sparsely located with large distances in between. Optical layer add/drop flexibility is not the most important concern. Most of the traffic between mega datacenters is machine generated with huge volume. Therefore, spectral efficiency, capacity, and ultra-long reach (6000km+) is important. Inter-datacenter long haul optical networks must minimize the number of regenerators in the field. Regenerators cause significant Op-Ex penalty. Sometimes, there simply may not be any space and power available between remotely located mega datacenters for regeneration purposes. The common requirement between long-haul carrier networks and inter-datacenter optical networks is smooth upgradability.

Coherent optical receivers [8] with electronic digital signal processing techniques make use of joint phase, amplitude and polarization encoding to achieve high spectral efficiency. Phase modulation also results in larger ASE noise tolerance and longer unregenerated reach. Such systems hold the promise for capacity growth in the next few years. State-of-the-art commercial coherent receivers use polarization multiplexed quaternary phase shift keying (PM-QPSK) modulation. Such systems can achieve 100Gb/s transmission on 50GHz ITU channel grid, and a total capacity of 8Tb/s in C-band. Other benefits of coherent receivers include high tolerance to PMD and CD. Practically all the PMD and CD effects can be corrected with linear digital electronic filters. High PMD tolerance allows carries to reuse bad-PMD field fibers without worrying about the stochastic signal fading due to PMD effects. The extremely high CD tolerance renders DCF (dispersion compensation fiber) unnecessary. The saving of DCF results in about 12 to 20% reduction in network transmission latency. In a 3000km link, this represents a 3ms one-way delay, which is significant for protocols requiring 3-way handshake such as TCP [10].

The spectral efficiency of coherent PM-QPSK modulation is 4 bits/symbol. Higher spectral efficiency requires more complex modulation schemes and much higher signal-to-noise ratio (SNR) according to Shannon's theory [9]. However, optical non-linearity effects limit the power that can be launched into the optical fiber, and the achievable SNR. Researches indicated that 400Gb/s might be the practical limit for single-carrier modulation implementations [11]. Going further, instead of keeping increasing signal constellation size on a single-carrier, multicarrier modulation methods such as OFDM appears to be more practical. A 1.2Tb/s OFDM transmission using 300GHz bandwidth was demonstrated in [12]. To ensure smooth adoption of these new modulation formats, transmission links should be designed free of bandwidth-limiting optical elements such as channelized ITU grid ROADMs[1]. In fact, Shannon's theory indicated that capacity grows linearly with respect to spectral width and only logarithmically with respect to SNR. Spectrum is therefore a very precious resource, and artificial spectral-limiting elements on transmission links, which will choke the long term capacity growth, should be avoided.

[1] Variable bandwidth ROADMs which does not introduce cascaded bandwidth narrowing are okay.

For the same reason that capacity is linear with respect to spectral width, to satisfy the growing datacenter interconnect bandwidth demands, it is also important to explore other spectral regions inside the optical fiber. Majority of the DWDM system manufactured today makes use of the C-band, which maximizes the SNR. The next logical band of choice for capacity expansion is L-band, which offers similar spectral capacity as C-band. EDFA can be tailored to work in the L-band and optical loss in L-band still remains very small. In addition, L-band wavelengths suffer from less nonlinearity effects in LEAF fibers, which have a very large embedded deployment base in the field.

Looking forward, other spectral regions such as S-band are also equally worth exploring as we quickly run out of C-band capacity in optical fiber [13], and long-haul fibers still remain scarce. Raman amplification is an important technology in exploring new optical spectra. Unlike EDFA which only operates in C and L bands, Raman amplifiers can be designed to operate at any wavelength. Besides, distributed Raman amplifiers help to improve unregenerated system reach by maintaining SNR and lowering the launch power requirement (hence lower nonlinear effect). On the other hand, as the optical spectrum used widens and the number of wavelengths increases, stimulated Raman scattering (SRS) effect which gives all the aforementioned benefits, can results in significant gain tilts [14]. Gain tilt management is therefore an important engineering issue to be considered in system designs in order to ensure the future upgradability of other transmission bands such as S and L bands.

From an infrastructure perspective, new fiber strands will help to alleviate the capacity crunch for datacenter networking and allow datacenter operators to avoid deploying bleeding-edge transmission technologies which often carry significant cost premium. To ensure maximum final capacity and transmission reach, new fiber deployments should focus on ultra-large-effective-area fibers which have lower loss and lower nonlinearity.

4. Conclusion

To conclude, fiber optics is the bloodstream for warehouse-scale computing. Remote mega datacenter operation requires both short-reach optical interconnects and long-haul optical networks. Short-reach intra-datacenter interconnects with photonic integration are important for building scalable datacenter clusters. To realize the advantage of photonic integration, optical transceiver power must scale down in synchronous with the space consumption. For long-haul transmission, because of the scarcity of long haul-fiber and operational cost consideration, maximizing system reach and fiber capacity is of utmost importance. Exploring new spectrum in the optical fiber and deploying more new fibers with ultra-large effective core areas are both important to cost-effectively satisfy the future bandwidth demands between remote mega datacenters.

Acknowledgements

I would like to thank my colleagues Bikash Koley, Valey Kamalov, Xiaoxue Zhao and my manager Vijay Gill for helpful discussions.

References

1. L.A. Barroso and U. Hölzle. *The Datacenter as a Computer – an Introduction to the Design of Warehouse-Scale Machines*, Morgan & Claypool Publishers, 2009. http://www.morganclaypool.com/doi/pdf/10.2200/S00193ED1V01Y200905CAC006
2. B, Koley, "Requirements for Data Center Interconnects," paper TuA2, 20th Annual Workshop on Interconnections within High Speed Digital Systems, Santa Fe, New Mexico, 3 – 6 May 2009.
3. C. E. Leiserson, IEEE Transactions on Computers, Vol 34, October 1985, pp 892-901
4. K. Hwang, "Advanced Computer Architecture: Parallelism, Scalability, Programmability," McGraw-Hill, New York, 1993
5. J. Kim et. Al, "Microarchitecture of a High-Radix Router," http://cva.stanford.edu/publications/2005/johnkim_hrr.pdf
6. http://www.sun.com/products/networking/datacenter/ds3456/ds3456_wp.pdf
7. N. Magen, et al, "Interconnect-power dissipation in a microprocessor," in Proc. 2004 Int. Workshop System Level Interconnect Prediction (ACM 2004), Session Interconnect Anal. SoCs Microprocess., pp 7-13.
8. J. Kahn, and E. Ip, "Principles of Digital Coherent Receivers for Optical Communications," paper OTuG5, OFC/NFOEC 2009
9. C.Shannon, Bell Systems Technical J., Vol.27, p.379 (1948)
10. W. R. Stevens, *TCP/IP illustrated, Volume 1*, Addison Wesley, 1994
11. T. Pfau et al., IEEE J. Lightwave Technology, Vol. 27, No. 8, April 2009, pp. 989-999
12. S. Chandrasekhar, et al,"Transmission of 1.2Tb/s 24-carrier no-guard-interval coherent OFDM superchannel over 7200-km of ultra-large-area fiber," ECOC 2009, PD2.6.
13. René-Jean Essiambre, et al, OFC/NFOEC 2009, Paper OThL1.
14. C. F. Lam, unpublished.