Optical Fiber Technology 17 (2011) 363–367

Contents lists available at ScienceDirect

Optical Fiber Technology

www.elsevier.com/locate/yofte

100GbE and beyond for warehouse scale computing interconnects

Bikash Koley *, Vijay Vusirikala

Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA

ARTICLE INFO

Article history:
Available online 28 July 2011

Keywords:
100 Gigabit
Ethernet
Warehouse scale computer
Internet
Datacenter
WDM
Cluster

ABSTRACT

Increasing broadband penetration in the last few years has resulted in a dramatic growth in innovative, bandwidth-intensive applications that have been embraced by the consumers. Coupled with this consumer trend is the migration from local compute/storage model to a cloud computing paradigm. As computation and storage continues to move from desktops to large internet services, computing platforms running such services are transforming into warehouse scale computers. 100 Gigabit Ethernet and beyond will be instrumental in scaling the interconnection within and between these ubiquitous warehouse scale computing infrastructures. In this paper, we describe the drivers for such interfaces and some methods of scaling Ethernet interfaces to speeds beyond 100GbE.

1. Introduction

As computation continues to move into the cloud, the computing platforms are no longer stand-alone servers but homogeneous interconnected computing infrastructures hosted in mega-datacenters. These warehouse scale computers (WSCs) provide a ubiquitous interconnected compute platform as a shared resource for many distributed services, and therefore are very different from traditional rack-full of collocated servers in a datacenter [1]. Interconnecting such WSCs in a cost-effective yet scalable way is a unique challenge that is being addressed through network design and technology transformation, which in turn is leading to the evolution of the modern internet. The central core of the internet, which was dominated by traditional backbone providers, is now connected by hyper giants offering rich content, hosting, and CDN (Content Distribution Network) services [2]. It is not difficult to imagine that the network is moving towards more and more direct connection from content providers to content consumers with the traditional core providers facing disintermediation.

Table 1 lists the ATLAS top-10 inter-domain Autonomous Systems (AS) in the public internet in 2007 and 2009. We see that content providers such as Google and Comcast, which were not ranked in 2007 now occupy prominent places in 2009. It should be noticed that this reports only accounts for publicly measureable bandwidth between AS's where the measurement was taken. Left uncounted here are three types of traffic: (1) traffic inside datacenters, (2) the backend bandwidths used to interconnect datacenters and operate the content distribution networks, and (3) Virtual Private Network (VPN) traffic. These data demonstrate the transformation from the original focus on network connectivity by traditional carriers to a focus on content by the non-traditional companies. New internet applications such as cloud computing and CDN are now reshaping the network landscape: Content providers and cloud computing operators such have now become the major driving forces behind large-capacity optical network deployments [3].

2. Intra-datacenter connectivity

A WSC is a massive computing infrastructure built with homogeneous hardware and system software arranged in racks and clusters interconnected by massive networking infrastructure [1]. Fig. 1 shows common architecture of a WSC. A set of commodity servers are arranged into racks and interconnected through a top of rack (TOR) switch. Rack switches are connected to cluster switches which provide connectivity between racks and form the cluster-fabrics for warehouse scale computing.

Ideally, one would like to have an intra-datacenter switching fabric with sufficient bi-sectional bandwidth to accommodate non-blocking connection from every server to every other server in a datacenter, so that applications do not require location awareness within a WSC infrastructure. However, such a design would be prohibitively expensive. More commonly, interconnections are aggregated with hierarchies of distributed switching fabrics with an oversubscription factor for communication across racks (Fig. 2) [4].

Intra-datacenter networking takes advantage of a fiber rich environment to drive very large bandwidth within and between clusters. However, fiber infrastructure itself is becoming a significant cost

* Corresponding author.
E-mail address: bkoley@google.com (B. Koley).

1068-5200/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.yofte.2011.06.008

Table 1
Atlas top-10 public internet bandwidth generating domains [1].

(a) Top-10 in 2007			(b) Top-10 in 2009		
Rank	Provider	Percentage	Rank	Provider	Percentage
1	Level(3)	5.77	1	Level(3)	9.41
2	Global Crossing	4.55	2	Global Crossing	5.7
3	ATT	3.35	3	Google	5.2
4	Sprint	3.2	4		
5	NTT	2.6	5		
6	Cogent	2.77	6	Comcast	3.12
7	Verizon	2.24	7		
8	TeliaSonera	1.82	8		
9	Savvis	1.35	9	*Intentionally omitted*	
10	AboveNet	1.23	10		

driver for such large WSC infrastructures. To address this, reuse of existing fiber-infrastructure and scaling of cross-sectional bandwidth by increasing per-port bandwidth is becoming critical. Introduction of higher port-speed optical interfaces always go through the natural evolution of bleeding-edge technology (e.g. 100GbE today) to maturity (e.g. 10GbE today, was bleeding edge 10 years back), with a gradual reduction in power-consumption per gigabit/second of interconnect [4] (Fig. 3). Broadly, one can break this technology evolution down to three stages:

(a) Bleeding edge: 10× speed increase is obtained for 20× increase in power consumption (e.g. 100GbE 10 × 10 MSA modules consume 14 W as compared to a 10GbE SFP + consuming < 1 W).
(b) Parity: 10× speed increase is obtained for 10× power consumption.
(c) Maturity: 10× speed increase is obtained for 4× power consumption (e.g. 10GbE today as compared to 1GbE interfaces).

Increase of per-port bandwidth directly translates into reduction of radix for the individual switches [4]. As a result, larger number of switching-nodes or fabric stages may become necessary to build the same cross-sectional bandwidth. Fig. 4 illustrates the example of a cluster fabric with 10 Tbps cross-sectional bandwidth. If the fabric is built with a switching node capable of 1 Tbps switching bandwidth, use of increasingly higher speed interfaces lead to step-function jumps in power consumption as larger number of stages are introduced due to radix constraint.

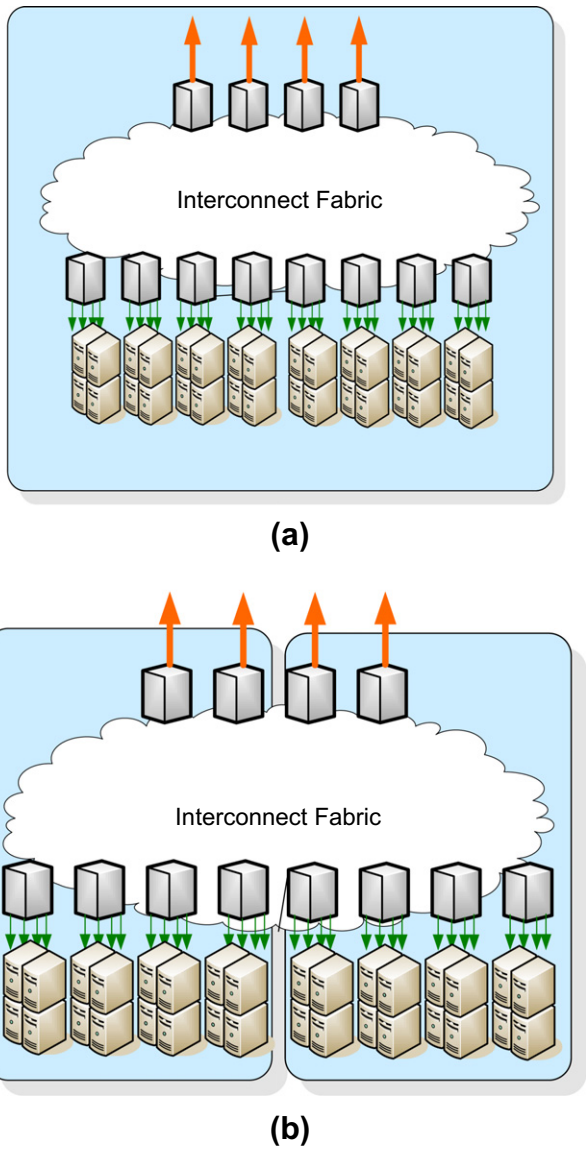

Fig. 2. Hierarchies of intra-datacenter cluster-switching interconnect fabrics (a) within a single building (b) across multiple buildings.

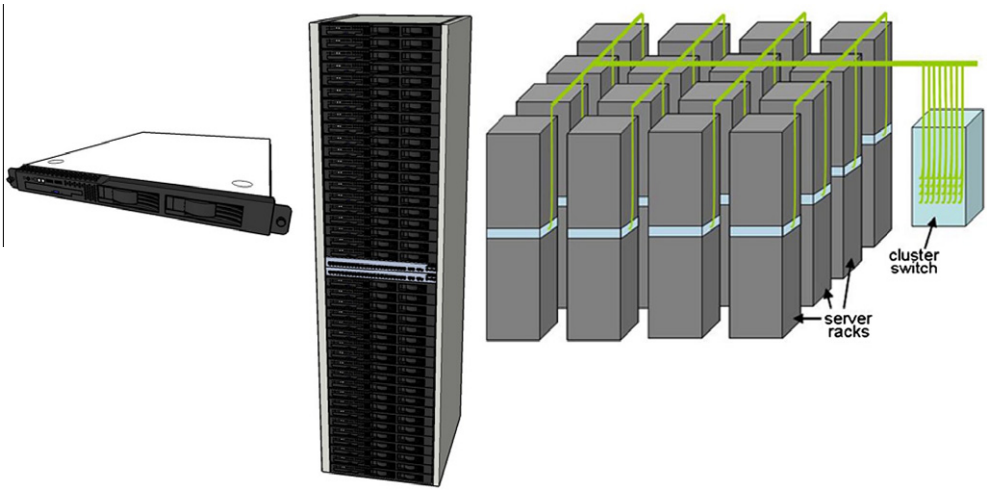

Fig. 1. Typical elements in a Warehouse scale computer.

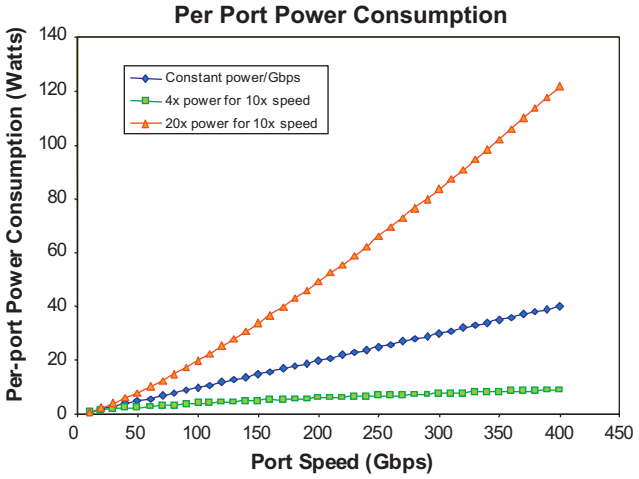

Fig. 3. Evolution of relative power consumption with port-speed for intra-datacenter interconnects at three different technology maturity levels.

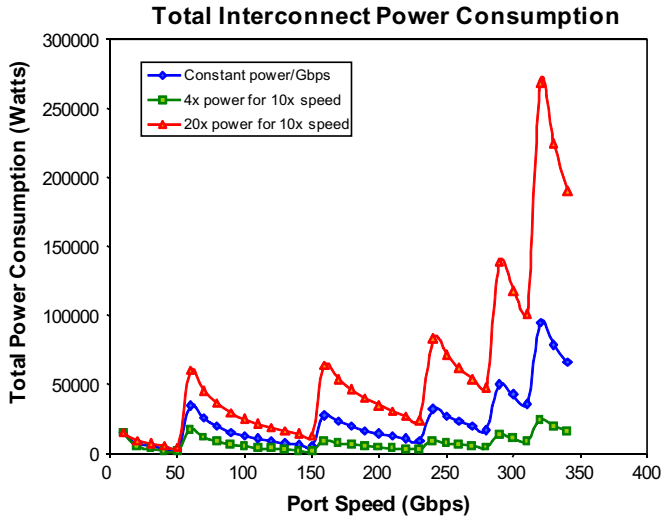

Fig. 4. Total power consumption in a 10 Tbps fabric built with nodes capable of 1 Tbps switching with various interface speeds; three technology maturity curves in-terms of power consumption are considered.

In order to meet requirements of bandwidth and power scaling in warehouse scale computing infrastructures, 100 Gbps and faster optical interfaces must target 4× power consumption for 10× speed as guiding design principle.

3. Inter-datacenter connectivity

A WSC infrastructure can span multiple datacenters. Consequently the cluster aggregation switching fabric will span multiple datacenters as well as shown in Fig. 5.

Typically inter-datacenter connection fabrics are implemented over a fiber-scarce physical layer as the link distances are tens to thousands of kilometers. The fiber-scarcity and limit of available BW on metro and long-haul fiber links lead to a very undesirable available BW between clusters, as illustrated in Fig. 6. If capacity per fiber-pair is not maximized, a bottleneck is introduced due to high oversubscription for inter-datacenter communication [1].

Acceleration of broadband penetration and uptake of internet based applications with rich multi-media contents have led to >40% compound annual growth rate of internet traffic [2] (Fig. 7), with 9 exabytes of traffic volume per month. While the exponential growth of internet traffic drives bandwidth demand for inter-datacenter networks, the Moore's-law growth of processing and storage capacity [5] utilized in the WSC infrastructure drives bandwidth at an even faster pace. Extrapolating the average CAGR of 60% seen in processing-power and storage capacity, one can see that Ethernet standard and port-speeds have kept up well with internet-scale traffic growth but are falling behind Moore's-law (machine-to-machine) traffic growth (Fig. 8.).

Therefore, the need for 100 Gbps and beyond interconnect technologies are immediate for inter-datacenter connections. Various emerging technology building blocks offer the potential for this cost-effective capacity scaling as described below:

(a) *Higher capacity per fiber*: Optical transport solutions that increase the maximum capacity per fiber beyond today's commercially available 8 Tb/s (based on 80 channels of 100 Gbps transmission in C-band). Published literature has shown a roadmap to continued fiber capacity scaling using a number of approaches for increasing the spectral range and spectral efficiency [6,7]. These include higher data rates, higher-order modulation, OFDM, multiple transmission bands etc.

(b) *Unregenerated reach*: As the transmission data rates increase, the unregenerated reach typically decreases due to the higher OSNR required. The use of techniques such as Soft-Decision FEC will help bridge the gap. In addition, techniques for maximizing optical link OSNR across the transmission spectral range such as optimized Raman amplification, tilt control, spectral equalization, per-span launch power adjustment, can be used to increase the maximum unregenerated reach.

(c) *Variable Rate Optics*: With coherent optical transmission systems, it is possible to have a variable transmission rate that is based on the link quality and condition. For example, for shorter links or links with "good" fiber types, the additional optical link margin can be used to transmit higher data rates. This type of variable rate transmission has to be tightly integrated with the packet layer and managed at the system/network level to realize the overall throughput maximization.

(d) *Flexible grid ROADM*: Current commercially available ROADMs are based on 50 GHz or 100 GHz ITU grid spacing. These fixed grid ROADMs become a limitation for future capacity scaling and network flexibility. Emerging flexible grid or gridless ROADMs [8,9], which provide the ability to arbitrarily determine spectral pass bands and spacing between pass bands, enable two key functionalities: (i) support for higher data rates in a spectrally efficient way by packing the wavelengths in a manner determined by the spectral content of the waves rather than the limitations of the ITU grid and (ii) flexibility for arbitrary add/drop of wavelengths in a manner independent of the underlying data rate or modulation scheme.

(e) *Large Core Fibers*: With the ability of coherent systems to compensate for fiber impairments such as chromatic dispersion and polarization mode dispersion, the major remaining fiber impairments that limit transmission are the fiber attenuation and fiber non-linearities. Recent advances in large-core (~110 μm^2), low-attenuation (<0.17 dB/km) fibers demonstrate the capability to increase transmission distance for a given fiber capacity [10]. The large effective area enables a lower power density which helps alleviate penalties from fiber non-linearities.

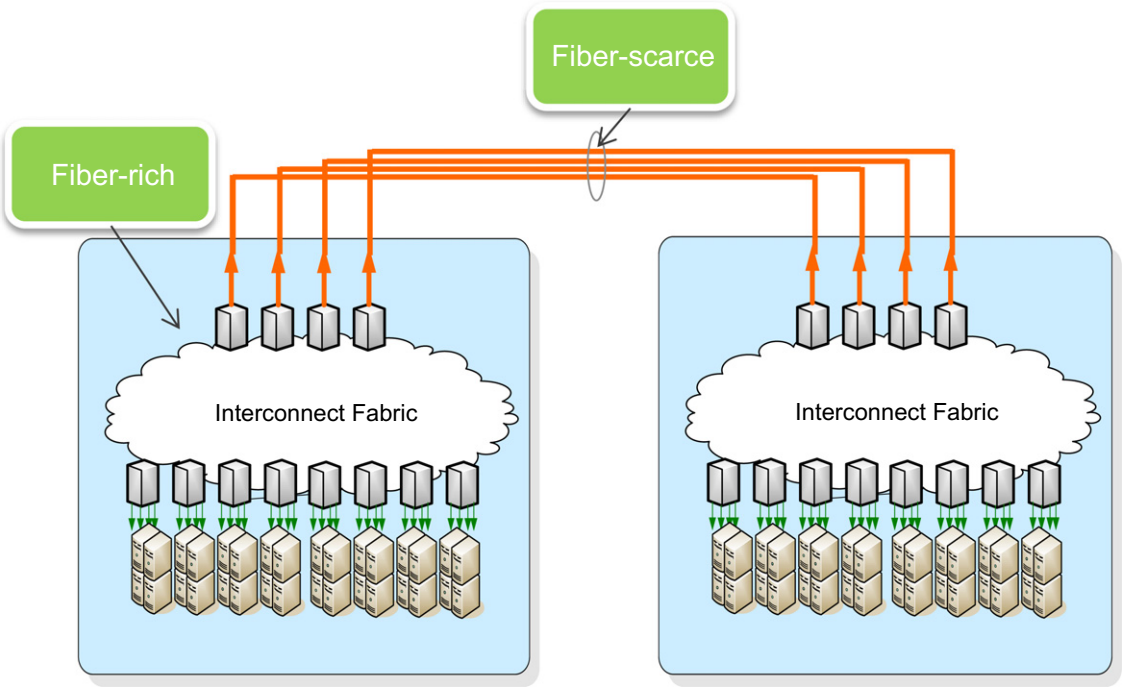

Fig. 5. Inter-datacenter networks connecting multiple WSCs.

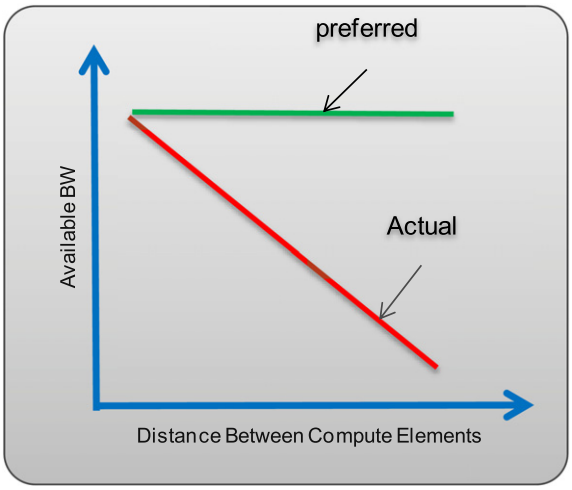

Fig. 6. Inter-dataceneter available bandwidth as a function of distance between the compute elements.

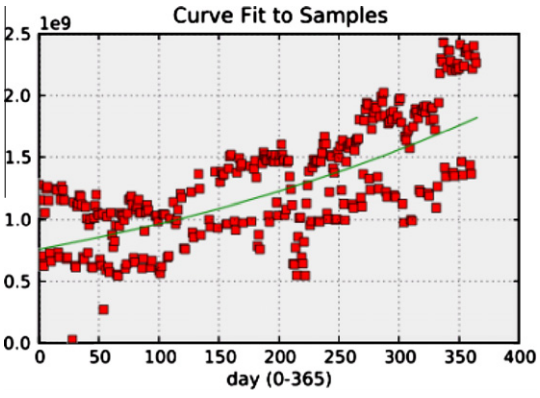

Fig. 7. A > 40% CAGR of internet traffic [3].

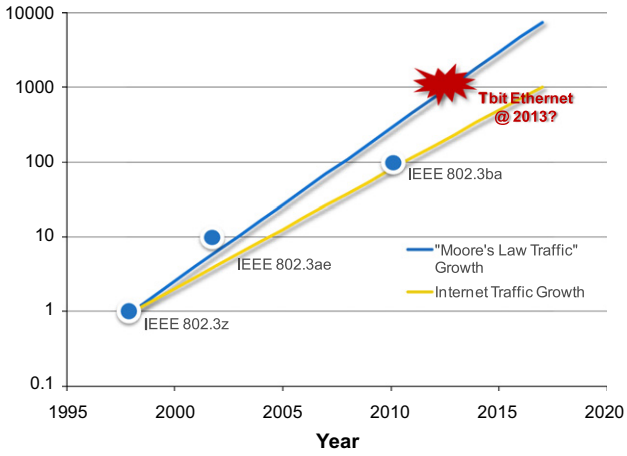

Fig. 8. Ethernet standards and port-speeds compared to internet and extrapolated Moore's-law (machine-to-machine) traffic growth.

4. Conclusions

Advent of warehouse scale computing has been driving the need for bandwidth within and between datacenters. While intra-datacenter connections can take advantage of a fiber-rich physical layer, need for fiber-scarce inter-datacenter connections will drive the adoption of 100GbE and beyond in the massive WSC environments. Deployment of network technologies beyond 100GbE will be needed within the next three to five years for WSC interconnects.

References

[1] L.A. Barroso, 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] C. Labovitz et al., ATLAS Internet Observatory 2009 Annual Report. <http://www.nanog.org/meetings/nanog47/presentations/Monday/Labovitz_ObserveReport_N47_Mon.pdf>.

[3] C.F. Lam, H. Liu, B. Koley, X. Zhao, V. Kamalov, V. Gill, Fiber optic communication technologies: what's needed for datacenter network operations, IEEE Commun. Mag. 48 (7) (2010).
[4] B. Koley, Requirements for data center interconnects, in: 20th Annual Workshop on Interconnections within High Speed Digital Systems, Santa Fe, New Mexico, 3–6 May 2009, Paper TuA2.
[5] Truskowski Morris, The evolution of storage systems, IBM Syst. J. 42 (2) (2003).
[6] René-Jean Essiambre, Gerhard Kramer, Peter J. Winzer, Gerard J. Foschini, Bernhard Goebel, Capacity limits of optical fiber networks, J. Lightw. Technol. 28 (2010) 662–701.
[7] K. Roberts, Digital coherent optical communications beyond 100 Gb/s, in: Signal Processing in Photonic Communications, OSA Technical Digest (CD), Optical Society of America, 2010. Paper JTuA1.
[8] C.F. Lam, W.I. Way, A System's View of Metro and Regional Optical Networks, Photonics West, San Jose, CA, January 29, 2009.
[9] M. Jinno et al., IEEE Commun. Mag. (2009) 66–73.
[10] R. Chen, M. O'Sullivan, C. Ward, S. Asselin, M. Belanger, Next generation transmission fiber for coherent systems, in: Optical Fiber Communication Conference, OSA Technical Digest (CD), Optical Society of America, 2010. Paper OTuI1.