7A3-1 (Invited)

100GbE and Beyond for Warehouse Scale Computing

Bikash Koley*, Vijay Vusirikala*, Cedric Lam*, Vijay Gill*

* Google Inc, USA

***Abstract*--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.**

I. 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-data-centers. 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 data-center [1]. Interconnecting such WSCs in a cost-effective yet scalable way is a unique challenge that needs to be addressed.

II. 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]. Figure 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.

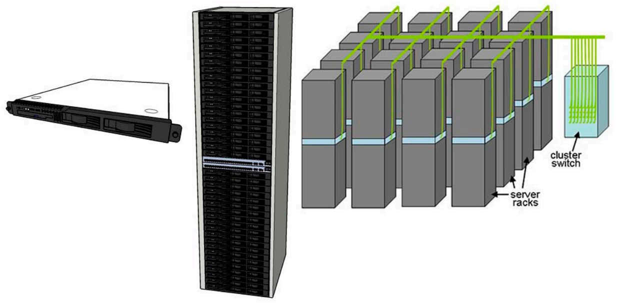
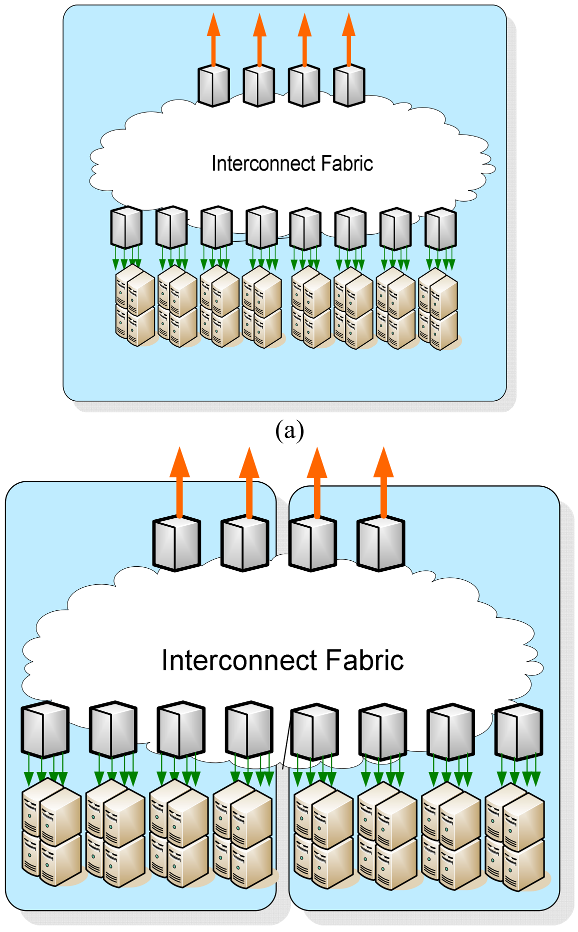


Fig. 1. Typical elements in a Warehouse Scale Computer

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) [2].



(a)

(b)

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

Intra-datacenter networking takes advantage of a fiber rich environment to drive very large bandwidth within and between clusters.

III. Inter-Datacenter Connectivity

A WSC infrastructure can span multiple data-centers.

Consequently the cluster aggregation switching fabric will span multiple data-centers as well as shown in Fig. 3.

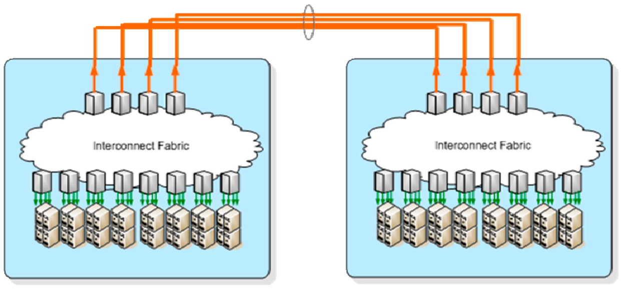


Fig. 3. Inter-datacenter networks connecting multiple WSCs

Typically inter-datacenter connection fabrics are implemented over a fiber-scarce physical layer as the link distances are tens to hundreds of kilometers. 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 [3] (Fig. 3), 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 [4] 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. 5.) .

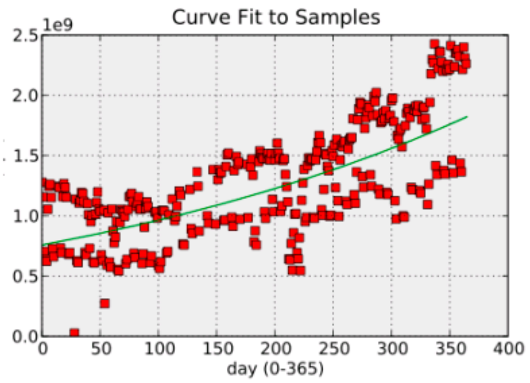


Fig. 4. > 40% CAGR of internet traffic [3]

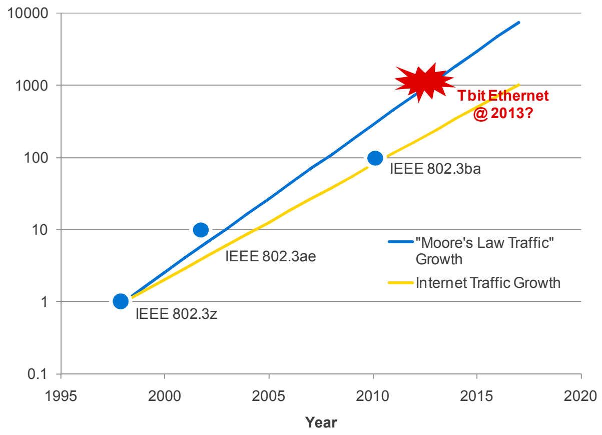


Fig. 5. Ethernet standards and port-speeds compared to Internet and extrapolated Morre's Law (machine-to-machine) traffic growth

Therefore, the need for Ethernet standard supporting speed 100Gbps is immediate for inter-datacenter connections.

IV. 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 Ethernet technology beyond 100GbE will be needed within the next three to five years for WSC interconnects.

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. Labovitz et al: ATLAS Internet Observatory 2009 Annual Report http://www.nanog.org/meetings/nanog47/presentations/Monday/Labovitz_ObserveReport_N47_Mon.pdf

[4] Morris,Truskowski, "The evolution of storage systems", IBM Systems Journal, Vol 42, No 2, 2003