Impact of Emerging Technologies and Topology Selection in Large Scale Data Centers Design

G. Guelbenzu[1], O. Raz[1], and H. J. S. Dorren[1]

[1] Eindhoven University of Technology, COBRA Research Institute, Den Dolech 2, 5600MB, Eindhoven, The Netherlands

We investigate the impact of emerging technologies in large scale data centers implementation, and the effect of network topology selection, by comparing an indirect topology and a direct topology. We conclude that large scale data center networks find many benefits from high radix optically connected switches. We further show that folded-clos scales better than hypercube with increasing number of servers.

Introduction

Over the past 20 years performance of computing systems, including data centers, has been increasing with a factor 1000 every 10 years. Amdahl scaling implies that the system bandwidth scales accordingly [1], and therefore, the system bandwidth of data center networks should approach exascale by 2024.

One important change in network design is the recent commercial availability [2, 3] of switch ASICs supporting 64-ports and 128-ports. 192-ports switches that can act as an optical overlay on an Ethernet network have been also been developed [4]. A second important consideration is that the increasing workload demand combined with a tree-like architecture (with limited east-west data traffic capabilities) leads to increased interest in flattened networks [5], or networks with a flat bisection bandwidth and a flat topology.

Many direct and indirect topologies have been proposed for use in data center networks. Among the direct networks, the (generalized) hypercube [6, 7], is especially interesting because it exploits the connection of several servers per switch (improving the poor scaling of direct networks), while providing full connectivity between switches in every dimension. Among the indirect networks, the folded-clos (sometimes also referred to as fat-tree [8, 9]) is also especially interesting, because it enables networks of any size with full bisection bandwidth, using commodity switches of any number of ports. It is therefore important to compare the scaling of hypercube and folded-clos topologies, using switches with an increasing number of ports, and to investigate how suitable they are to provide flattened networks.

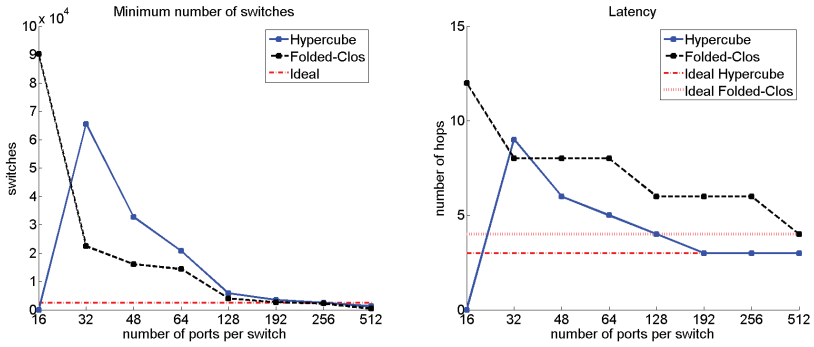
In this paper, we compare these two topologies for switches of increasing radix and find that both are able to provide a full, flat bisection bandwidth, with a reasonably flat topology for switches of radix 256 and above. In general folded-clos scales better than hypercube, providing solutions with less number of switches; and hypercube provides solutions with lower latency than folded-clos. By freeing our architecture from constraints on the numbers of server per rack (typically 40) we find that we can drastically lower the number of switches in the network at the expense of the bisection bandwidth leading us to the conclusion that scaling out switch ASICs may be more beneficial than scaling up.

Topologies comparison

In this study, we look into folded-clos and hypercube networks which are able to interconnect more than 100000 servers divided into racks with full bisection bandwidth. This number of servers is required for exascale performance networks, since at present, petascale performance networks are built with tens of thousands of servers [10]. Full bisection bandwidth is imposed to have a congestion-free network [2].

In order to compare the topologies we use two performance metrics: cost and latency. As a first, rough approximation for the cost, we can use the required number of switches to interconnect the servers. Figure 1a shows the required number of switches for our interconnection network, comparing hypercube and folded-clos topologies for switches with different number of ports. From Figure 1a we conclude that both topologies benefit from increasing the number of ports in the switches, providing solutions with fewer switches. Folded-Clos is the topology that is able to provide solutions for switches of any number of ports. In addition, we conclude that folded-clos always gives solutions with fewer switches than hypercube. For instance, with commercially available 128-ports switches, folded-clos requires 4000 switches, while hypercube needs 5832 switches. We have also plotted the results for switches of 256-ports and 512 ports, and we conclude that, from 256-ports switches, both topologies are able to provide the ideal amount of 2500 switches with present rack configurations. This number can be further reduced if we allow the interconnection of more than 40 servers per rack.

Another important performance metric of the topologies is latency. We use the diameter of the network measured in hops, as it provides an upper bound to the worst-case latency of the network [11]. In the ideal case, a hypercube will require two dimensions and three hops, and a folded-clos will require two levels and four hops. Figure 1b shows the number of hops for our interconnection network, comparing hypercube and folded-clos topologies for switches with different number of ports. We find that with increasing number of ports, fewer hops are needed and that hypercube always provides solutions with less hops than folded-clos. Hypercube reaches its ideal latency of three hops with 192-ports switches, while folded-clos reaches its ideal latency of four hops only using 512-ports switches.



(a) (b)

Figure 1: (a) Required number of switches to interconnect 100000 servers with hypercube and folded-clos topologies (b) Required number of hops for 100000 servers with hypercube and folded-clos topologies

Challenges and opportunities in scaling CMOS switching ASICs

Above we have compared the number of switches and latency of hypercube and folded-clos networks, using the common assumption of populating every rack with 40 servers [12]. Recently micro-servers have been suggested as a viable low cost alternative for traditional servers in high performance computers [13]. For such low power and small volume building blocks one can imagine populating racks in data centers with more than 40 servers. It is therefore of interest to explore how the number of switches in a data center network can be further reduced if more servers are connected to each switch.
As a case study we examine how the number of switches in a hypercube network scales when full bisection bandwidth is not maintained. In Figure 2 we sketch the number of switches needed as a function of the number of servers attached to a switch (i.e.: number of servers per rack). We can see that by adding up to 47 servers per rack, full bisection bandwidth is maintained, and a further drop in cost of 12% is achieved compared to a 40 servers per rack configuration. As more servers are placed into each rack, the bisection bandwidth decreases: with half bisection bandwidth only 55% of the switches are needed, and with quarter bisection bandwidth only 36% of the switches are required.

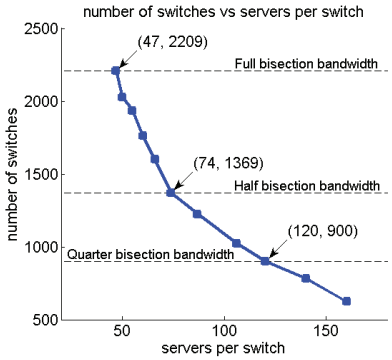


Figure 2: Number of switches as a function of servers per rack for hypercube network using switches with 256 ports

Discussion

The radix of a CMOS ASIC switch is ultimately limited by the number of SerDeses that can be integrated inside it. This is driven by the limited number of pins in a BGA package and the limited power supply [14]. The recent trend of increasing the serial line rate for these SerDeses is only making the problem worse as driving the traces in PCB at higher frequencies requires more power. Increasing the total number of pins is very challenging. One way of freeing more pins for I/O purposes is to reduce the power consumption (freeing power and ground pins). This can be accomplished by bringing the photonic interconnects closer to the switch ASIC or place them on top of [15]. It appears that higher radix switches (scaling out) may have more to offer in terms of system benefits than higher serial line rate (scaling up). This requires designers of ASICs to follow a policy of replacing power pins with data pins supporting short distance electrical transmission lines. In this way a path to flatter and lower cost DC networks can be defined.

Conclusions

In this study, we compare hypercube and folded-clos topologies for the implementation of exascale performance DCNs, using cost and latency as performance metrics. Both topologies are able to provide flattened networks with full and flat bisection bandwidth. In terms of cost, folded-clos provides cheaper solutions than hypercube. In terms of latency, hypercube provides better solutions than folded-clos. We further show that substantial cost savings can be obtained if more than 40 servers are placed in each rack and the bisection bandwidth can be compromised. We conclude that from the total system perspective, the use of optical interconnects supports scaling out (more pins) of switch ASICs. This may prove to be more beneficial than scaling up (higher line rate) as it leads to flatter networks.

Acknowledgements

This work has been supported by the FP7 European Project COSIGN (FP7- 619572).

References

[1] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities", in *Proceedings of the April 18-20, 1967, spring joint computer conference on - AFIPS'67 (Spring)*, 1967, p. 483.

[2] "Broadcom BCM56840_PLUS Switching Technology", [Online]. Available: http://www.broadcom.com/collateral/pb/56840_PLUS-PB00-R.pdf.

[3] "Broadcom BCM56850 StrataXGS Trident II Switching Technology", [Online]. Available: http://www.broadcom.nl/collateral/pb/56850-PB03-R.pdf.

[4] "Polatis - Series 6000 - Single Mode Low Loss All Optical Switch up to 192x192 ports", [Online]. Available:http://www.polatis.com/polatis-series-6000-optical-matrix-switch-192x192-sdn-enabled-industry-leading-performace-lowest-loss-switches.asp.

[5] A. Hendel, "Large-Scale Data Center Networks: Component and Technology Insights", in *OFC, Th4J.1*, 2014.

[6] L. N. Bhuyan et al, Generalized Hypercube and Hyperbus Structures for a Computer Network", *IEEE Trans. Comput.*, vol. C-33, no. 4, pp. 323-333, Apr. 1984.

[7] J. H. Ahn et al, "HyperX", in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis - SC '09*, 2009, p. 1.

[8] C. E. Leiserson, "Fat-trees: Universal networks for hardware-efficient supercomputing", *IEEE Trans. Comput.*, vol. C-34, no. 10, pp. 892-901, Oct. 1985.

[9] M. Al-Fares et al "A scalable, commodity data center network architecture", in *Proceedings of the ACM SIGCOMM 2008 conference on Data communication - SIGCOMM '08*, 2008, vol. 38, no. 4, p. 63.

[10] A. Greenberg et al "Towards a next generation data center architecture: scalability and commoditization", *ACM Work. Program. routers extensible Serv. tomorrow*, pp. 57-62, 2008.

[11] D. Abts et al, "High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities", *Synth. Lect. Comput. Archit.*, vol. 6, no. 1, pp. 1-115, Mar. 2011.

[12] L. A. Barroso et al, "The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines", *Synth. Lect. Comput. Archit.*, vol. 4, no. 1, pp. 1-108, Jan. 2009.

[13] R. P. Luijten et al, "The DOME embedded 64 bit microserver demonstrator", in *Proceedings of 2013 International Conference on IC Design & Technology (ICICDT)*, 2013, pp. 203-206.

[14] "2013 ITRS Summary", [Online]. Available: http://www.itrs.net/Links/2013ITRS/Summary2013.htm.

[15] K. Hasharon et al, "A High End Routing Platform for Core and Edge Applications Based on Chip To Chip Optical Interconnect", in *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2013*, 2013, p. OTu3H.2.