Core Techniques on Designing Network for Data Centers: A Comparative Analysis

HE Xiaobo[1, a]

[1]Chongqing water resources and electric engineering college, Chongqing 402160,China

[a]hexiaobo@126.com

Keywords: Data Center; Core Techniques; Network Structure; Comparative Analysis;

Abstract. With the fast development of data science and Internet technology, the bursting need for data storage and analysis makes it necessary to build up data centers. In this paper, we discuss the core techniques on designing network. Advances in data intensive computing and high performance computing facilitate rapid scaling of data center networks, resulting in a growing body of research exploring new network architectures that enhance scalability, cost effectiveness and performance. Understanding the tradeoffs between these different network architectures could not only help data center operators improve deployments, but also assist system designers to optimize applications running on top of them. We analyze the state-of-the-art network architectures of famous data centers and propose our novel core techniques. With the implementation and experiment, we verify the robustness of our proposed technology.

Introduction

The computer industry has been actively building large scale data centers that deliver enormous computation power and storage capacity needed by data-intensive applications [1]. By tens of thousands of nodes cluster has become in recent years. With the increasing scale of the network, reducing overall system infrastructure costs and achieve higher levels of performance has become a problem order data center operators. Data center architectures often have different end goals that require optimization of different characteristics. If the workload is compute-intensive, data centers need to be equipped with powerful nodes. For communication-intensive workloads, data center networks play a critical role in delivering performance while making sure that costs are affordable. Most of the existing work, proposed based on single scale network architecture and topology with specific targets and analysis of network structure. To the best of our knowledge, a whole variety of comparative analysis of network structure does not exist in the literature. While cost comparison analysis is useful to analyze different data center architectures [2], we note that quantifying and comparing other dimensions such as scalability and power can yield further insights.

In this work, we conduct a comparative analysis of several representative data center network architectures. We present a list of key metrics to depict performance and cost, and analyze our representative architectures in terms of these metrics. The specific contributions of our work are: (1) We comprehensively compare contemporary popular and representative data center topologies by analyzing significant metrics in data centers including scalability, latency and hop counts, path diversity, cost and power. (2) We evaluate different topological structure, network throughput typical data center traffic mode using the minimum network simulation. To the best of our knowledge, this is the first work, compare the throughput of topological structure of various influence on the overall system of data center. (3) With in-depth analysis, we give recommendations for practical data center topology implementation based on different network sizes.

Data Center Network Classification

Switch-only Topologies. (1) Multi-tiered Network: Multi-tiered design is a traditional data center architecture that is commonly used in many medium-to-large enterprises. A three-tiered topology contains core switches at the root level, aggregation switches at the middle level, and access level

© 2015. The authors - Published by Atlantis Press

switches connected to the hosts (see Figure 1). In this work, we assume that all of the core layer and aggregation layer switches use 40 Gigabit Ethernet ports. Each access switches with Gigabit Ethernet port connection to a host and a 10 Gigabit Ethernet uplink aggregation switch. A basic multi layers network parameters were oversubscribed ratio. To the best of our knowledge, there is no standard definition of oversubscription rate; the researchers tend to use their own definition, may be a specific topology. For example, in a tree topology, the over subscription rate is usually defined as the uplink to downlink bandwidth than the bandwidth. We adopt a more generalized definition: oversubscription is injected into the network bandwidth than network. (2) Fat Tree Network: The design of fat tree network is motivated by the fact that the price differential between high end switch (switches with higher link bandwidth or higher number of ports) and low-end switches is considerably large. The main idea behind the fat tree topology structure is to replace high-end switches in multi-level interconnection topology some low-end switches. The main difference between fat tree is the convergence layer and core layer switches all is a group of low interconnect replaced end switch. Each subset is called a pod in Figure 2. As the number of uplinks and downlinks for each pod are equal, fat tree has full bisection bandwidth. (3) Flattened Butterfly Network: Flattened butterfly topology [3] was originally proposed for on-chip interconnection network. Figure 3 shows an 8-ary 2-flat FBFLY. Each square in the figure represents a switch, and each of the 8 switches interconnects with the other 7 switches. In addition, each switch links with 8 host nodes (i.e., servers).

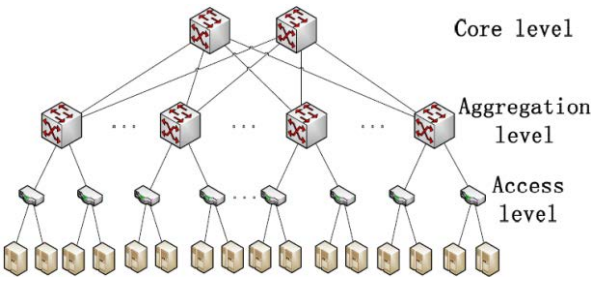

Fig. 1 The Multi-tiered Network

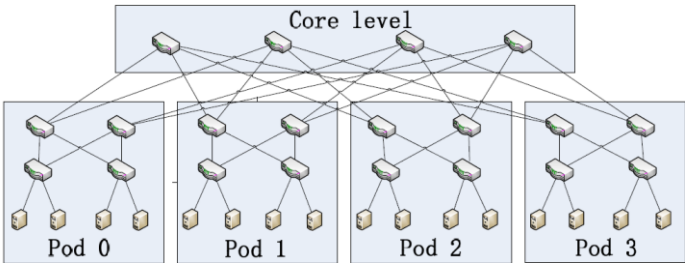

Fig. 2The Fat Tree Network

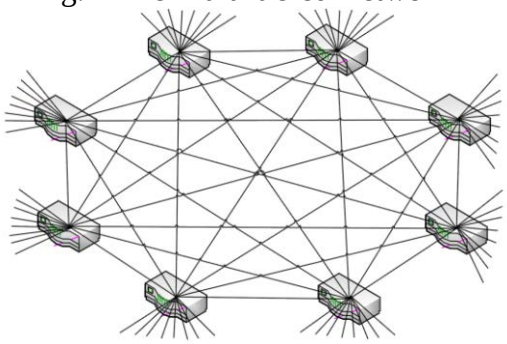
Fig. 3The Flattened Butterfly

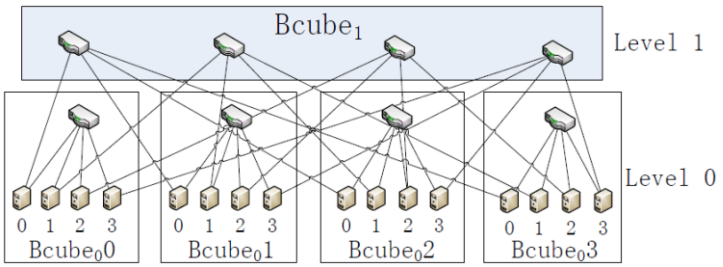

Fig. 4The BCube1

Hybrid Topologies. (1) Camcube: In server-based data center architectures, the data center is created using a set of servers, where each server typically has a multi-core processor, and a high-performance network interface card (NIC) with multiple ports. The servers are not only end hosts, but also perform packet forwarding and routing. (2) BCube1: The BCube1 architecture [4] uses both switches and servers for routing traffic. Figure 4 shows a sample BCube1 architecture.

Experiment and Comparison Metrics

The Scalability Metrics. Scalability is the ability of a system, to handle a growing amount of work in a capable manner or its ability to be enlarged to accommodate that growth. Comparing different topologies scalability, topological structure we need to set the oversubscribed is the same. We will discuss later, the over subscription rate is the main factor affecting network scalability; it is basically a metric to quantify the network bandwidth between all the host sharing. (a) For multi-tiered network the number of switch ports per host can be written as formula 1. (b) For fat tree topology, the

network's oversubscription ratio is a fixed 1:1, the number of switch ports per host can be written as formula 2. (c) For FBFLY, refer to formula 3. The figure 5 illustrate the detailed plot.

$$\left(4f \cdot f/2 + 4f^2 + 8f \cdot f/2\right)/2f^2 \tag{1}$$

$$\left(\frac{5}{4}f^2 \cdot \sum 4f_i\right)/f^3 \tag{2}$$

$$\left((k+1)n^k \cdot n\right)/n^{k+1} \tag{3}$$

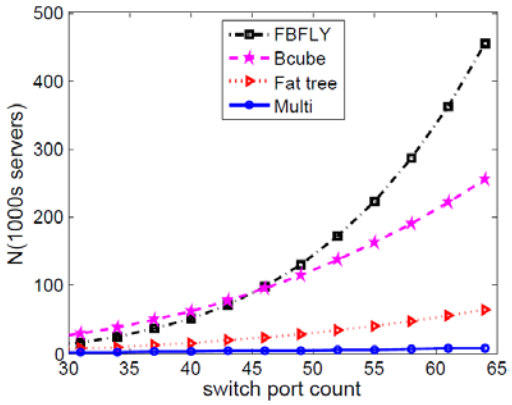

Fig. 5The Number of Hosts

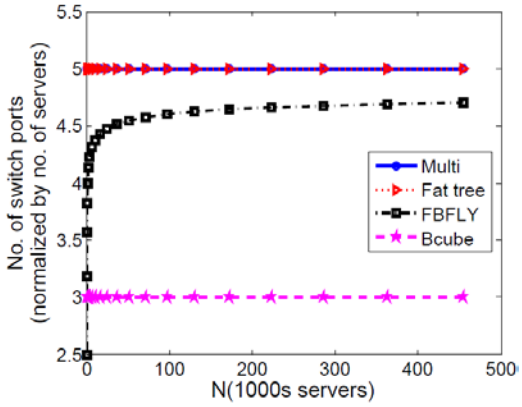

Fig. 6The Number of Switch Ports

The Path Diversity Metrics. Path diversity is an important metric for two reasons: a multiplicity of paths can improve load balance and enhance throughput by distributing the traffic load, and the network is more immune to link and switch failures. In this paper, we define path diversity as the number of different shortest paths between a pair of hosts. We consider both the maximum (over all host pairs) number of shortest paths between a pair of hosts, and the average number (over all host pairs) of shortest paths. These paths are not necessarily disjoint. As a second measure, we consider the number of disjoint paths (not necessarily shortest), both maximum and average over all host pairs. In order to maintain fairness of the comparison in terms of path diversity, we need to set oversubscription of the networks the same. Here we set it to be 1:1 for all architectures to equalize the performance.

The Hop Count Metrics. The average hop count is the average number of hops on the shortest path between a pair of hosts (averaged over all host pairs). This metric is useful in deducing packet latencies. Similar to the comparison with respect to path diversity, we choose an oversubscription ratio of 1:1 to normalize the performance, and the parameters are the same as those for path diversity comparison. For multi-tiered and fat tree topologies, we use the same parameters p,q,r we defined earlier for the path diversity comparison. The average hop count for multi-tiered topology is:

$$\frac{2p + 4q + 6(r - p - q)}{r} \tag{4}$$

For FBFLY and BCube, the hop counts are calculated using Hamming distance. The comparison results are shown in Figure 6. We notice that for smaller number of hosts, FBFLY and BCube have lower hop count than others making them more suitable for smaller scale data centers.

The Throughput Metrics. In a network, throughput (or accepted traffic) is the rate (bits/sec) at which traffic is delivered to the destination nodes [3]. In our experiment, we show normalized throughput over maximum achievable injection bandwidth. We performed simulations using Mini-net [4], a software network emulator, and network that statically analyzes network constructs and features. We ran our simulations on a Xeon x5472 3.0GHz quad-core CPU machine with 8G DRAM. At present, we have a switch topologies, such as simulators requires considerable restructuring efforts it for other topologies work. We compare our topological results in a star network host each pair of connected by a single none use of special double blocking switch link. In our experiments, each topology with link capacity 16 host 10 Mbps (in addition to the extent of aggregation link capacity of 40 Mbps in a multilevel topology). We study the topologies using several types of workloads: hotspot, random and stride. Each host in Mini-net runs a shell program and

communication among programs is modeled for the above-mentioned workload patterns. The average throughput for the workloads in each architecture is measured by averaging over three independent runs. Later we would like to use the distance metrics [5-7] to modify the throughput metrics.

Conclusion and Summary

In this work, we presented a comparative analysis of several popular core data center network topologies such as Fat tree, Multi-tiered networks, Flattened Butterfly, Camcube and BCube. The metrics that we chose for comparison were scalability, path diversity, hop count, throughput and some more potential ones. We find that different topologies scale differently for various metrics, and we conclude that data center designers have to consider such characteristics to maximize their performance while minimizing cost and power. In the near future, we will study energy optimization strategies and application-specific constraints to better understand data center needs and design.

References

[1] L. Popa, S. Ratnasamy, G. Iannaccone, A. Krishnamurthy, and I. Stoica, "A cost comparison of datacenter network architectures," in Co-NEXT, 2010.

[2] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Vl2: a scalable and flexible data center network," in SIGCOMM, 2009.

[3] R. Niranjan Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "Portland: a scalable fault-tolerant layer 2 data center network fabric," in SIGCOMM, 2009.

[4] J. Mudigonda, P. Yalagandula, and J. C. Mogul, "Taming the flying cable monster: A topology design and optimization framework for data-center networks," in USENIX ATC, 2011.

[5] Wang, Haoxiang, Ferdinand Shkjezi, and Ela Hoxha. "Distance metric learning for multi-camera people matching." Advanced Computational Intelligence (ICACI), 2013 Sixth International Conference on. IEEE, 2013.

[6] Kireeva, Natalia V., et al. "Impact of distance-based metric learning on classification and visualization model performance and structure–activity landscapes." Journal of computer-aided molecular design 28.2 (2014): 61-73.

[7] Huang, Xiaoman, and Mark A. Friedl. "Distance metric-based forest cover change detection using MODIS time series." International Journal of Applied Earth Observation and Geoinformation 29 (2014): 78-92.