SPRINGER BRIEFS IN COMPUTER SCIENCE

Yang Liu
Jogesh K. Muppala
Malathi Veeraraghavan
Dong Lin
Mounir Hamdi

Data Center Networks

Topologies, Architectures and Fault-Tolerance Characteristics

Springer

SpringerBriefs in Computer Science

For further volumes:
http://www.springer.com/series/10028

Yang Liu • Jogesh K. Muppala
Malathi Veeraraghavan • Dong Lin
Mounir Hamdi

Data Center Networks

Topologies, Architectures and Fault-Tolerance Characteristics

Springer

Yang Liu
Jogesh K. Muppala
Dong Lin
Mounir Hamdi
Department of Computer Science
and Engineering
The Hong Kong University of Science
and Technology
Kowloon, Hong Kong, SAR

Malathi Veeraraghavan
Department of Electrical
and Computer Engineering
University of Virginia
Charlottesville, VA, USA

ISSN 2191-5768 ISSN 2191-5776 (electronic)
ISBN 978-3-319-01948-2 ISBN 978-3-319-01949-9 (eBook)
DOI 10.1007/978-3-319-01949-9
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013948240

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Large-scale data centers form the core infrastructure support for the ever expanding cloud based services. Thus the performance and dependability characteristics of data centers will have significant impact on the scalability of these services. In particular, the data center network needs to be agile and reconfigurable in order to respond quickly to ever changing application demands and service requirements. Significant research work has been done on designing the data center network topologies in order to improve the performance of data centers.

In this book, we present a detailed overview of data center network architectures and topologies that have appeared in the literature recently. We start with a discussion on various representative data center network topologies, and compare them with respect to several properties in order to highlight their advantages and disadvantages. Thereafter, we discuss several routing algorithms designed for these architectures, and compare them based on various criteria: the basic algorithms to establish connections, the techniques used to gain better performance and the mechanisms for fault-tolerance. A good understanding of the state-of-the-art in data center networks would enable the design of future architectures in order to improve performance and dependability of data centers.

Hong Kong, P. R. China — Yang Liu
Hong Kong, P. R. China — Jogesh K. Muppala
Charlottesville, VA, USA — Malathi Veeraraghavan
Hong Kong, P. R. China — Dong Lin
Hong Kong, P. R. China — Mounir Hamdi

Contents

Acronyms

ABT	Aggregated Bottleneck Throughput
APL	Average Path Length
ARP	Address Resolution Protocol
BFD	Bidirectional Forwarding Detection
BSR	BCube Source Routing
CDN	Component Decomposition Number
DCN	Data Center Networks
DFR	DCell Fault-tolerant Routing
ECMP	Equal Cost Multi-Path
EoR	End of Row
IBA	InfiniBand
IP	Internet Protocol
IPv4/IPv6	Internet Protocol Version 4/6
LCS	Largest Component Size
LDM	Location Discovery Message
LDP	Location Discovery Protocol
LISP	Locator-Identifier Split Protocol
MAC	Medium Access Control
MTBF	Mean Time Between Failures
MTTR	Mean Time To Repair
NIC	Network Interface Card
OSPF	Open Shortest Path First
PBB	Provider Backbone Bridging
RFR	Routing Failure Rate
RSM	Replicated State Machine
SCS	Smallest Component Size
TCP	Transmission Control Protocol
ToR	Top of Rack
TRILL	Transparent Interconnection of Lots of Links
UTP	Unshielded Twisted Pair
VLAN	Virtual Local Area Networks

VLB	Valiant Load Balancing
VM	Virtual Machine
WDM	Wavelength Division Multiplexing

Chapter 1
Introduction

1.1 Introduction

Data center infrastructure design has recently been receiving significant research interest both from academia and industry, in no small part due to the growing importance of data centers in supporting and sustaining the rapidly growing Cloud-based applications including search (e.g., Google, Bing), video content hosting and distribution (e.g., YouTube, NetFlix), social networking (e.g., Facebook, Twitter), and large-scale computations (e.g., data mining, bioinformatics, indexing). For example, the Microsoft Live online services are supported by a Chicago-based data center, which is one of the largest data centers ever built, spanning more than 700,000 square feet. In particular, cloud computing is characterized as the culmination of the integration of computing and data infrastructures to provide a scalable, agile and cost-effective approach to support the ever-growing critical IT needs (in terms of computation, storage, and applications) of both enterprises and the general public [2, 8].

Massive data centers providing storage form the core of the infrastructure for the Cloud [8]. It is thus imperative that the data center infrastructure, including the data center network, be well designed so that both the deployment and maintenance of the infrastructure is cost-effective. With data availability and security at stake, the role of the data center is more critical than ever.

The topology of the network interconnecting the servers has a significant impact on the agility and reconfigurability of the data center infrastructure to respond to changing application demands and service requirements. Today, data center networks primarily use top of rack (ToR) switches that are interconnected through end of row (EoR) switches, which are in turn connected via core switches. This approach leads to significant bandwidth oversubscription on the links in the network core [1]. This prompted several researchers to suggest alternate approaches for scalable cost-effective network architectures. According to the reconfigurability of the topology after the deployment of the DCN, there are fixed topology and flexible topology networks. Fixed topology networks can be further classified into

Y. Liu et al., *Data Center Networks*, SpringerBriefs in Computer Science,
DOI 10.1007/978-3-319-01949-9_1, © The Author(s) 2013

two categories: tree-based topologies such as fat-tree[1] and Clos Network [9], and recursive topologies such as DCell [10], BCube [11]. Flexible topologies such as c-Through [14], Helios [7] and OSA [6] enable reconfiguration of their network topology at run time based on the traffic demand. Every approach is characterized by its unique network topology, routing algorithms, fault-tolerance and fault recovery approaches.

Our primary focus in this book is examining data center network topologies that have been proposed in research literature. We start with a discussion on the current state-of-the-art in data center network architectures. Then we examine various representative data center topologies that have been proposed in the research literature, and compare them on several dimensions to highlight the advantages and disadvantages of the topologies. Thereafter, we discuss the routing protocols designed for these topologies, and compare them based on various criteria, such as the algorithms used to gain better performance and the mechanisms for fault-tolerance. Our goal is to bring out the salient features of the different approaches such that these could be used as guidelines in constructing future architectures and routing techniques for data center networks.

We note that other researchers have conducted thorough surveys on other important issues about data center networks, such as routing in data centers [5], and data center virtualization [3]. Kachris et al. published a survey focusing on optical interconnects for data centers [12]. They cover some of the topologies that will be discussed in this paper. Wu et al. made comparisons of some existing DCN architectures [15]. Zhang et al. compared DCN architectures from the perspectives of congestion notification algorithms, TCP incast and power consumption [16].

1.2 Data Center Applications

Cloud computing introduces a paradigm shift in computing where businesses and individuals are no longer required to own and operate dedicated physical computing resources to provide services to their end-users over the Internet. Cloud computing relieves its users from the burdens of provisioning and managing their own data centers and allows them to pay for resources only when they are actually needed and used. Cloud computing can be provided by public clouds (e.g., Amazon EC2, Microsoft Azure) or by private clouds maintained and used within an organization. The services provided by the cloud range from IaaS (infrastructure as a service) where the cloud user requests one or more virtual machines hosted on the cloud resources to SaaS (software as a service) where the cloud user is able to use a service (e.g., Customer Relationship Management software) hosted on cloud resources. The use of cloud computing technology is increasing rapidly and, according to industry estimates, the global cloud computing market is expected to exceed $100 billion by 2015.

Massive data centers hosting the computing, storage, and communication resources form the core of the support infrastructures for cloud computing. With

the proliferation of cloud computing and cloud based services, data centers are becoming increasingly large with massive number of servers and massive amount of storage. The entire infrastructure is orchestrated by the Data Center Network to work as a cohesive whole. Server virtualization is increasingly employed to make the data center flexible and adapt to varying demands. In addition, network virtualization is also being considered [3].

Data centers typically run two types of applications: outward facing (e.g., serving web pages to users) applications, and internal computations (e.g., MapReduce for web indexing). In general multiple services run concurrently within a data center, sharing computing and networking resources. Workloads running on a data center are often unpredictable: Demand for new services may spike unexpectedly, and place undue stress on the resources. Furthermore, server failures are to be expected given the large number of servers used [4].

Greenberg et al. [8] define *agility* as an important property, requiring the data center to support running of any service on any server at any time as per the need. This can be supported by turning the servers into a single large fungible pool, thus letting services dynamically expand and contract their footprint as needed. This is already well-supported in terms of both storage and computation, for example by Google's GFS, BigTable, MapReduce etc. This requires the infrastructure to project the illusion to the services of equidistant end-points with non-blocking communication core supporting unlimited workload mobility. The net result is increased service developer productivity, and lower cost while achieving high performance and reliability. In the following section we discuss how these application requirements influence data center network design.

1.3 Data Center Network Requirements

Mysore et al. [13] list the following requirements for scalable, easily manageable, fault-tolerant and efficient Data Center Networks (DCN):

1. Any VM may migrate to any physical machine without the need for a change in its IP address
2. An administrator should not need to configure any switch before deployment
3. Any end host should efficiently communicate with any other end host through any available paths
4. No forwarding loops
5. Failures will be common at this scale, and hence their detection should be rapid and efficient

They then elaborate on the implications of the above requirements on network protocols: a single Layer 2 fabric for entire data center (1&2), MAC forwarding tables with hundreds of thousands entries (3), and efficient routing protocols which disseminate topology changes quickly to all points (5).

The above requirements on the network infrastructure imply that it should provide uniform high capacity between any pair of servers. Furthermore, the capacity between servers must be limited only by their NIC speeds. Topology independence for addition and removal of servers should be supported. Performance isolation should be another requirement, which means that traffic from one application should be unaffected by others. Ease of management supporting "Plug-and-Play" (Layer 2 semantics) should be the norm. The network should support flat addressing, so that any server can have any IP address. Legacy applications depending on broadcast must work unhindered.

1.4 Summary

In this chapter, we presented a basic introduction to data center networks. We briefly mentioned the typical characteristics of data center applications that dictate the expectations from the underlying data center networks. We then briefly reviewed the data center network requirements presented in the literature, which need to be met by future data center networking architectures. The subsequent chapters provide a detailed coverage of various data center network topologies, their performance and fault-tolerance characteristics.

References

1. Al-Fares, M., Loukissas, A., Vahdat, A.: A scalable, commodity data center network architecture. In: Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication, Seattle, pp. 63–74. ACM (2008)
2. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M.: Above the clouds: a berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley (2009). http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html
3. Bari, M., Boutaba, R., Esteves, R., Granville, L., Podlesny, M., Rabbani, M., Zhang, Q., Zhani, M.: Data center network virtualization: a survey. IEEE Commun. Surv. Tutor. **PP**(99), 1–20 (2012). doi:10.1109/SURV.2012.090512.00043
4. Barroso, L., Hölzle, U.: The datacenter as a computer: an introduction to the design of warehouse-scale machines. Synth. Lect. Comput. Archit. **4**(1), 1–108 (2009)
5. Chen, K., Hu, C., Zhang, X., Zheng, K., Chen, Y., Vasilakos, A.: Survey on routing in data centers: insights and future directions. IEEE Netw. **25**(4), 6–10 (2011)
6. Chen, K., Singla, A., Singh, A., Ramachandran, K., Xu, L., Zhang, Y., Wen, X., Chen, Y.: OSA: an optical switching architecture for data center networks with unprecedented flexibility. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, San Jose, pp. 18–18. USENIX Association (2012)
7. Farrington, N., Porter, G., Radhakrishnan, S., Bazzaz, H., Subramanya, V., Fainman, Y., Papen, G., Vahdat, A.: Helios: a hybrid electrical/optical switch architecture for modular data centers. In: ACM SIGCOMM Computer Communication Review, vol. 40, pp. 339–350. ACM, New York, NY, USA (2010)

8. Greenberg, A., Hamilton, J., Maltz, D.A., Patel, P.: The cost of a cloud: research problems in data center networks. SIGCOMM Comput. Commun. Rev. **39**(1), 68–73 (2009). doi:http://doi.acm.org/10.1145/1496091.1496103
9. Greenberg, A., Hamilton, J.R., Jain, N., Kandula, S., Kim, C., Lahiri, P., Maltz, D.A., Patel, P., Sengupta, S.: VL2: a scalable and flexible data center network. SIGCOMM Comput. Commun. Rev. **39**(4), 51–62 (2009). doi:http://doi.acm.org/10.1145/1594977.1592576
10. Guo, C., Wu, H., Tan, K., Shi, L., Zhang, Y., Lu, S.: DCell: a scalable and fault-tolerant network structure for data centers. ACM SIGCOMM Comput. Commun. Rev. **38**(4), 75–86 (2008)
11. Guo, C., Lu, G., Li, D., Wu, H., Zhang, X., Shi, Y., Tian, C., Zhang, Y., Lu, S.: BCube: a high performance, server-centric network architecture for modular data centers. ACM SIGCOMM Comput. Commun. Rev. **39**(4), 63–74 (2009)
12. Kachris, C., Tomkos, I.: A survey on optical interconnects for data centers. IEEE Commun. Surv. Tutor. **14**(4), 1021–1036 (2012). doi:10.1109/SURV.2011.122111.00069
13. Niranjan Mysore, R., Pamboris, A., Farrington, N., Huang, N., Miri, P., Radhakrishnan, S., Subramanya, V., Vahdat, A.: Portland: a scalable fault-tolerant layer 2 data center network fabric. ACM SIGCOMM Comput. Commun. Rev. **39**(4), 39–50 (2009)
14. Wang, G., Andersen, D., Kaminsky, M., Papagiannaki, K., Ng, T., Kozuch, M., Ryan, M.: c-Through: part-time optics in data centers. In: ACM SIGCOMM Computer Communication Review, vol. 40, pp. 327–338. ACM, New York, NY, USA (2010)
15. Wu, K., Xiao, J., Ni, L.: Rethinking the architecture design of data center networks. Front. Comput. Sci. **6**, 596–603 (2012). doi:10.1007/s11704-012-1155-6. http://dx.doi.org/10.1007/s11704-012-1155-6
16. Zhang, Y., Ansari, N.: On architecture design, congestion notification, TCP incast and power consumption in data centers. IEEE Commun. Surv. Tutor. **15**(1), 39–64 (2012)

Chapter 2
Data Center Network Topologies: Current State-of-the-Art

2.1 Typical Data Center Network Topology

A typical data center network configuration is shown in Fig. 2.1. In this configuration, the end hosts connect to top of rack (ToR) switches typically using a 1 GigE link. The ToR switches typically contain 48 GigE ports connecting to the end hosts, and up to four 10 GigE uplinks. The ToR switches sometimes connect to one or more end of row (EoR) switches. The design of the data center network topology is to provide rich connectivity among the ToR switches so that the requirements set out in Sect. 1.3 are satisfied.

Forwarding of packets may be handled either at Layer 3 or Layer 2 depending on the architecture. If the Layer 3 approach is used, then IP addresses are assigned to hosts hierarchically based on their directly connected switch. Standard intra-domain routing protocols, eg. OSPF, can be used for routing. However this approach incurs large administration overhead. An alternative is to use a Layer 2 approach. Here address assignment is based on flat MAC addresses and forwarding is done accordingly. This approach incurs less administrative overhead. However this approach suffers from poor scalability and low performance. A middle ground between Layer 2 and Layer 3 is to use virtual LANs (VLANs). This approach is feasible for smaller scale topologies, however it suffers from the resource partitioning problem [5].

Two types of traffic shown in Fig. 2.1, viz., North-South traffic and East-West traffic, place different demands on the networking infrastructure. North-South traffic corresponds to communication between the servers and the external world. East-West traffic is the internal communication among the servers. Depending on the type of application (outward facing or internal computation), one or the other type of traffic is dominant.

End host virtualization is extensively used in today's data centers, enabling physical servers to support multiple virtual machines (VM). One consequence of this approach is the need to support a large number of addresses and VM migrations (e.g. vMotion). In a layer 3 fabric, migrating a VM to a different switch changes

Y. Liu et al., *Data Center Networks*, SpringerBriefs in Computer Science,
DOI 10.1007/978-3-319-01949-9_2, © The Author(s) 2013

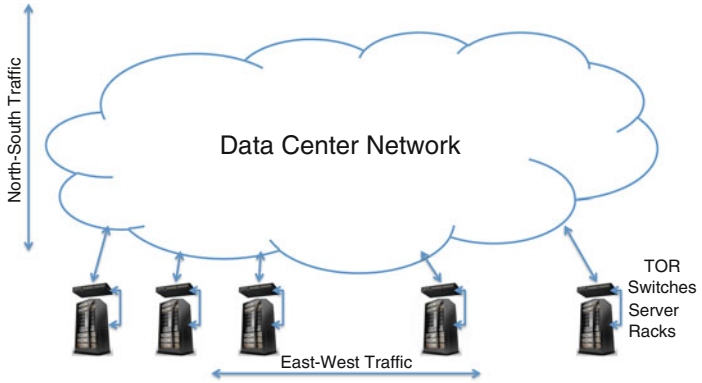

Fig. 2.1 A typical data center network topology

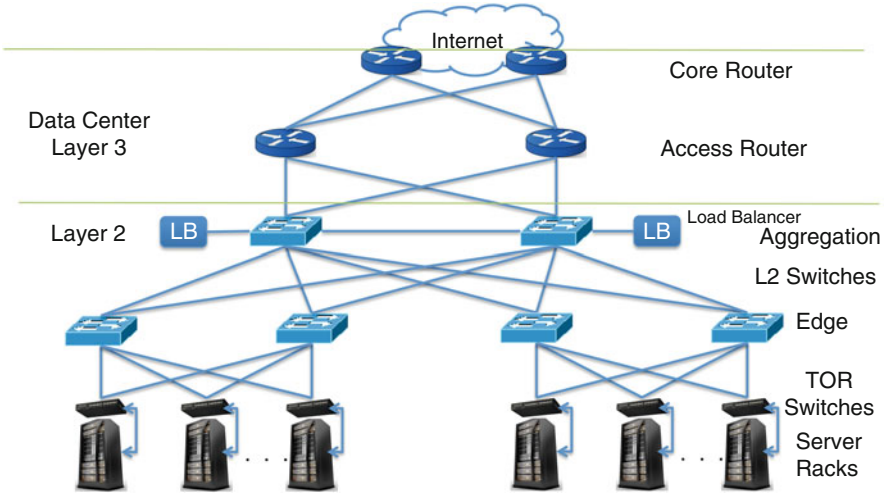

Fig. 2.2 Cisco's recommended DCN topology

the VM's IP address. In a layer 2 fabric, migrating a VM incurs ARP overhead, and requires forwarding on millions of flat MAC addresses.

2.1.1 Tree-Based Topology

Tree-based topologies have been the mainstay of data center networks. As an example, Cisco [2] recommends a multi-tier tree-based topology as shown in Fig. 2.2. ToR switches connect to edge switches, and edge switches in turn connect to aggregation switches. Aggregation switches in turn are connected to the core.