Data Center Networking

Deke Guo

Data Center Networking

Network Topologies and Traffic Management in Large-Scale Data Centers

清華大学出版社
TSINGHUA UNIVERSITY PRESS

Springer

Deke Guo
School of System Engineering
National University of Defense Technology
Changsha, Hunan, China

ISBN 978-981-16-9367-0 ISBN 978-981-16-9368-7 (eBook)
https://doi.org/10.1007/978-981-16-9368-7

Jointly published with Tsinghua University Press
The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the print book from: Tsinghua University Press.

© Springer Nature Singapore Pte Ltd. 2022
This work is subject to copyright. All rights are reserved by the Publishers, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publishers, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publishers nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publishers remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Background

The data center aims to interconnect the hardware resources, such as large-scale servers and network facilities, with dedicated topologies. Thus, data center acts as a massive resource pool, including computing, storage, network and other resources, to provide elastic services for various applications.

Data center networking (DCN) plays an essential role for the design of data centers. DCN aims to connect all switches and servers together using an outstanding topology, so as to achieve better network performance. Many applications hosted by data centers cause intensive data interaction across multiple servers. DCN becomes the bottleneck of data centers, and affects user's experience directly. Therefore, it is urgent to bring more innovations into the field of DCN, so as to promote the development of data centers and its applications.

The research about DCN has become a highlight field in the academia as well as the industry communities. There still exist a series of fundamental challenges for speeding the development of the data centers. Specifically, this book focuses on two challenges, including the topology design and the collaborative management of traffic inside each data center.

This book deeply discusses the design and optimization methods of several novel network topologies for data centers, so as to realize the high bandwidth, high fault tolerance and high scalability. Furthermore, this book reports the cooperative transmission mechanisms of several correlated traffic patterns in a data center, which can significantly reduce the bandwidth consumption. The majority content in the book come from our recent papers published in journals or conferences, and systematically report representative research results in the field of data center network.

Content Organization

The book consists of ten chapters, which can be divided into three parts. Part I presents the background and overview of data centers and then discusses the state-of-the-art network topologies of data centers, in Chaps. 1 and 2, respectively.

Chapter 1 reports the evolution trend of data centers, illustrates the application fields of data centers, and summarizes the important challenges of data center networking.

Chapter 2 proposes a new taxonomy that divides current data center topologies into five categories, including the switch-centric, server-centric, modular, random, and wireless topologies.

Part II introduces our four network topologies of data centers in Chaps. 3–6. They are the server-centric network topology HCN, modular network topology DCube, hybrid network topology R3, and wireless network topology VLCcube.

Chapter 3 introduces two server-centric network topologies, HCN and BCN, based on the compound graph theory. They are constructed using homogeneous servers, each of which just has two NICs, and commodity switches.

Chapter 4 introduces a family of intra-module network topologies, called DCube. Each DCube interconnects a large number of dual-port servers and low-end switches. Many DCube modules can further form a new modular data center.

Chapter 5 introduces a general design methodology for DCN topologies using the compound graph theory, and discusses a hybrid network topology R3, which is the compound graph of a structured topology and a random topology.

Chapter 6 proposes a hybrid network topology, VLCcube, by introducing extra visible light communication (VLC) links into data centers. Specifically, VLCcube augments Fat-Tree, a representative DCN in production data centers.

Part III focuses on the cooperative transmission of correlated flows inside a data center. Managing those correlated flows is very essential for efficiently utilizing the network resource and speeding the data applications.

Chapter 7 introduces the cooperative transmission management problem of Incast. The basic idea is to perform the in-network data aggregation, during the transmission phase as early as possible rather than just at the receiver side.

Chapter 8 considers how to realize the in-network aggregation of correlated Shuffle transfer, so that the bandwidth consumption can be considerably reduced. This chapter also introduces the scalable traffic forwarding mode based on Bloom filters.

Chapter 9 makes the first step towards the study of in-network aggregation of uncertain Incast transfer.

Chapter 10 introduces the cooperative traffic management of multicast. This chapter proposes the minimum cost forest (MCF) model for such a multicast transfer with uncertain senders.

The author's research work has been partially supported by the National Natural Science Foundation of China under Grant No. U19B2024.

Changsha, China Deke Guo
September 2021

Contents

Part I Basic Knowledge

1 Introduction of Data Center 3
1.1 Evolution of Data Centers 3
1.1.1 Basic Concept and Intuitive Classification 3
1.1.2 Demands for Data Centers from Cloud Computing 6
1.1.3 Demands for Data Centers from Big Data Applications 8
1.1.4 The Development of New Generation Data Centers 9
1.2 Fundamental Services Offered by Data Centers 13
1.2.1 Storage Service 13
1.2.2 Computing Services 14
1.2.3 Big Data Applications 15
1.3 Challenges for Data Center Networks 17
1.3.1 Customization of Network Functionality 18
1.3.2 High Scalability of Data Center Networks 19
1.3.3 Efficient Multiplexing of Network Resources 20
1.3.4 Network Virtualization of Data Centers 21
1.3.5 Cooperative Transmission of Correlated Traffic 22
References 23

2 State-of-the-Art DCN Topologies 25
2.1 Introduction 25
2.2 Switch-Centric DCN Topologies 27
2.2.1 Tree-Like Topologies 28
2.2.2 Flat Topologies 30
2.2.3 Optical Switching Topologies 33
2.3 Server-Centric DCN Topologies 36
2.3.1 Compound Graph-Based DCN Topologies 36
2.3.2 Non-Compound Graph-Based DCN Topologies 39

2.4 Modular DCN Topologies 41
2.4.1 Intra-module Network Topology 41
2.4.2 Inter-module Network Topology 42
2.5 Random DCN Topologies 43
2.5.1 Small-World Network-Based DCN Topology 45
2.5.2 Random Regular Graph-Based DCN Topology 45
2.5.3 Scale-Free Network-Based DCN Topology 46
2.6 Wireless DCN Topologies 48
2.6.1 60 GHz-Based Hybrid DCN Topology 48
2.6.2 60 GHz-Based Wireless DCN Topology 49
2.6.3 FSO-Based DCN Topology 50
2.6.4 VLC-Based DCN Topology 50
2.7 Evolution and Future Trends of DCN Topology Design 51
2.7.1 Evolution of DCN Topology Design 51
2.7.2 Future Trends of DCN Topology Design 52
References 53

Part II Novel Data Center Network Structures

3 HCN: A Server-Centric Network Topology for Data Centers 59
3.1 Introduction 59
3.2 The Design of HCN Topology 61
3.2.1 Basic Theories of the Compound Graph 61
3.2.2 The Construction Methodology of HCN 63
3.3 The Design of BCN Topology 65
3.3.1 Description of BCN Topology 65
3.3.2 The Construction Methodology of BCN 67
3.4 Routing Mechanism of BCN 71
3.4.1 Single-path for the Unicast Traffic 71
3.4.2 Multi-paths for Unicast Traffic 73
3.4.3 Fault-Tolerant Routing in BCN 75
3.5 Performance Evaluation 78
3.5.1 Network Order 78
3.5.2 Low Network Diameter and Server Degree 80
3.5.3 Connectivity and Path Diversity 80
3.5.4 Evaluation of the Path Length 82
3.6 Discussion 84
3.6.1 Extension to More Server Ports 84
3.6.2 Locality-Aware Task Placement 84
3.6.3 Impact of the Server Routing 84
References 85

4 DCube: A Family of Network Topologies for Containerized Data Centers 87
4.1 Introduction 87
4.2 The DCube Topology 89
4.2.1 Design Idea of DCube 89
4.2.2 H-DCube 90
4.2.3 M-DCube 91
4.3 Single-Path Routing for the Unicast Traffic 92
4.3.1 Single-Path Routing in H-DCube 93
4.3.2 Single-Path Routing in M-DCube 94
4.4 Multi-paths Routing for the Unicast and Multicast Traffic 97
4.4.1 Multi-path Routing for the Unicast Traffic in H-DCube 98
4.4.2 Multi-path Routing for the Unicast Traffic in M-DCube 99
4.4.3 Speedup for the Multicast Traffic 101
4.5 Analysis and Evaluation 102
4.5.1 Speedup for the Unicast and Multicast Traffic 103
4.5.2 Aggregate Bottleneck Throughput 104
4.5.3 Qualification of Cost and Cabling Complexity 105
4.5.4 Summary 108
4.6 Discussions 108
4.6.1 Locality-Aware Task Placement 108
4.6.2 Extension to More Server Interfaces 108
4.6.3 Impact of Server Routing 109
References 109

5 R3: A Hybrid Network Topology for Data Centers 111
5.1 Introduction 111
5.2 The Design Methodology of Hybrid Topologies 112
5.2.1 Overview of Network Topologies 113
5.2.2 R3: A Hybrid Topology Based on the Compound Graph Theory 114
5.2.3 Deployment Strategy for Data Centers with a Hybrid Topology 117
5.3 Efficient Routing Methods of R3 117
5.3.1 Edge Coloring Based Identifier Allocation 118
5.3.2 Identifier-Based Routing Algorithm 119
5.4 Topology Optimization 121
5.4.1 Impact Factors of the R3 Topology 122
5.4.2 Optimization Strategy of Topologies 122
5.5 Incremental Expansion of R3 124
5.5.1 Expansion within an Existing Random Cluster 125
5.5.2 Expansion by Adding an Extra Random Cluster 125

5.6 Performance Evaluation 127
5.6.1 The Routing Flexibility 127
5.6.2 The Cabling Cost 128
5.6.3 The Network Performance 129
5.7 Discussion 130
References 131

6 VLCcube: A Network Topology for VLC Enabled Wireless Data Centers 133
6.1 Introduction 133
6.1.1 Motivation 133
6.1.2 Related Work 135
6.2 The Design of VLCcube 136
6.2.1 Feasibility of Introducing VLC Links into DCNs 136
6.2.2 The Interference Among Transceivers 137
6.2.3 Topology Design of VLCcube 138
6.3 Routing and Congestion Aware Flow Scheduling in VLCcube 142
6.3.1 Hybrid Routing Scheme in VLCcube 142
6.3.2 Problem Formulation of Flow Scheduling 143
6.3.3 Scheduling the Batched Flows 144
6.3.4 Online Scheduling of Flows 147
6.4 Performance Evaluation 149
6.4.1 Settings of Evaluations 149
6.4.2 Topological Properties of VLCcube 150
6.4.3 Network Performance Under Trace Flows 151
6.4.4 Network Performance Under Stride-2k Flows 152
6.4.5 Network Performance Under Random Flows 153
6.4.6 Impact of Congestion Aware Flow Scheduling 154
6.5 Discussion 155
References 156

Part III Traffic Cooperation Management in Data Centers

7 Collaborative Management of Correlated Incast Transfer 161
7.1 Introduction 161
7.2 In-Network Aggregation of an Incast Transfer 163
7.2.1 Feasibility of the Inter-flow Data Aggregation 163
7.2.2 Minimal Incast Tree 165
7.2.3 Inter-flow Data Aggregation on an Incast Tree 166
7.3 Efficient Building Method of an Incast Tree 166
7.3.1 Online Construction of an Incast Tree 168
7.3.2 Construction of the Minimal Incast Tree 171
7.3.3 Dynamical Behaviors of Senders 173
7.3.4 Dynamical Behaviors of the Receiver 173

7.4 Discussion 174
7.4.1 Extension to General Incast Transfers 174
7.4.2 Extension to Other Network Topologies 175
7.4.3 Impact on the Job Execution Time 175
7.5 Performance Evaluation 176
7.5.1 The Prototype Implementation 176
7.5.2 Impact of the Data Center Size 178
7.5.3 Impact of the Incast Transfer Size 179
7.5.4 Impact of the Aggregation Ratio 180
7.5.5 Impact of the Distribution of Incast Members 182
References 183

8 Collaborative Management of Correlated Shuffle Transfer 185
8.1 Introduction 185
8.2 In-Network Aggregation of Shuffle Transfers 186
8.2.1 Problem Statement 186
8.2.2 Construction of an Incast Aggregation Tree 188
8.2.3 Construction of a Shuffle Aggregation Subgraph 192
8.2.4 The Fault-Tolerance of Shuffle Aggregation Subgraph 195
8.3 Scalable Forwarding Schemes for Performing In-Network Aggregation 196
8.3.1 General Forwarding Scheme 196
8.3.2 In-Switch Bloom Filter Based Forwarding Scheme 197
8.3.3 In-Packet Bloom Filter Based Forwarding Scheme 198
8.4 Performance Evaluation 200
8.4.1 The Prototype 200
8.4.2 Impact of the Data Center Size 201
8.4.3 Impact of the Shuffle Transfer Size 203
8.4.4 Impact of the Aggregation Ratio 204
8.4.5 The Size of Bloom Filter in Each Packet 206
References 207

9 Collaborative Management of Uncertain Incast Transfer 209
9.1 Introduction 209
9.2 Overview of Aggregating Uncertain Incast Transfer 211
9.2.1 Problem Statement of Uncertain Incast Transfer 211
9.2.2 Aggregating a Deterministic Incast Transfer 212
9.2.3 Aggregating an Uncertain Incast Transfer 213
9.3 Aggregation Tree Building Method for Uncertain Incast Transfer 214
9.3.1 The Diversity of In-Network Aggregation Gain 215
9.3.2 Initializing Senders for an Uncertain Incast Transfer 215
9.3.3 Aggregation Tree Building Method for Uncertain Incast 220

9.4 Performance Evaluation 223
9.4.1 Evaluation Methodology and Scenarios 223
9.4.2 Impact of the Incast Transfer Size 224
9.4.3 Impact of the Data Center Size 226
9.4.4 Impact of the Aggregation Ratio 227
9.4.5 Impact of Distributions of Incast Members 228
9.4.6 Impact of the Receiver Diversity 230
References 231

10 Collaborative Management of Correlated Multicast Transfer 233
10.1 Introduction 233
10.2 Related Work 235
10.3 Problem Statement of Uncertain Multicast 236
10.3.1 Observations 236
10.3.2 Problem Statement 238
10.3.3 Mixed Integer Linear Programming 239
10.4 Efficient Building Methods of MCF 241
10.4.1 Primary Approximation Method 241
10.4.2 Enhanced Approximation Method 244
10.5 Performance Evaluation 247
10.5.1 Implementation of the Uncertain Multicast in SDN Testbed 247
10.5.2 Evaluation Based on Small-Scale Experiments 248
10.5.3 Evaluation Based on Large-Scale Simulations 251
References 255