

SpringerBriefs in Computer Science

Series Editors

Stan Zdonik

Peng Ning

Shashi Shekhar

Jonathan Katz

Xindong Wu

Lakhmi C. Jain

David Padua

Xuemin Shen

Borko Furht

V.S. Subrahmanian

Martial Hebert

Katsushi Ikeuchi

Bruno Siciliano

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

© The Author(s) 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, 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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

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
Hong Kong, P. R. China
Charlottesville, VA, USA
Hong Kong, P. R. China
Hong Kong, P. R. China

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

Contents

| | | |
|----------|---|----|
| 1 | Introduction | 1 |
| 1.1 | Introduction | 1 |
| 1.2 | Data Center Applications | 2 |
| 1.3 | Data Center Network Requirements | 3 |
| 1.4 | Summary | 4 |
| | References | 4 |
| 2 | Data Center Network Topologies: Current State-of-the-Art | 7 |
| 2.1 | Typical Data Center Network Topology | 7 |
| 2.1.1 | Tree-Based Topology | 8 |
| 2.1.2 | Clos Network | 9 |
| 2.2 | Data Center Network Technologies | 10 |
| 2.3 | Problems and Issues with Current Approaches | 12 |
| 2.4 | Summary | 13 |
| | References | 13 |
| 3 | Data Center Network Topologies: Research Proposals | 15 |
| 3.1 | Classification of Topologies | 15 |
| 3.2 | Fixed Tree-Based Topologies I | 16 |
| 3.2.1 | Basic Tree | 16 |
| 3.2.2 | Fat-Tree | 17 |
| 3.3 | Fixed Recursive Topologies II | 18 |
| 3.3.1 | DCell | 18 |
| 3.3.2 | BCube | 19 |
| 3.4 | Flexible Topologies | 20 |
| 3.4.1 | c-Through | 21 |
| 3.4.2 | Helios | 22 |
| 3.4.3 | OSA | 22 |
| 3.5 | Comparison of Topologies | 23 |
| 3.5.1 | Comparison of Scale | 23 |
| 3.5.2 | Comparison of Performance | 25 |

| | | |
|----------|--|-----------|
| 3.5.3 | Performance Evaluation Using Simulation | 26 |
| 3.5.4 | Hardware Redundancy of Data Center Network Topologies .. | 27 |
| 3.6 | Potential New Topologies | 29 |
| 3.7 | Summary | 30 |
| | References | 30 |
| 4 | Routing Techniques | 33 |
| 4.1 | Introduction | 33 |
| 4.2 | Fixed Tree-Based Topologies I | 34 |
| 4.2.1 | Fat-Tree Architecture of Al-Fares et al. [1]..... | 34 |
| 4.2.2 | PortLand and Hedera: Layer-2 Network Based on a Tree Topology | 36 |
| 4.2.3 | VL2 | 37 |
| 4.3 | Fixed Recursive Topologies II | 39 |
| 4.3.1 | DCell | 39 |
| 4.3.2 | BCube | 40 |
| 4.4 | Summary | 41 |
| 4.4.1 | Addressing | 41 |
| 4.4.2 | Centralized and Distributed Routing | 42 |
| | References | 43 |
| 5 | Performance Enhancement | 45 |
| 5.1 | Introduction | 45 |
| 5.2 | Centralized Flow Scheduling in Fat-Tree Architecture of Al-Fares et al. [1] | 45 |
| 5.3 | Hedera's Flow Scheduling for Fat-Tree | 46 |
| 5.4 | Random Traffic Spreading of VL2 | 47 |
| 5.5 | BCube Source Routing | 47 |
| 5.6 | Traffic De-multiplexing in c-Through..... | 48 |
| 5.7 | Summary of Performance Enhancement Schemes | 48 |
| 5.7.1 | Use of Multiple Paths for Performance Enhancement | 49 |
| 5.7.2 | Flow Scheduling | 49 |
| | References | 50 |
| 6 | Fault-Tolerant Routing | 51 |
| 6.1 | Introduction | 51 |
| 6.2 | Failure Models | 51 |
| 6.2.1 | Failure Type | 51 |
| 6.2.2 | Failure Region | 52 |
| 6.2.3 | Failure Neighborhood | 52 |
| 6.2.4 | Failure Mode | 52 |
| 6.2.5 | Failure Time..... | 53 |
| 6.2.6 | Taxonomy of Faults..... | 53 |
| 6.2.7 | Evaluation of Fault-Tolerance Characteristics..... | 54 |
| 6.3 | Link Failure Response in Fat-Tree | 59 |
| 6.4 | Fault-Tolerant Routing in PortLand and Hedera..... | 60 |

| | |
|--|-----------|
| 6.5 DCell Fault-Tolerant Routing | 61 |
| 6.6 Fault-Tolerant Routing in BCube..... | 62 |
| 6.7 Summary of Fault-Tolerant Routing Algorithms | 62 |
| References | 64 |
| 7 Conclusions | 65 |
| Index | 67 |

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 |