Casimer DeCusatis

HANDBOOK OF FIBER OPTIC DATA COMMUNICATION

A PRACTICAL GUIDE TO OPTICAL NETWORKING



Fourth Edition

Handbook of Fiber Optic Data Communication

A Practical Guide to Optical Networking

FOURTH EDITION

EDITED BY

Casimer DeCusatis

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Table of Contents

Cover image

Title page

Copyright

Preface to the Fourth Edition

A Historical View of Fiber Data Communications

Part I: Technology Building Blocks

Chapter 1. Transforming the Data Center Network

- 1.1 Properties Of A Network Reference Architecture
- 1.2 Next-Generation Data Center Networks
- References

Chapter 2. Transceivers, Packaging, and Photonic Integration

- 2.1 Introduction
- 2.2 Transceivers
- 2.3 System-Level Integration
- 2.4 Future Trends In Electro-Optical Packaging
- References

Chapter 3. Plastic Optical Fibers for Data Communications

- 3.1 Introduction
- 3.2 A POF Taxonomy
- 3.3 PMMA-SI-POF

3.4 PF-GI-POF

References

Chapter 4. Optical Link Budgets

4.1 Fundamentals Of Fiber Optic Communication Links

4.2 Basic Link Budget Analysis For Network Designers

4.3 Figures Of Merit

4.4 Advanced Link Budget Analysis: Optical Power Penalties

4.5 Link Budgets With Optical Amplification

References

Additional References

Optical Link Budget Models And Specifications Available Online

References On Reflection Noise

Case Study. Deploying Systems Network Architecture (SNA) in IP-Based Environments: The Mainframe Network as a TCP/IP Server

Introduction

Overview

Factors Contributing To The Continued Use Of SNA

How Do We Modernize SNA To SNA Over IP?

SNA Over IP Networks

Solutions

Conclusion

Chapter 5. Optical Wavelength-Division Multiplexing for Data Communication Networks

5.1 Basics Of Wavelength-Division Multiplexing

5.2 WDM Systems And Networks

5.3 Optical Transport Network

References

Case Study. A More Reliable, Easier to Manage TS7700 Grid Network

Introduction

TS7700 Grid Network

High Availability

The XYZ Company's TS7700 Integrated Grid

Summary

Chapter 6. Passive Optical Networks (PONs)

6.1 Passive Optical Networks

6.2 Relevant PON Variants And Standards

6.3 PON Deployment References

References

Part II: Protocols and Industry Standards

Chapter 7. Manufacturing Environmental Laws, Directives, and Challenges

7.1 Introduction

7.2 Worldwide Environmental Directives, Laws, And Regulations

7.3 Restriction Of Hazardous Substances

References

Case Study. Energy Efficient Networking for Data Centers

Chapter 8. Fibre Channel Standard

8.1 Introduction

8.2 Fibre Channel Overview And Basic Structure

8.3 Storage Area Networks

8.4 How Fibre Channel Leverages Optical Data Communications

8.5 Summary

Web Resources And References

Chapter 9. Lossless Ethernet for the Data Center

9.1 Introduction To Classic Ethernet

9.2 Ethernet Physical Layer

9.3 Gigabit Ethernet

9.4 Lossless Ethernet

References

Case Study. FCoE Delivers a Single Network for Simplicity and Convergence

Chapter 10. Metro and Carrier Class Networks: Carrier Ethernet and OTN

10.1 Evolution: The Roots Of Modern Networks

10.2 Ethernet Virtual LANs

10.3 Network Evolution Using Carrier Class Ethernet And OTN

10.4 Carrier Ethernet: Standardized Services, Scalable, Reliable, Quality Of Service, And Service Management

10.5 Optical Transport Networking: A Transparent Optical Protocol

10.6 The Packet-Optical Network

Resources

Chapter 11. InfiniBand, iWARP, and RoCE

11.1 Introduction

11.2 InfiniBand Architecture

11.3 IB Network

11.4 Communication Mechanisms

11.5 Layered Architecture

11.6 RDMA Over Converged Ethernet (RoCE)

11.7 8 IWARP

References

Part III: Network Architectures and Applications

Chapter 12. Disaster Recovery and Data Networking

12.1 Introduction

12.2 Data Consistency: The BASE-ACID Model

12.3 Examples Of BASE-ACID Methodology

12.4 IBM Parallel Sysplex And GDPS

12.5 Time Synchronization In Disaster Recovery Solutions

12.6 Cloud Backup And Recovery

12.7 Container Data Centers

References

Case Study. Using Business Process Modeling Notation and Agile Test-Driven Design Methodology

for Business Continuity Planning

Chapter 13. Network Architectures and Overlay Networks

13.1 STP And MC-LAG

13.2 Layer 3 Versus Layer 2 Designs For Cloud Computing

13.3 The Open Data Center Interoperable Network

13.4 Cisco FabricPath

13.5 Juniper Qfabric

13.6 Virtual Network Overlays

Acknowledgments

References

Chapter 14. Networking for Integrated Systems

14.1 IBM PureSystems

14.2 Cisco Virtual Computing Environment And Unified Computing System Solutions

14.3 Sun/Oracle Exalogic

14.4 Hewlett-Packard Matrix

14.5 Hitachi Data Systems Unified Compute Platform

14.6 Huawei FusionCube

References

Case Study. The Network That Won *Jeopardy!*—Watson Supercomputing

Case Study. NYSE Euronext Data Center Consolidation Project

Goal

Design Considerations

Design

Result

Chapter 15. Cloud Computing Data Center Networking

15.1 Introduction

15.2 Cloud Characteristics

15.3 Cloud Facilities

15.4 Cloud Architecture

15.5 Cloud Computing Data Center Network Trends

15.6 Conclusion

Acknowledgments

References

Chapter 16. Hypervisors, Virtualization, and Networking

16.1 Virtualization

16.2 PowerVM

16.3 VMware

16.4 Xen

16.5 Kernel-Based Virtual Machine

16.6 Z/VM

16.7 Virtual Switches

References

Case Study. Open Standards for Cloud Networking

Chapter 17. Software-Defined Networking and OpenFlow

17.1 Introduction

17.2 SDN Architecture I: Overview

17.3 SDN Architecture II: Data Plane

17.4 SDN Architecture III: Control Plane

17.5 Example Application: WAN TE

17.6 SDN In Optical Networks

17.7 Conclusion

References

Chapter 18. Emerging Technology for Fiber Optic Data Communication

18.1 Introduction

18.2 Architecture Of All-Optical Networks

18.3 Tunable Transmitter

18.4 Tunable Receiver

18.5 Optical Amplifier

18.6 Wavelength Multiplexer/Demultiplexer

18.7 Wavelength Router

18.8 Wavelength Converter

18.9 Summary

References

Appendix A. Measurement Conversion Tables

Appendix B. Physical Constants

Appendix C. The 7-Layer OSI Model

Appendix D. Network Standards and Data Rates

Organization Of Major Industry Standards

SONET/SDH

Ethernet

10G Ethernet WAN PHY

Ethernet First Mile Standards

Appendix E. Fiber Optic Fundamentals and Laser Safety

References

Index

Copyright

Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

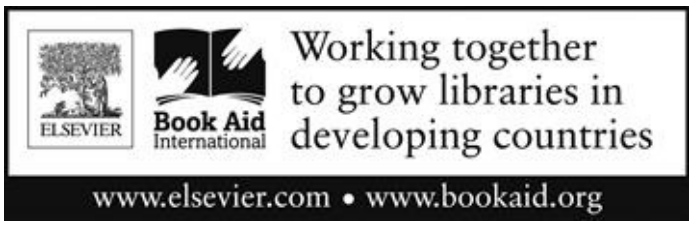
Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-401673-6

Typeset by MPS Limited, Chennai, India www.adi-mps.com

Printed and bound in United States of America

13 14 15 16 10 9 8 7 6 5 4 3 2 1



Preface to the Fourth Edition

In previous editions of this Handbook, I have tried to summarize the importance of optical networkin to the data communication field with the following bit of poetry:

> ***SONET[1]on the Lambdas[2]***
>
> *When I consider how the light is bent*
>
> *By fibers glassy in this Web World Wide,*
>
> *Tera- and Peta-, the bits fly by*
>
> *Are they from Snell and Maxwell sent*
>
> *Or through more base physics, which the Maker presents*
>
> *(lambdas of God?) or might He come to chide*
>
> *"Doth God require more bandwidth, light denied?"*
>
> *Consultants may ask; but Engineers to prevent*
>
> *that murmur, soon reply "The Fortune e-500 do not need*
>
> *mere light alone, nor its interconnect; who requests*
>
> *this data, if not clients surfing the Web?" Their state*
>
> *is processing, a billion MIPS or CPU cycles at giga-speed.*
>
> *Without fiber optic links that never rest,*
>
> *The servers also only stand and wait.*
>
> C. DeCusatis, with sincere apologies to Milto

Of course, I am certainly not the only network engineer to be inspired by classic literature. Those you who prefer Joyce Kilmer (http://www.poetry-archive.com/k/trees.html) to Milton might enjoy t following well-known piece, crafted by Radia Perlman while she was inventing the Spanning Tr Protocol (https://www.cs.washington.edu/education/courses/461/08wi/lectures/p44-perlman.pd which you can also hear set to music online with Ms. Perlman on piano and her daughter on voca

(http://www.youtube.com/watch?v=iE_AbM8ZykI).

Algorhyme

I think that I shall never see

A graph more lovely than a tree.

A tree whose crucial property

Is loop-free connectivity.

A tree that must be sure to span

So packets can reach every LAN.

First, the root must be selected.

By ID, it is elected.

Least-cost paths from root are traced.

In the tree, these paths are placed.

A mesh is made by folks like me,

Then bridges find a spanning tree.

The radical changes currently taking place in data networking are perhaps reflected in another poe by Radia's son, Ray Perlner, who wrote this tribute to the TRILL protocol:

Algorhyme V2

I hope that we shall one day see

A graph more lovely than a tree.

A graph to boost efficiency

While still configuration-free.

A network where RBridges can

Route packets to their target LAN.

The paths they find, to our elation,

Are least cost paths to destination!

With packet hop counts we now see,

The network need not be loop-free!

RBridges work transparently,

Without a common spanning tree.

The potential replacement of spanning tree protocols (whether by TRILL or other options) is on one of the many changes in this field that led us to believe the time was right to once again update th Handbook. This is a very interesting time to be a network engineer, as the field experiences perhaps i greatest upheavals since Metcalf first introduced the basic principles of Ethernet. Since the fir edition of this book was published over 10 years ago, I have tried to continually incorporate feedbac and comments from readers to improve the book and ensure that it continues to provide a singl indispensable reference for the optical data communication field. You will still find a single referenc for all the leading data center networking protocols and technologies, as well as many new chapte dealing with issues that did not exist when the last edition was published (including the TRIL protocol mentioned earlier). A series of all new case studies discuss real-world applications of th technology. It has become apparent that network virtualization is the next big frontier, and we are ju beginning to see the full potential of data center networking—application aware, distan independent, infinitely scalable, user-centric networks that catalyze real-time global computin advanced streaming multimedia, distance learning, telemedicine, and a host of other applications. W hope that those who build and use these networks will benefit in some measure from this book.

An undertaking such as this would not be possible without the concerted efforts of man contributing authors and a supportive staff at the publisher, to all of whom I extend my deepe gratitude. As always, this book is dedicated to my mother and father, who first helped me see th wonder in the world; to the memory of my godmother Isabel; and to my wife, Carolyn, and m daughters Anne and Rebecca, without whom this work would not have been possible.

Dr. Casimer DeCusatis

Editor, Poughkeepsie, New York

December 2012

[1]Synchronous Optical Network.

[2]The Greek symbol “lambda” or λ is commonly used in reference to an optical wavelength.

[3]The original author of the classic sonnet “On his blindness.”

A Historical View of Fiber Data Communications*

When I wrote the first edition of *Understanding Fiber Optics* in 1987, fiber had recently become th backbone of the North American telecommunications network, where it transmitted 417 Mb/s throug single-mode fiber. Developers were working on the next generation to transmit 1.7 Gb/s on land. Th laying of the first transatlantic fiber cable, TAT-8, was a year away. Local area networks did not nee fiber at their 1987 data rates of 1–10 Mb/s, but developers had hopes for the coming generation 100-Mb/s transmission. Fiber-to-the-home systems had been demonstrated to groups of 150 homes Japan and Canada, but were far too costly for general installations.

Today state-of-the-art backbone networks carry 100 Gb/s on each wavelength-division multiplexe (WDM) optical channel in a single fiber. Optical amplifiers have replaced the electro-optic repeate used in first-generation backbone systems. Transatlantic submarine cable capacity increased s rapidly that when TAT-8 suffered an undersea failure in 2002, it was not worth repairing. Fiber cam to my home in suburban Boston several years ago, and now provides me with Internet access 25 Mb/s, four orders of magnitude faster than the 1200 baud my first modem pumped through an age pair of copper wires in 1987.

It has been a remarkable run for fiber, with the growth of transmission capacity rivaling the rise computer power. Indeed, fiber communications and computer chips complement each oth wonderfully; we need both to make the Internet hum with information and to bring the virtual world our fingertips.

The new technology described in this volume reflects the global importance of da communications and information processing. The Internet has become a part of our life in th developed world, and new mobile devices are bringing connectivity to the developing world. Althoug fibers are fixed in place, they provide vital links in a mobile world. The key allure of storing data "the cloud" is that users can access that data from anywhere. Often that is through mobile device with wireless connections to the network that make the "cloud" metaphor seem appropriate. Yet th big pipes carrying data into the cloud are fiber cables in the global backbone network. And fibers a the data plumbing in the server farms and storage area networks that are the physical reality of th cloud.

Reaching these high data capacities has required continuing innovation in fiber and optic technology, from the nuts and bolts of packaging, light sources, detectors, and fibers to new concep for networking and data transmission. Innovation seeks cost-effectiveness as well as hig performance; VCSELs have become standard light sources, and plastic fibers offer the potential low cost for short links.

Interestingly, some of the latest and greatest high-performance innovations are revivals technology earlier abandoned as impractical. Old-timers may remember that coherent communicatio systems, the optical counterpart of heterodyne radio, were supposed to be the next great thing back

1987 because they promised higher-speed transmission over longer distances. But within a few year coherent transmission was blown away by the combination of fiber amplifiers and WDM, whic multiplied system data rates and distance spans.

Now coherent transmission is back, powering commercial systems transmitting 100 Gb/s line rat on a single optical channel. Developers gave coherent transmission a new chance when they needed squeeze 100 Gb/s data streams into the 50 GHz bands assigned to WDM channels. Coherent systen can detect modulation of light phase and polarization, squeezing more data into a limited bandwidth the same way as cellular phone transmitters. Coherent transmission also allowed digital processing compensate for signal impairments such as chromatic dispersion, avoiding the need for optic dispersion compensation. The next challenges will be to push to higher data rates to handle th inevitable growth in traffic volume and to continue refining the network architecture and application

This volume documents the progress so far and looks to further developments.

Jeff Hecht

Auburndale, Massachusetts

September 2012

*By Jeff Hecht, author of *City of Light: The Story of Fiber Optics* and *Understanding Fiber Optics*.

PART I

Technology Building Blocks

OUTLINE

Chapter 1 Transforming the Data Center Network

Chapter 2 Transceivers, Packaging, and Photonic Integration

Chapter 3 Plastic Optical Fibers for Data Communications

Chapter 4 Optical Link Budgets

Case Study Deploying Systems Network Architecture (SNA) in IP-Based Environments

Chapter 5 Optical Wavelength-Division Multiplexing for Data Communication Networks

Case Study A More Reliable, Easier to Manage TS7700 Grid Network

Chapter 6 Passive Optical Networks (PONs)

CHAPTER 1

Transforming the Data Center Network

Casimer DeCusatis, IBM Corporation, 2455 South Road, Poughkeepsie, NY

In recent years, there have been many fundamental changes in the architecture of modern data center New applications have emerged, including cloud computing, big data analytics, real-time stoc trading, and more. Workloads have evolved from a predominantly static environment into one th changes over time in response to user demands, often as part of a highly virtualized, multitenant da center. In response to these new requirements, data center networks have also undergone significa change. Conventional network architectures, which use Ethernet access, aggregation, and core tie with a separate storage area network, are not well suited to modern data center traffic patterns. Th chapter reviews the evolution from conventional network architectures into designs better suited dynamic, distributed workloads. This includes flattening the network, converging Ethernet wi storage and other protocols, and virtualizing and scaling the network. Effects of oversubscriptio latency, higher data rates, availability, reliability, energy efficiency, and network security will b discussed.

Keywords

network; fabric; cloud; virtualize; Ethernet; SAN

In recent years, there have been many fundamental and profound changes in the architecture modern data centers, which host the computational power, storage, networking, and applications th form the basis of any modern business [1–7]. New applications have emerged, including clou computing, big data analytics, real-time stock trading, and more. Workloads have evolved from predominantly static environment into one which changes over time in response to user demand often as part of a highly virtualized, multitenant data center. In response to these new requirement data center hardware and software have also undergone significant changes; perhaps nowhere is th more evident than in the data center network.

In order to better appreciate these changes, we first consider the traditional data center architectu and compute model, as shown in Figure 1.1. While it is difficult to define a “typical” data cente Figure 1.1 provides an overview illustrating some of the key features deployed in many enterpri class networks and Fortune 1000 companies today, assuming a large number of rack or blade serve using x86-based processors. Figure 1.1 is not intended to be all inclusive, since there are mar variations on data center designs. For example, in some applications such as high-performan computing or supercomputing, system performance is the dominant overriding design considerati Mainframes or large enterprise compute environments have historically used more virtualization ar based their designs on continuous availability, with very high levels of reliability and serviceabilit Telecommunication networks are also changing, and the networks that interconnect multiple da centers are taking on very different properties from traditional approaches such as frame relay. Th

includes the introduction of service-aware networks, fiber to the home or small office, and passiv optical networks. A new class of ultralow-latency applications has emerged with the advent of rea time financial transactions and related areas such as telemedicine. Further, the past few years hav seen the rise of warehouse-scale data centers, which serve large cloud computing applications fro Google, Facebook, Amazon, and similar companies; these applications may use custom-designe servers and switches. While many of these data centers have not publicly disclosed details of the designs at this time, it is a good assumption that when the data center grows large enough to consun electrical power equivalent to a small city, energy-efficient design of servers and data cent heating/cooling become major considerations. We will consider these and other applications later i this book; for now, we will concentrate on the generic properties of the data center network shown i Figure 1.1. Although data centers employ a mixture of different protocols, including InfiniBand, Fibr Channel, FICON, and more, a large portion of the network infrastructure is based on some variation Ethernet.

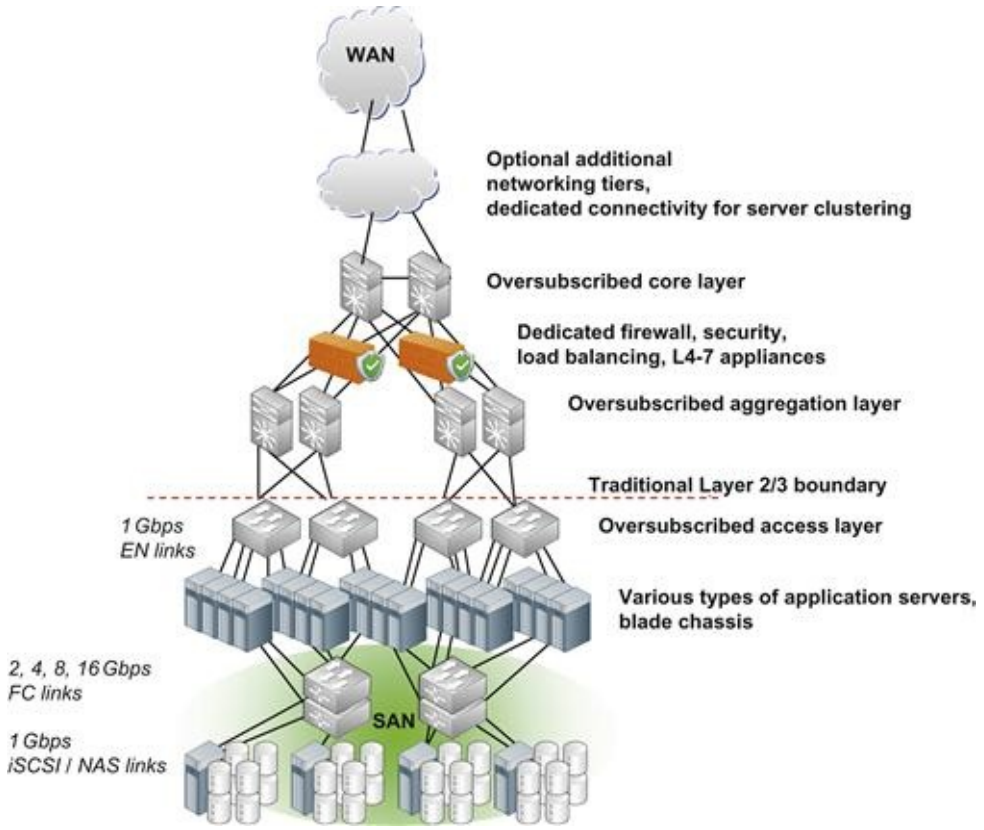


FIGURE 1.1 Design of a conventional multitier data center network.

Historically, as described in the early work from Metcalf [8], Ethernet was first used to interconne "stations" (dumb terminals) through repeaters and hubs on a shared data bus. Many stations wou listen to a common data link at the same time and make a copy of all frames they heard; fram intended for that station would be kept, while others would be discarded. When a station needed transmit data, it would first have to check that no other station was transmitting at the same time; on the network was available, a new transmission could begin. Since the frames were broadcast to a stations on the network, the transmitting station would simultaneously listen to the network for th same data it was sending. If a data frame was transmitted onto an available link and then heard by th transmitting station, it was assumed that the frame had been sent to its destination; no furth checking was done to insure that the message actually arrived correctly. If two stations accidentall

began transmitting at the same time, the packets would collide on the network; the station would the cease transmission, wait for a random time interval (the backoff interval), and then attempt retransmit the frame.

Over time, Ethernet evolved to support switches, routers, and multiple link segments, with high level protocols such as TCP/IP to aid in the recovery of dropped or lost packets. Spanning tr protocol (STP) was developed to prevent network loops by blocking some traffic paths, at the expen of network bandwidth. However, this does not change the fact that Ethernet was inherently a "be effort" network, in which a certain amount of lost or out-of-order data packets are expected by desig Retransmission of dropped or misordered packets was a fundamental assumption in the Ethern protocol, which lacked guaranteed delivery mechanisms and credit-based flow control concepts suc as those designed into Fibre Channel and InfiniBand protocols.

Conventional Ethernet data center networks [9] are characterized by access, aggregation, service and core layers, which could have three, four, or more tiers of switching. Data traffic flows from th bottom tier up through successive tiers as required, and then back down to the bottom tier, providin connectivity between servers. Since the basic switch technology at each tier is the same, the TCP/I stack is usually processed multiple times at each successive tier of the network. To reduce cost an promote scaling, oversubscription is typically used for all tiers of the network. Layer 2 and 3 functio are separated within the access layer of the network. Services dedicated to each application (firewall load balancers, etc.) are placed in vertical silos dedicated to a group of application servers. Finally, th network management is centered in the switch operating system; over time, this has come to include wide range of complex and often vendor proprietary features and functions. This approach was ver successful for campus Local Area Networks (LANs), which led to its adoption in most data center despite the fact that this approach was never intended to meet the traffic patterns or performanc requirements of a modern data center environment.

There are many problems with applying conventional campus networks to modern data cent designs. While some enterprise data centers have used mainframe-class computers, and thus take advantage of server virtualization, reliability, and scalability, many more use x86-based server which have not supported these features until more recently. Conventional data centers have consiste of lightly utilized servers running a bare metal operating system or a hypervisor with a small numb of virtual machines (VMs). The servers may be running a mix of different operating system including Windows, Linux, and UNIX. The network consists of many tiers, where each lay duplicates many of the IP/Ethernet packet analysis and forwarding functions. This adds cumulativ end-to-end latency (each network tier can contribute anywhere from 2 to 25 µs) and requir significant amounts of processing and memory. Oversubscription, in an effort to reduce latency an promote cost-effective scaling, can lead to lost data and is not suitable for storage traffic, whic cannot tolerate missing or out of order data frames. Thus, multiple networks are provided for bo Ethernet and Fibre Channel (and to a lesser degree for other specialized applications such as serv clustering or other protocols such as InfiniBand). Each of these networks may require its ow dedicated management tools, in addition to server, storage, and appliance management. Server typically attach to the data center network using lower bandwidth links, such as 1 Gbit/s Ethernet an either 2, 4, 8, or 16 Gbit/s Fibre Channel storage area networks (SANs).

The network design shown in Figure 1.1 is well suited for applications in which most of the da traffic flows in a north–south direction either between clients and servers or between the servers an the wide area network (WAN). This was the case in campus LANs, which used a central wiring clos to house networking equipment and distributed data traffic using an approach similar to the electric

wiring system in an office building. However, in modern data centers with large numbers of VMs p server, an increasing amount of data traffic flows between servers (so-called east–west traffic). It h been estimated that as much as 75% of the traffic in cloud computing environments follows th approach. Conventional multitier networks were never intended to handle these traffic patterns and a result often suffer from suboptimal performance.

Conventional networks do not scale in a cost-effective or performance-effective manner. Scalin requires adding more tiers to the network, more physical switches, and more physical servi appliances. Management functions also do not scale well, and IPv4 addresses may become exhauste as the network grows. Network topologies based on STP can be restrictive for modern applications ar may prevent full utilization of the available network bandwidth. The physical network must b manually rewired to handle changes in the application workloads, and the need to manually configu features such as security access makes these processes prone to operator error. Further, convention networks are not optimized for new features and functions. There are unique problems associated wi network virtualization (significantly more servers can be dynamically created, modified, or destroye which is difficult to manage with existing tools). Conventional networks also do not easily provide f VM migration (which would promote high availability and better server utilization), nor do the provide for cloud computing applications such as multitenancy within the data center.

Attempting to redesign a data center network with larger Layer 2 domains (in an effort to facilita VM mobility) can lead to various problems, including the well-known “traffic trombone” effect. Th term describes a situation in which data traffic is forced to traverse the network core and back agai similar to the movement of a slide trombone, resulting in increased latency and lower performance. some cases, traffic may have to traverse the network core multiple times, or over extended distance further worsening the effect. In a conventional data center with small Layer 2 domains in the acce layer and a core IP network, north–south traffic will be bridged across a Layer 2 subnet between th access and the core layers, and the core traffic will be routed east–west, so that packets normal traverse the core only once. In some more modern network designs, a single Layer 2 domain stretched across the network core, so the first-hop router may be far away from the host sending th packet. In this case, the packet travels across the core before reaching the first-hop router, then bac again, increasing the latency as well as the east–west traffic load. If Layer 3 forwarding implemented using VMs, packets may have to traverse the network core multiple times befo reaching their destination. Thus, inter-VLAN traffic flows with stretched or overlapping Layer domains can experience performance degradation due to this effect. Further, if Layer 2 domains ar extended across multiple data centers and the network is not properly designed, traffic flows betwee sources and destinations in the same data center may have to travel across the long distance lin between data centers multiple times.

Installation and maintenance of this physical compute model requires both high capital expense ar high operating expense. The high capital expense is due to the large number of underutilized serve and multiple interconnect networks. Capital expense is also driven by multitier IP networks, and th use of multiple networks for storage, IP, and other applications. High operational expense is driven b high maintenance and energy consumption of poorly utilized servers, high levels of manual networ and systems administration, and the use of many different management tools for different parts of th data center. As a result, the management tasks have been focused on maintaining the infrastructu and not on enhancing the services that are provided by the infrastructure to add business value.

1.1 Properties of a network reference architecture

Modern data centers are undergoing a major transition toward a more dynamic infrastructure. Th allows for the construction of flexible IT capability that enables the optimal use of information support business initiatives. For example, a dynamic infrastructure would consist of highly utiliz servers running many VMs per server, using high-bandwidth links to communicate with virtu storage and virtual networks both within and across multiple data centers. As part of the dynam infrastructure, the role of the data center network is also changing in many important ways, causi many clients to reevaluate their current networking infrastructure. Many new industry standards ha emerged and are being implemented industrywide. The accelerating pace of innovation in this area h also led to many new proposals for next-generation networks to be implemented within the next fe years.

Proper planning for the data center infrastructure is critical, including consideration of such facto as latency and performance, cost-effective scaling, resilience or high availability, rapid deployment new resources, virtualization, and unified management. The broad interest that IT organizations ha in redesigning their data center networks is driven by the desire to reduce cost (both capital a operating expense) while simultaneously implementing the ability to support an increasingly dynam (and in some cases highly virtualized) data center. There are many factors to consider whe modernizing the design of a data center network.

There are several underlying assumptions about data center design inherent in the network referen architecture. For example, this architecture should enable users to treat data center computi storage, services, and network resources as fully fungible pools that can be dynamically and rapid partitioned. While this concept of a federated data center is typically associated with cloud computi environments, it also has applications to enterprise data centers, portals, and other common use case Consider the concept of a multitenant data center, in which the tenants represent clients, sharing common application space. This may include sharing data center resources among different divisio of a company (accounting, marketing, research), stock brokerages sharing a real-time trading engi government researchers sharing VMs on a supercomputer, or clients sharing video streaming from content provider. In any sort of multitenant data center, it is impractical to assume that t infrastructure knows about the details of any application or that applications know about details of t infrastructure. This is one of the basic design concepts that lead to simplicity, efficiency, and securi in the data center.

As another example, this architecture should provide connectivity between all available data cent resources with no apparent limitations due to the network. An ideal network would offer infini bandwidth and zero latency, be available at all times, and be free to all users. Of course, in practi there will be certain unavoidable practical considerations; each port on the network has a fixed upp limit on bandwidth or line rate and minimal, nonzero transit latency, and there are a limited number ports in the network for a given level of subscription. Still, a well-designed network will minimi these impacts or make appropriate trade-offs between them in order to make the network a seamles transparent part of the data processing environment. This is fundamental to realizing hig performance and cost efficiency. Another key capability of this new compute model involv providing a family of integrated offerings, from platforms that offer server, storage, and networki resources combined into a simple, turnkey solution to network and virtualization building blocks th scale to unprecedented levels to enable future cloud computing systems.

Network switches must support fairly sophisticated functionality, but there are several options f

locating this intelligence (i.e., control plane functionality) within a data center network. First, w could move the network intelligence toward the network core switches. This option has som economic benefits but limits scalability and throughput; it is also not consistent with the design of dynamic, workload-aware network, and requires significant manual configuration. Second, we cou move the intelligence toward the network edge. This is a technically superior solution, since provides improved scale and throughput compared with the intelligent core option. It also enabl better interaction between the servers and the network. This approach may face economic challenge since the intelligence needs to be embedded within every edge switch in the network. A third approac is to move the network intelligence into the attached servers. This provides relatively large scale an high throughput, at least for some applications. This option decouples the physical network from th network provisioning and provides for a dynamic logical network that is aware of the requirements f workload mobility. Over time, emerging industry standards for software-defined networking (SDN and network overlays will increasingly provide for dynamic provisioning, management flexibility, an more rapid adoption of new technologies.

Each of these approaches is a nontrivial extension of the existing data center network; collectivel they present a daunting array of complex network infrastructure changes, with far-reachin implications for data center design. In the following sections, we discuss in detail the key attributes a next-generation network architecture, including the problems solved by these approaches. Th discussion of next-generation network requirements follows best practices for the design of standard based networks, including the open data center interoperable network (ODIN), which has bee endorsed by many industry leading companies [10–13]. Later in this book, we will discuss bo industry standard and vendor proprietary approaches to delivering these features.

1.1.1 Flattened, Converged Networks

Classic Ethernet networks are hierarchical, as shown in Figure 1.1, with three, four, or more tie (such as the access, aggregation, and core switch layers). Each tier has specific design consideration and data movement between these layers is known as multitiering. The movement of traffic betwee switches is commonly referred to as “hops” in the network (there are actually more precise technic definitions of what constitutes a “hop” in different types of networks, which will be discussed in mo detail later). In order for data to flow between racks of servers and storage, data traffic needs to trav up and down a logical tree structure as shown in Figure 1.1. This adds latency and potentially creat congestion on interswitch links (ISLs). Network loops are prevented by using STP, which allows onl one active path between any two switches. This means that ISL bandwidth is limited to a single logic connection, since multiple connections are prohibited. To overcome this, link aggregation group (LAGs) were standardized, so that multiple links between switches could be treated as a single logic connection without forming loops. However, LAGs have their own limitations, for example, they mu be manually configured on each switch port.

Many clients are seeking a flattened network that clusters a set of switches into one large (virtua switch fabric. This would significantly reduce both operating and capital expense. Topologically, “flat” network architecture implies removing tiers from a traditional hierarchical data center netwo such that it collapses into a two-tier network (access switches, also known as top of rack (TOR switches, and core switches). Most networking engineers agree that a flat fabric also implies th connected devices can communicate with each other without using an intermediate router. A fla network also implies creating larger Layer 2 domains (connectivity between such domains will st

require some Layer 3 functionality). This flat connectivity simplifies the writing of applications sinc there is no need to worry about the performance hierarchy of communication paths inside the da center. It also relieves the operations staff from having to worry about the “affinity” of applicatio components in order to provide good performance. In addition, it helps prevent resources in a da center from becoming stranded and not efficiently usable. Flatter networks also include elimination STP and LAG. Replacing the STP protocol allows the network to support a fabric topology (tree, rin mesh, or core/edge) while avoiding ISL bottlenecks, since more ISLs become active as traffic volum grows. Self-aggregating ISL connections replace manually configured LAGs.

Flattening and converging the network reduces capital expense through the elimination of dedicate storage, cluster and management adapters and their associated switches, and the elimination traditional networking tiers. Operating expense is also reduced through management simplification b enabling a single console to manage the resulting converged fabric. Note that as a practic consideration, storage traffic should not be significantly oversubscribed, in contrast to convention Ethernet design practices. The use of line rate, nonblocking switches is also important in a converge storage network, as well as providing a forward migration path for legacy storage. Converging an flattening the network also leads to simplified physical network management. While convention data centers use several tools to manage their server, storage, network, and hypervisor elements, be practices in the future will provide a common management architecture that streamlines th discovery, management, provisioning, change/configuration management, problem resolution an reporting of servers, networking, and storage resources across the enterprise. Such a solution helps optimize the performance and availability of business-critical applications, along with supporting I infrastructure. It also helps to ensure the confidentiality and data integrity of information assets, whi protecting and maximizing data utility and availability. Finally, converged and flattened data center may require new switch and routing architectures to increase the capacity, resiliency, and scalabilit of very large Layer 2 network domains.

The datacom industry has begun incrementally moving toward the convergence of fabrics that use to be treated separately, including the migration from Fibre Channel to Fibre Channel over Ethern (FCoE) and the adoption of Remote Direct Memory Access (RDMA) over Ethernet standards for high performance, low-latency clustering. This will occur over time and available industry data suggest that Fibre Channel and iSCSI or network attached storage (NAS) are not expected to go away anytim soon. Within the WAN, many telecommunication or Internet service providers are migrating towar Ethernet-based exchanges, which replace conventional Asynchronous Transport Mode (ATM) Synchronous Optical Network (SONET) / Synchronous Digital Hierarchy (SDH) protocols, and pack optical networks are emerging as an important trend in the design of these systems.

It should be pointed out that as this book goes to press, the transition toward an Ethernet dominate network is well under way, but remains far from complete. Despite strong interest in converging th entire data center onto Ethernet and related protocols such as FCoE and RDMA over Converge Ethernet (RoCE), the data center continues to be a multiprotocol environment today and some feel will remain so for the foreseeable future. According to Bundy [14], in 2011 Fibre Channel shippe more optical bandwidth than any other protocol (84 petabytes/s). Ethernet bandwidth was secon highest, with just under 73 petabytes/s. That same year, Fibre Channel shipped over 11.7 millio Small form factor pluggable (SFP)+ optical transceivers, of which 7.7 million support 8 Gbit/s da rates. While only 156,000 transceivers capable of 16 Gbit/s data rates were shipped, this number wa projected to grow dramatically in 2012 (by 268%). In terms of performance, Fibre Channel continu to lead industry benchmarks for high-performance VM applications [15]. Ethernet was the ne