Interconnecting Data Centers Using VPLS

Ensure Business Continuance on Virtualized Networks by Implementing Layer 2 Connectivity Across Layer 3

Nash Darukhanawalla, CCIE® No. 10332
Patrice Bellagamba

ciscopress.com

Interconnecting Data Centers Using VPLS

Nash Darukhanawalla, Patrice Bellagamba

Published by:
Cisco Press
800 East 96th Street
Indianapolis, IN 46240 USA

Printed in the United States of America

First Printing June 2009

Library of Congress Cataloging-in-Publication Data:

Darukhanawalla, Nash, 1963-

Interconnecting data centers using VPLS / Nash Darukhanawalla and Patrice Bellagamba.

p. cm.

ISBN-13: 978-1-58705-992-6 (pbk.)

ISBN-10: 1-58705-992-4 (pbk.)

1. Extranets (Computer networks) 2. Internetworking (Telecommunication) 3. Electronic data processing departments. 4. Cisco IOS. I. Bellagamba, Patrice, 1959- II. Title.

TK5105.875.E87D37 2010

004.6—dc22

2009020207

ISBN-13: 978-1-58705-992-6

ISBN-10: 1-58705-992-4

Warning and Disclaimer

This book is designed to provide information about interconnecting data centers using Virtual Private LAN Service (VPLS). Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied.

The information is provided on an “as is” basis. The authors, Cisco Press, and Cisco Systems, Inc., shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book or from the use of the discs or programs that may accompany it.

The opinions expressed in this book belong to the author and are not necessarily those of Cisco Systems, Inc.

Trademark Acknowledgments

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Cisco Press or Cisco Systems, Inc., cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Corporate and Government Sales

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact: **U.S. Corporate and Government Sales** 1-800-382-3419 corpsales@pearsontechgroup.com

For sales outside the United States please contact: **International Sales** international@pearsoned.com

Feedback Information

At Cisco Press, our goal is to create in-depth technical books of the highest quality and value. Each book is crafted with care and precision, undergoing rigorous development that involves the unique expertise of members from the professional technical community.

Readers' feedback is a natural continuation of this process. If you have any comments regarding how we could improve the quality of this book, or otherwise alter it to better suit your needs, you can contact us through email at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

Publisher: Paul Boger

Associate Publisher: Dave Dusthimer

Executive Editor: Mary Beth Ray

Managing Editor: Patrick Kanouse

Senior Development Editor: Christopher Cleveland

Project Editor: Jennifer Gallant

Editorial Assistant: Vanessa Evans

Book Designer: Louisa Adair

Composition: Mark Shirar

Cisco Representative: Eric Ullanderson

Cisco Press Program Manager: Anand Sundaram

Copy Editor: Keith Cline

Technical Editors: Cesar Carballes and Yves Louis

Indexer: Brad Herriman

Proofreader: Apostrophe Editing Services

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Introduction

This book presents Virtual Private LAN Service (VPLS)-based solutions that provide a high-speed, low-latency network and Spanning Tree Protocol (STP) isolation between data centers. The book also includes detailed information about issues that relate to large Layer 2 bridging domains and offers guidance for extending VLANs over Layer 3 networks using VPLS technology.

The solutions presented in this book have been validated under the Cisco Validated Design System Assurance program. All solutions were validated with a wide range of system tests, including system integration, fault and error handling, and redundancy. Testing also verified the end-to-end flow of both unicast and multicast unidirectional traffic. Voice, using components of the Cisco Unified Communications solution, was also implemented and verified.

The solutions in this book were developed because globalization, security, and disaster recovery considerations are driving divergence of business locations across multiple regions. In addition, organizations are looking to distribute workload between computers, share network resources effectively, and increase the availability of applications. With the ultimate goal of eliminating all downtime and sharing data across regions, enterprises are deploying geographically dispersed data centers to minimize planned or unplanned downtime, whether it is caused by a device failure, security breach, or natural disaster.

As data centers grow in size and complexity, enterprises are adopting server virtualization technologies to achieve increased efficiency and use of resources. In addition to providing resource optimization, virtualization strategies offer data protection, which enables enterprises to build disaster recovery solutions and provide high availability, scalability, flexibility, and business continuity.

Server virtualization technologies include the following:

- **VMotion**: Allows an individual virtual machine (such as Windows Server) to be dynamically moved to another VMware server. A dedicated VLAN is required for VMotion traffic so that virtual machines can be moved without affecting users. In addition, the group of servers that VMs are balanced between must be in the same Layer 2 domain, because attributes such as an IP address cannot change when a virtual machine moves. Therefore, all VMware servers, including the source VMware server, must have connections to the same VLAN.

- **NIC teaming**: Servers that include only one network interface card (NIC) are susceptible to many single points of failure, such as a failure of the NIC, its network cable, or the access switch to which it connects. A solution developed by NIC vendors, NIC teaming eliminates this single point of failure. In this solution, special drivers allow two NICs to be connected to separate access switches or to separate line cards on the same access switch. If one NIC fails, the other NIC assumes the IP address of the server and takes over operation without disruption. Types of NIC teaming solutions include active/standby and active/active. All solutions require Layer 2 adjacency for the teamed NICs.

- **Server clusters:** High-availability (HA) server clusters have become key components of IT strategies as organizations need to increase processing power, distribute workloads between computers, share network resources effectively, and increase the availability of applications. HA clusters typically are built with two separate networks: a public network to access the active node of the cluster from outside the data center, and a private network to interconnect the nodes of the cluster for private communications. The private network connection is also used to monitor the health and status of each node in the HA cluster using the heartbeat system. This solution requires that the network is capable of handling any kind of failure without causing a split-brain condition that could lead to duplicate instances of services and even the corruption of data on shared storage devices. The private network is a nonrouted network that shares the same Layer 2 VLAN between the nodes of the cluster even when extended between multiple sites.

These virtualization technologies have resulted in an expansion of Layer 2 domains, which in turn has increased the spanning-tree domain at the network level. STP was developed to handle a network with a small diameter, so an enterprise network with geographically dispersed data centers needs an effective solution for Layer 2 connectivity between multiple sites.

Also, during the process of migrating physical servers from one location to another, it is much easier to extend the Layer 2 VLAN and maintain the original configuration of the systems, thus avoiding IP address renumbering. Even during a phased migration period, when just part of the server farm is being relocated, the Layer 2 adjacency is often required across the entire server farm to ensure business continuity.

As data center resources and security requirements continue to grow, organizations must connect multiple data centers over larger distances. As a result, organizations are facing additional challenges such as maintaining the high availability of applications and dealing with complex multisite interconnections.

Objective of This Book

This book provides design guidance, configuration examples, and best practices for deploying a single IP/MPLS-based network to interconnect data centers ensuring high availability Layer 2 connectivity with STP isolation in the core. Customers who have already deployed a separate optical network for Layer 2 extension can also take advantage of these solutions to reduce infrastructure and maintenance costs.

This book addresses issues that are related to large Layer 2 bridging domains and provides guidance for extending VLANs using VPLS technology.

This book also examines in detail the technologies that offer such solutions, explains the benefits and trade-offs of various solutions, and describes a variety of deployment options. The deployment model that an organization chooses depends on the complexity of the requirements, the protocols currently deployed in the data centers, the scalability required, and many other factors.

Who Should Read This Book

This book is intended for systems professionals and system engineers who are designing solutions for interconnecting data centers that ensure high availability Layer 2 connectivity and STP isolation. Service providers that offer metro Ethernet leased-line aggregation and Layer 2 transport services can also benefit from these solutions that provide large-scale Layer 2 extension.

Cisco Validated Design Program

The Cisco Validated Design (CVD) program designs, tests, and documents systems and solutions to facilitate faster, more reliable, and more predictable customer deployments. The program includes Cisco Validated Design and CVD System Assurance.

Cisco Validated Design

Cisco Validated Designs are systems or solutions that have been validated through architectural review and proof-of-concept testing in a Cisco lab. Cisco Validated Design provides guidance for deploying new technologies or for applying enhancements to existing infrastructure.

CVD System Assurance

The Cisco Validated Design System Assurance is a program that identifies systems that have undergone architectural- and customer-relevant testing.

A CVD certified design is a highly validated and customized solution that meets the following criteria:

- Reviewed and updated for general deployment
- Achieves the highest levels of consistency and coverage within the Cisco Validated Design program
- Solution requirements successfully tested and documented with evidence to function as detailed within a specific design in a scaled, customer representative environment
- Zero observable operation impacting defects within the given test parameters; that is, no defects that have not been resolved either outright or through software change, redesign, or workaround

For more information about Cisco CVD program, refer to http://tinyurl.com/6gxuk2.

How This Book Is Organized

The material in this book is presented in a building-block fashion that takes you from the legacy deployement models for data center interconnect (DCI) and problems associated with extending Layer 2 networks, through VPN technologies, to various Multiple Spanning Tree (MST)-, Embedded Event Manager (EEM)- and generic routing encapsulation (GRE)-based deployment models, and beyond. Although this book is intended to be read cover to cover, it is designed to be flexible and allow you to easily find information that applies to your needs.

The chapters cover the following topics:

- **Chapter 1, "Data Center Layer 2 Interconnect":** This chapter provides an overview of high availability clusters. It also explains DCI legacy deployment models and problems associated with extending Layer 2 networks.
- **Chapter 2, "Appraising Virtual Private LAN Service":** This chapter discusses Layer 2 and Layer 3 VPN technologies and provides introductions to VPLS, pseudowires, EEM, and MPLS.
- **Chapter 3, "High Availability for Extended Layer 2 Networks":** This chapter focuses on design components such as maximum transmission unit (MTU), core routing protocols, and convergence optimization techniques to achieve high availability.
- **Chapter 4, "MPLS Traffic Engineering":** This chapter explains the implemetation of MPLS-TE for load repartition of Layer 2 VPN traffic over parallel links. It also introduces Fast Reroute (FRR) for faster convergence.
- **Chapter 5, "Data Center Interconnect: Architecture Alternatives":** This chapter highlights several options for implementing DCI. In addition, this chapter provides guidance for selecting an appropriate solution based on your requirements (such as scalability and ease of implementation).
- **Chapter 6, "Case Studies for Data Center Interconnect":** This chapter provides case studies that relate to the DCI solutions that this book describes.
- **Chapter 7, "Data Center Multilayer Infrastructure Design":** This chapter highlights the Cisco data center multitier model and provides information about network topology, hardware, software, and traffic profiles used for validating the designs in this book.
- **Chapter 8, "MST-Based Deployment Models":** This chapter covers "MST in pseudowire" and "isolated MST in N-PE" solutions and provides configuration details for implementing these solutions.

- **Chapter 9, "EEM-Based Deployment Models":** This chapter explains "EEM semaphore protocol," which was developed to achieve N-PE redundancy in the absence of ICCP. In addition, this chapter describes various EEM-based VPLS and Hierarchical VPLS (H-VPLS) solutions, provides in-depth theory about the operation of each solution, and provides configuration details.
- **Chapter 10, "GRE-Based Deployment Models":** This chapter provides VPLS and H-VPLS solutions over an IP network using VPLSoGRE (VPLS over GRE).
- **Chapter 11, "Additional Data Center Interconnect Design Considerations":** This chapter introduces other technologies or issues that should be considered when designing DCI solutions.
- **Chapter 12, "VPLS PE Redundancy Using Inter-Chassis Communication Protocol":** This chapter introduces ICCP protocols and provides various redundancy mechanisms and sample configurations.
- **Chapter 13, "Evolution of Data Center Interconnect":** This chapter provides a brief overview of the emerging technologies and the future of DCI.
- **Glossary:** This element provides definitions for some commonly used terms associated with DCI and the various deployment models discussed in the book.

The authors have also written several documents, including *Interconnecting Geographically Dispersed Data Centers, High Availability Clusters, Layer 2 Extension Between Remote Data Centers*. These documents cover a few key concepts from this book and are freely available on Cisco.com.

Chapter 1

Data Center Layer 2 Interconnect

Many enterprises are making fundamental changes to their business processes by using advanced IT applications to achieve enhanced productivity and operational efficiencies. As a result, the underlying network architecture to support these applications is evolving to better accommodate this new model.

As data availability becomes a critical requirement, many businesses are devoting more resources to ensure continuous operation. Enterprises are provisioning dedicated networks to guarantee performance metrics for applications without compromising security.

Although maintaining uninterruptible access to all data center applications is desirable, the economics of business-continuance require network operators to prioritize applications according to their importance to the business. As a result, data centers need a range of business-continuance solutions to accommodate this goal, from simple tape backup and remote replication to synchronous mirroring and mirrored distributed data centers.

Enterprises can enhance application resilience in several ways, including the following:

- Removing single points of server failure by deploying high-availability clusters or load-balancing technology across web and application servers
- Extending the deployment of these clusters in different data centers to protect against major disruptions

User access is as important as downtime protection and data recovery. Following a disruption, how long can the business afford for users to be without access to applications? Companies are employing technologies such as Global Site Selector that enable users to manually or automatically connect to an alternative site running the application they need.

Businesses run tens and often hundreds of applications, each of which might have differing continuance requirements, measured in a time-to-recovery and data-loss perspective. IT groups need to match the associated characteristics and cost of a business-continuance solution to the potential business and consider which technologies to deploy where problems impact data, applications, and user access.

Cisco delivers scalable, secure, and cost-effective technology that helps enterprises build end-to-end backup and recovery solutions and disaster recovery solutions. These solutions include the following:

- High-availability data center networking and storage-area networks for nonstop access to applications and data
- Synchronized distributed data centers for continuous service over WANs in the event of site disruptions
- Synchronous disk mirroring and replication over WANs for fast recovery and zero data loss
- Asynchronous data replication over IP networks for remote data protection
- Consolidated backup to tape or near-line disk and remote electronic vaulting over enterprise-wide storage networks for consistent protection of distributed data

Each of these solutions requires the appropriate network infrastructure to help ensure that user-specific availability, performance, distance, and latency requirements are met. In addition, enterprises require a resilient, integrated business-continuance network infrastructure to protect three key areas in the event of a disruption:

- Data
- Applications
- User access

Overview of High-Availability Clusters

High-availability (HA) clusters operate by using redundant computers or nodes that provide services when system components fail. Normally, if a server with a particular application crashes, the application is unavailable until the problem is resolved. HA clustering remedies this situation by detecting hardware/software faults, and immediately providing access to the application on another system without requiring administrative intervention. This process is known as *failover*.

HA clusters are often used for key databases, file sharing on a network, business applications, and customer services such as e-commerce websites. HA cluster implementations attempt to build redundancy into a cluster to eliminate single points of failure. These implementations include multiple network connections and data storage that connects via storage-area networks (SAN).

HA clusters usually are built with two separate networks:

- **The public network:** Used to access the active node of the cluster from outside the data center

- **The private network:** Used to interconnect the nodes of the cluster for private communications within the data center and to monitor the health and status of each node in the cluster

Public Network Attachment

For the public network (facing the nodes cluster), the server often is enabled by a dual-homing mechanism with one network interface card (NIC) configured in active state and one NIC configured in standby state. If a link to the active NIC fails, or the NIC loses connectivity with its default gateway, the operating system performs a failover. A NIC failover for a public network has no affect on cluster availability because the heartbeat mechanism and NICs in active/standby mode for public access are two separate hand-check mechanisms.

The network design must provide the highest availability for the LAN infrastructure. To achieve this goal, the teaming service or dual-homing should be distributed between different access switches, which in turn should be connected to different aggregation switches, as illustrated in Figure 1-1.

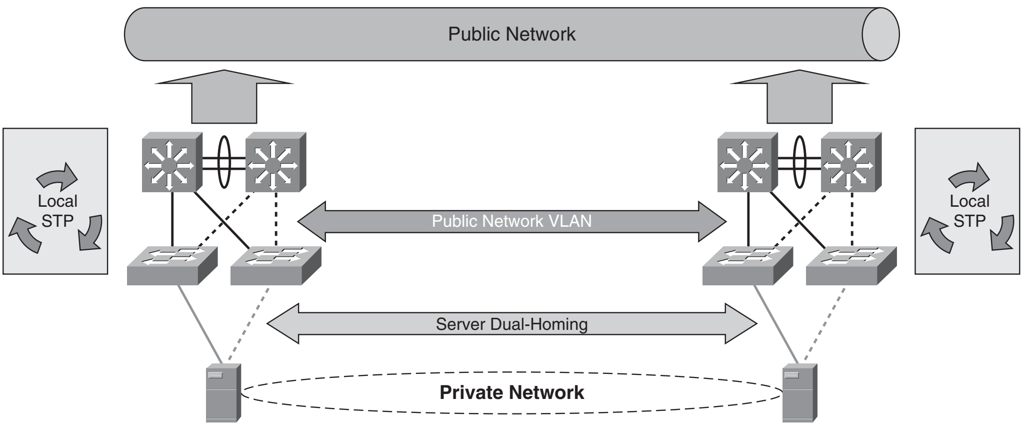

Figure 1-1 *Extending the public network.*

Private Network Attachment

The private network primarily carries cluster heartbeat, or keepalive, messages. Other server-to-server communications that occur on this private network include the following:

- Cluster data
- Cluster file system data
- Application data (back-end)

The private network is a nonrouted network that shares the same Layer 2 VLAN between the nodes of the cluster even when extended across multiple sites. In a campus cluster environment, heartbeats are sent via the private network from node to node of the HA cluster using a proprietary Layer 2 (nonroutable) protocol. The servers manage the I/O by sending traffic over all interfaces and by preventing traffic from being sent over a failing path. This approach provides resiliency in the event of a NIC failure on a server.

The heartbeat is the most important component of the cluster that uses the private network interconnection. However, if all paths go down for more than 10 seconds (applicable for most HA clusters), a *split-brain* situation can occur, which prompts the cluster framework to check the number of votes and decide which server or servers will continue as the members in the cluster. Nodes that lose cluster membership assume that they are isolated, and any applications that run on those nodes terminate. Surviving nodes know that the nonsurviving nodes have stopped, and the cluster will then restart the applications.

Although some HA cluster vendors recommend disabling Spanning Tree Protocol (STP) for the private interconnect infrastructure, such a drastic measure is neither necessary nor recommended when using Cisco Catalyst switches. In fact, Cisco has since provided the *PortFast* feature, which puts an access port into forwarding mode immediately after link up without losing loop-detection capabilities. To avoid connectivity delays, PortFast must be enabled on all access interfaces connecting cluster nodes. This rule also applies to any servers connected to the switch. The IEEE also defines the PortFast concept within the Rapid STP 802.1w standard under the *edge port* designation. In addition, Cisco supports Per-VLAN Spanning Tree, which maintains a spanning-tree instance for each VLAN configured in the network.

Note For detailed information about HA clusters, refer to the Windows HPC Server 2008 site at http://www.microsoft.com/HPC/.

For detailed information about STP PortFast configuration to resolve server/workstation startup connectivity delay, refer to the Cisco document Using PortFast and Other Commands to Fix Workstation Startup Connectivity Delays, available at http://tinyurl.com/2e29bw.

For detailed information about designing a network for extended HA clusters, refer to the following Cisco white paper A: "Technology and Networking Guide for High Availability Clusters Extended Across Multiple Data Centers," at http://tinyurl.com/cb4f3k.

Data Center Interconnect: Legacy Deployment Models

Several transport technologies are available for interconnecting the data centers, each of which provides various features and allows different distances, depending on factors such as the power budget of the optics, the lambda used for transmission, the type of fiber, and so forth.

Consider the features of the LAN and SAN switches that provide higher availability for the data center interconnect (DCI) before considering some of the available technologies. The convergence time required for the application also is important and should be evaluated.

The list that follows describes common transport options:

- **Dark fiber:** Dark fiber is a viable method for extending VLANs over data center or campus distances. The maximum attainable distance is a function of the optical characteristics (transmit power and receive sensitivity) of the LED or laser that resides in a small form-factor pluggable (SFP) or Gigabit Interface Converter (GBIC) transponder, combined with the number of fiber joins, and the attenuation of the fiber.
- **Coarse wavelength-division multiplexing (CWDM):** CWDM offers a simple solution to carry up to eight channels (1 Gbps or 2 Gbps) on the same fiber. These channels can carry Ethernet or Fiber Channel. CWDM does not offer protected lambdas, but client protection allows rerouting of the traffic on the functioning links when a failure occurs. CWDM lambdas can be added and dropped, allowing the creation of hub-and-spoke, ring, and meshed topologies. The maximum achievable distance is approximately 60 miles (100 km) with a point-to-point physical topology and approximately 25 miles (40 km) with a physical ring topology.
- **Dense wavelength-division multiplexing (DWDM):** DWDM enables up to 32 channels (lambdas), each of which can operate at up to 10 Gbps. DWDM networks can be designed either as multiplexing networks that are similar to CWDM or with a variety of protection schemes to guard against failures in the fiber plant. DWDM also offers more protection mechanisms (splitter protection and Y-cable protection), and the possibility to amplify the channels to reach greater distances.

Note For details about data center transport technologies, refer to Chapter 2 of *Data Center High Availability Clusters Design Guide*, available at http://tinyurl.com/ct4cw8.

In nearly all of these deployment models, costs associated with deploying and maintaining a dedicated optical network is one of the biggest concerns. Also, there is no STP isolation. Depending on the nature of the problem, issues in one data center will affect other data centers. Another disadvantage is the lack of load balancing across redundant paths due to blocked links in the core network.

Problems Associated with Extended Layer 2 Networks

A common practice is to add redundancy when interconnecting data centers to avoid split-subnet scenarios and interruption of the communication between servers, as illustrated in Figure 1-2. The split-subnet is not necessarily a problem if the routing metric makes one site preferred over the other. Also, if the servers at each site are part of a cluster and the communication is lost, mechanisms such as the quorum disk avoid a split-brain condition.

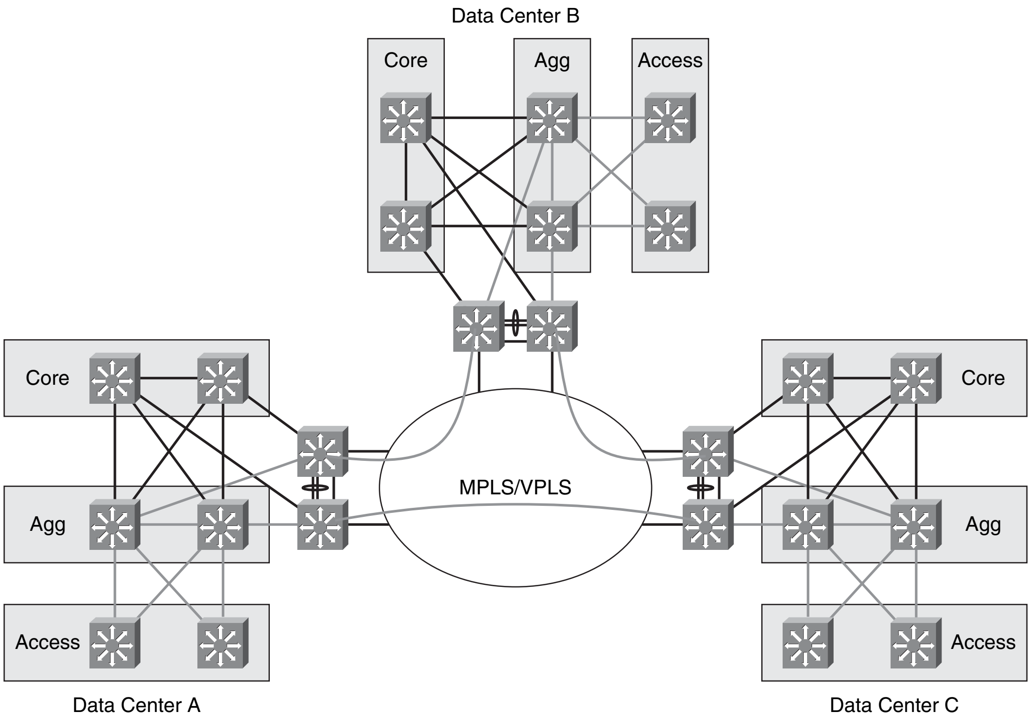

Figure 1-2 *Layout of multiple data center interconnect with redundant N-PEs in each data center.*

Adding redundancy to an extended Ethernet network typically means relying on STP to keep the topology loop free. STP domains should be reduced as much as possible and limited within the data center. Cisco does not recommend deploying the legacy 802.1d because of its old timer-based mechanisms that make the recovery time too slow for most applications, including typical clustering software.

An extended Layer 2 network does introduce some problems to contend with, however.

STP operates at Layer 2 of the Open Systems Interconnection (OSI) model, and the primary function STP is to prevent loops that redundant links create in bridge networks. By exchanging bridge protocol data units (BPDU) between bridges, STP elects the ports that eventually forward or block traffic.

The conservative default values for the STP timers impose a maximum network diameter of seven. Therefore, two bridges cannot be more than seven hops away from each other.

When a BPDU propagates from the root bridge toward the leaves of the tree, the age field increments each time the BPDU goes through a bridge. Eventually, the bridge discards the BPDU when the age field goes beyond maximum age. Therefore, convergence of the spanning tree is affected if the root bridge is too far away from some bridges in the network.

An aggressive value for the max-age parameter and the forward delay can lead to an unstable STP topology. In such cases, the loss of some BPDUs can cause a loop to appear. Take special care if you plan to change STP timers from the default value to achieve faster STP convergence.

Unlike legacy STP, Rapid STP (RSTP) converges faster because it does not depend on the timers to make a rapid transition. However, STP does not provide the required robustness for large-scale Layer 2 deployments:

- Network stability is compromised as a result of slow response to network failures (slow convergence). Even new spanning-tree developments such as RSTP and Multiple Spanning Tree (MST) assume good-quality physical connections such as dark fiber or WDM connections. These STP protocols are not built to accommodate frequent link-flapping conditions, high error rates, unidirectional failures, or nonreport of loss of signal. These typical and frequent behaviors of long- and medium-distance links could lead to STP slow convergence or even instability.
- The primary reason for multisite data centers is disaster recovery. However, because data centers typically require Layer 2 connectivity, failure in one data center can affect other data centers, which could lead to a blackout of all data centers at the same time.
- A broadcast storm propagates to every data center, which, if uncontrolled, could result in network-wide outage.
- STP blocks links, which prevents load balancing of traffic across redundant paths in the core network.

Note For understanding and tuning STP timers and the rules to tune them when absolutely necessary, refer to the Cisco document, "Understanding and Tuning Spanning Tree Protocol Timers," available at http://tinyurl.com/7ppqq.

Summary

This chapter provided an overview of HA clusters, legacy deployment models for interconnecting data centers, and problems related to extending Layer 2 networks. The solutions that this book presents address these issues in more detail and provide guidance for designing and deploying DCI.

Index

Symbols

A

B

C

D

E

F

I

K-L

M

N

O

P

Q

R

S

T

U

V-Z