

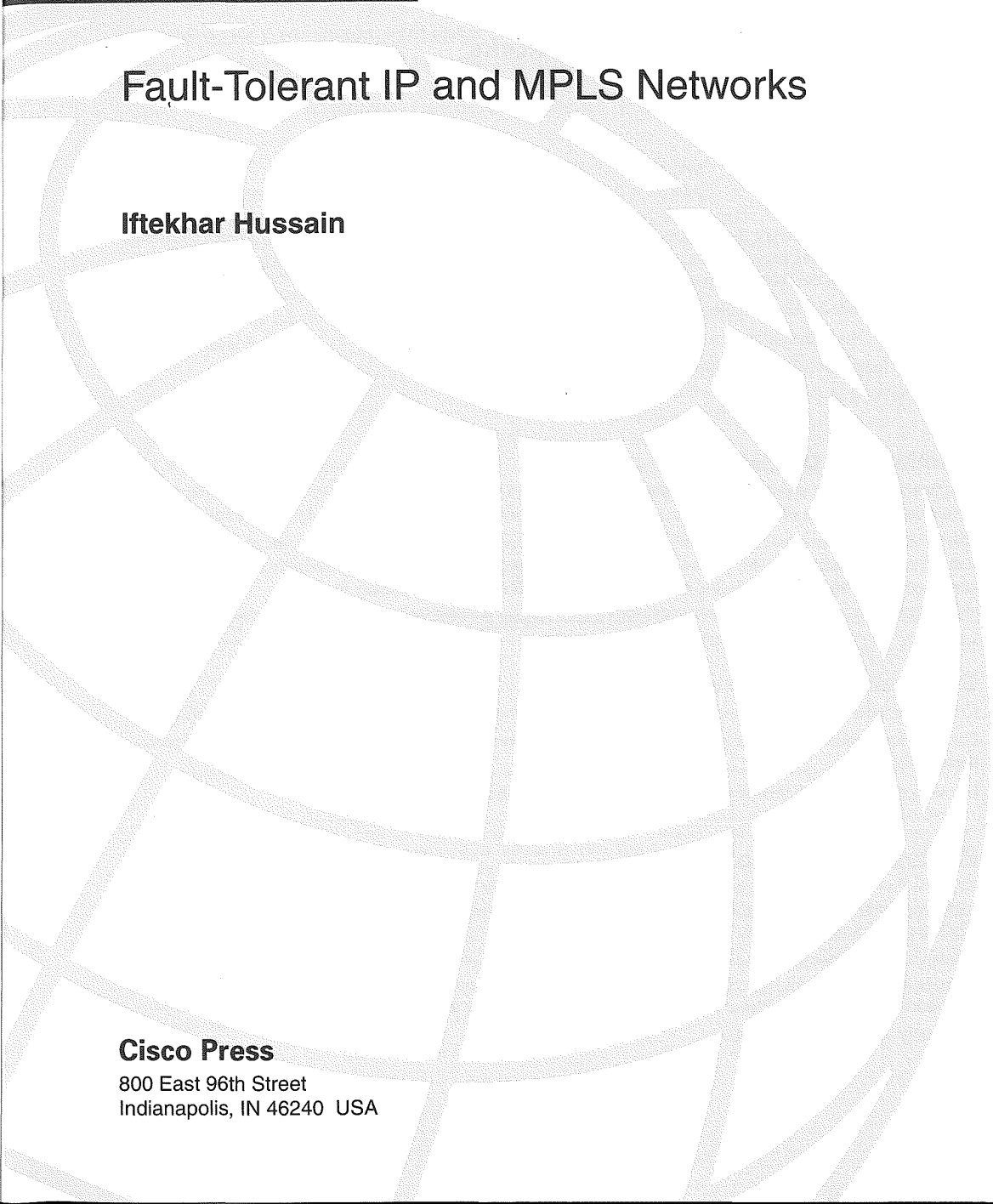


Fault-Tolerant IP and MPLS Networks

Design and deploy high availability IP and MPLS network architectures with this comprehensive guide

ciscopress.com

Iftekhar Hussain



Fault-Tolerant IP and MPLS Networks

Iftekhar Hussain

Cisco Press

800 East 96th Street
Indianapolis, IN 46240 USA

Fault-Tolerant IP and MPLS Networks

Iftekhhar Hussain

Copyright© 2005 Cisco Systems, Inc.

Published by:

Cisco Press

800 East 96th Street

Indianapolis, IN 46240 USA

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from the publisher, except for the inclusion of brief quotations in a review.

Printed in the United States of America 1 2 3 4 5 6 7 8 9 0

First Printing November 2004

Library of Congress Cataloging-in-Publication Number: 2002115120

ISBN: 1-58705-126-5

Warning and Disclaimer

This book is designed to provide information about IP and MPLS control-plane fault tolerance. 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.

Corporate and Government Sales

Cisco Press offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales.

For more information please contact: **U.S. Corporate and Government Sales** 1-800-382-3419 corpsales@pearson-techgroup.com

For sales outside the U.S. 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 e-mail at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

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.

Publisher
Editor-in-Chief
Executive Editor
Cisco Representative
Cisco Press Program Manager
Production Manager
Development Editor
Technical Editors

Editorial Assistant
Cover Designer
Composition
Copy Editor and Indexer
Proofreader

John Wait
John Kane
Jim Schachterle
Anthony Wolfenden
Nannette M. Noble
Patrick Kanouse
Betsey Henkels
Don Banks, Rudy Figaro,
Liem Nguyen, Mark Szczesniak
Tammi Barnett
Louisa Adair
Mark Shirar
Keith Cline
Tonya Cupp



Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: 31 0 20 357 1000
Fax: 31 0 20 357 1100

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems, Inc.
Capital Tower
168 Robinson Road
#22-01 to #29-01
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the
Cisco.com Web site at www.cisco.com/go/offices.

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia • Czech Republic
Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland • Israel • Italy
Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal
Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden
Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

Copyright © 2003 Cisco Systems, Inc. All rights reserved. CCIP, CCSP, the Cisco Arrow logo, the Cisco Powered Network mark, the Cisco Systems Verified logo, Cisco Unity, Follow Me Browsing, FormShare, iQ Net Readiness Scorecard, Networking Academy, and ScriptShare are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn, The Fastest Way to Increase Your Internet Quotient, and iQuick Study are service marks of Cisco Systems, Inc.; and Aironet, ASIST, BPX, Catalyst, CCDA, CCDP, CCIE, CCNA, CCNP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, the Cisco IOS logo, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Empowering the Internet Generation, Enterprise Solver, EtherChannel, EtherSwitch, Fast Step, GigaStack, Internet Quotient, IOS, IPTV, iQ Expertise, the iQ logo, LightStream, MGX, MICA, the Networkers logo, Network Registrar, Packet, PIX, Post-Routing, Pre-Routing, RateMUX, Registrar, SlideCast, SMARTnet, StrataView Plus, Stratum, SwitchProbe, TeleRouter, TransPath, and VCO are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and certain other countries.

All other trademarks mentioned in this document or Web site are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0303R)

Printed in the USA

About the Author

Iftekhar Hussain is a technical leader within the Internet Technologies Division at Cisco Systems. For the past several years, Iftekhar has been involved in the design of high availability-related aspects of IP/MPLS networks. He brings extensive industry experience to the subjects of networking and telecommunication, including switching, traffic management, and voice delivery over packet-switched networks. He has a Ph.D. in electrical and computer engineering from the University of California, Davis.

About the Technical Reviewers

Don Banks is a distinguished engineer at Cisco Systems in the IOS Technologies Division, where he is responsible for the development of the IOS high availability architecture. He leads a cross-functional team responsible for providing high availability for Cisco IOS-based platforms with the goal of 99.999 percent availability or better using fault-tolerant techniques. Don has been responsible for the architecture for stateful switchover (SSO/NSF), in-service software upgrade (ISSU) and follow-on high-availability architecture enhancements. Don has six patents filed in the area of high availability. Throughout his career of more than 30 years in the industry, Don has been involved in the design and implementation of new operating systems, networking protocols, storage networking, fault tolerance, as well as hardware platform design while working on research for an IBM Fellow, and at such companies as Amdahl, Tandem, Apple, and Auspex.

Rudy Figaro is a technical leader at Cisco. He works in MPLS and routing, leading software design and development. Rudy has created the architectural design for MPLS high availability. Rudy also led and worked on many software development projects including MPLS IP+ATM, MPLS-TE and Fast ReRoute, GMPLS, Layer 3 and Layer 2 VPNs, and IP multitable support. Recently Rudy has worked on an architecture for virtual routers with focus on high availability, performance, and scaling. Rudy also holds a patent for "Apparatus and Methods Providing Redundant Routing in a Switched Device."

Liem H. Nguyen is a technical leader in the core IP routing development area at Cisco Systems. Before coming to Cisco in 1997, Liem spent 11 years at IBM and 2 years at start-up NetEdge Systems in RTP, North Carolina, as a network communication software developer. He is an alumnus of North Carolina State University, where he obtained his bachelor of science degree in electrical engineering in 1984, and he is quite proud of his Wolfpack, who won the 1983 NCAA Basketball Championship under the reign of Jim Valvano.

Mark Szczesniak is a technical lead engineer for the Cisco Layer 3 Services Group. He is responsible for leading forwarding code development for MPLS. Mark has been with Cisco for 10 years, since the Lightstream acquisition.

Understanding High Availability of IP and MPLS Networks

Until recently, many service providers maintained and operated separate legacy circuit-switched and packet-switched networks. Traditionally, voice services have been offered over circuit-switched networks, commonly known as *Public Switched Telephone Networks* (PSTN). On the other hand, connectivity between enterprises for *virtual private network* (VPN) data applications has been provided over packet-switched networks such as Frame Relay (FR) and *Asynchronous Transfer Mode* (ATM). Of late, many service providers are migrating legacy Layer 2 and Layer 3 services to converged *Multiprotocol Label Switching* (MPLS)-enabled IP networks.¹ This migration toward a common multiservice IP/MPLS network is driven by the necessity to reduce the *capital expenditure* (capex) and *operational expenses* (opex) of both building and operating separate network infrastructures.

This chapter describes major sources of network failures and provides an overview of techniques that are commonly used to improve availability of IP/MPLS networks. In particular, this chapter outlines mechanisms for reducing network downtime due to control-plane failures.

Reliability and Availability of Converged Networks

For service providers, maintaining highly reliable and revenue-generating legacy service offerings is extremely important. So as much as possible, they are interested in migrating legacy services on to IP/MPLS infrastructures without cannibalizing revenue from these services. During migration, they also try to keep network downtime to a minimum (for example, in the order of a few minutes per year) to keep the cost of network outages in check. For example, a 1-minute network outage that affects 100 customers could cost a service provider several hundred thousand dollars.² Therefore, it is not surprising to know that network reliability and availability rank among the top concerns of the most service providers. In short, high availability of IP/MPLS networks is a prerequisite to offer reliable and profitable carrier-class services. A well-designed network element, such as a router, facilitates the building of highly available networks and reduces the capex and opex associated with redundant network infrastructures.

Defining Key Terms

Before proceeding further, it would be useful to define some key terms.

Availability and Unavailability

The phrase “availability of a system such as a router or network” denotes the probability (with values in the 0.0 to 1.0 range such as 0.1, 0.2, and so forth) that the system or network can be used when needed. Alternatively, the phrase describes the fraction of the time that the service is available. As a benchmark, carrier-class network equipment requires availability in the range of five-nines (0.99999), which means the equipment is available for service 99.999 percent of the time.

The term *unavailability* is defined as the probability that a system or network is not available when needed, or as the fraction of the time service is not available. An alternative and often more convenient expression (because of its additive properties) for unavailability is *downtime per year*. Downtime in units of minutes per year is obtained through multiplication of unavailability values by minutes in a year (365 days in a year times 24 hours in a day times 60 minutes in an hour). Service providers commonly use yet another expression for unavailability, especially when referring to voice calls. This term is *defects per million* (DPM). DPM measures the number of defective units (or number of failed call attempts) out of a sample size of one million units (1,000,000 call attempts). DPM is obtained by multiplying unavailability by 1,000,000. From these definitions, it follows that 0.99999 availability is equivalent to 0.00001 unavailability, 5.256 downtime per year, or 10 DPM.

Reliability and Its Relationship to Availability

The phrase “reliability of a system or network” is defined as the probability that the system or network will perform its intended function without failure over a given period of time. A commonly used measure of reliability is known as *mean time between failures* (MTBF), which is the average expected time between failures. A service outage caused by a failure is represented as mean time to repair (MTTR). That is the average time expected to be required to restore a system from a failure. MTTR includes time required for failure detection, fault diagnosis, and actual repair. Availability is related to MTBF and MTTR as follows:

$$\text{Availability} = \text{MTBF} / (\text{MTBF} + \text{MTTR})$$

This relationship shows that increasing MTBF and decreasing MTTR improves availability. This means that the availability of a router can be improved by increasing the reliability of its hardware and software components. Similarly, improving the reliability of its constituent elements such as routers, switches, and transport facilities can enhance the availability of a network.

In general, reliability is just one of several factors that can influence the availability of a system. For example, in addition to reliability of constituent network elements, network availability is strongly influenced by the fault-tolerance capability of the network elements, as described in the following section.

Fault Tolerance and Its Effect on Availability

Fault tolerance describes the characteristics of a system or component that is designed in such a way that, in the event of a component failure, a backup or “redundant” component immediately can take its place with no loss of service. Fault tolerance can be provided via software, hardware, or combination of the two. The switch between the failing component and the backup component is opaque to the outside world—from the view outside the system, no failure has occurred.

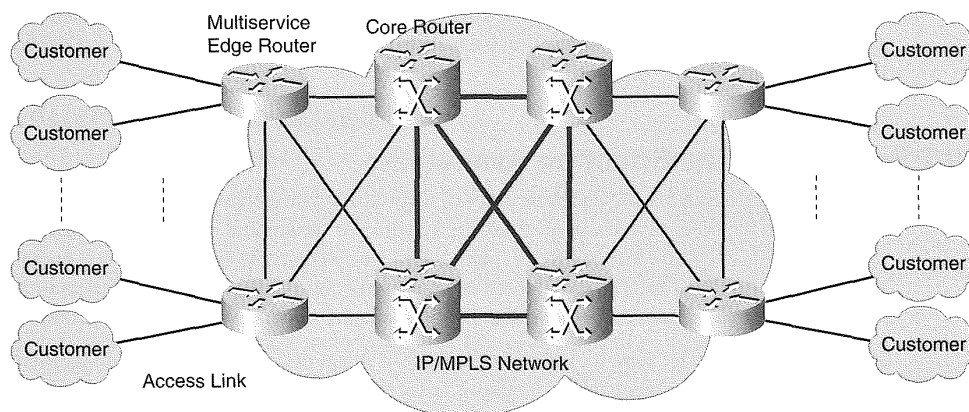
A network is said to be fault tolerant or survivable if it can maintain or restore an acceptable level of service performance during network failures. Network-level fault tolerance relies on software or hardware to quickly detect the failure and switch to a known backup path/link. The backup paths may be provided at multiple transport layers, including *wavelength-division multiplexing* (WDM), Synchronous Optical Network/Synchronous Digital Hierarchy (SONET/SDH), and MPLS.

As described in the previous section, improving MTBF can increase overall system availability. However, by using redundant components, one can reduce system downtime by orders of magnitude and get closer to the carrier-class goal of five-nines availability while keeping the MTBF and MTTR the same. The effectiveness of a redundancy scheme depends on its switchover success rate (the probability of a successful switchover from active to standby component when the active component fails). Generally, it is difficult to achieve a perfect (100 percent) switchover success rate. In practice, a redundancy scheme that can achieve a 99 percent or better switchover success rate is considered a good design.

To summarize, redundancy is one of the key building blocks for improving high availability. Redundancy not only prevents equipment failures from causing service outages, it also can provide a means for in-service planned maintenance and upgrade activities.

MPLS Network Components

An MPLS-based network consists of routers and switches interconnected via transport facilities such as fiber links (see Figure 1-1). Customers connect to the backbone (core) network through multiservice edge (MSE) routers. The backbone comprises the core routers that provide high-speed transport and connectivity between the MSE routers. An MSE router contains different types of line cards and physical interfaces to provide Layer 2 and Layer 3 services, including ATM, FR, Ethernet, and IP/MPLS VPNs.

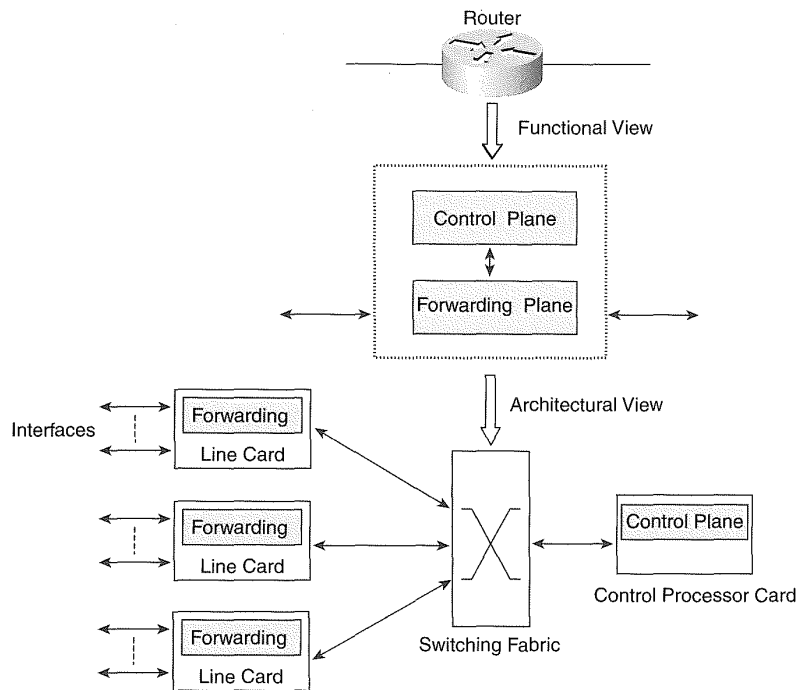
Figure 1-1 Converged IP/MPLS Network Architecture

In the incoming direction, line cards receive packets from external interfaces and forward them to the switching fabric (see Figure 1-2). In the outgoing direction, line cards receive packets from the switching fabric and forward them to the outgoing interfaces. The switching fabric, the heart of the router, is used for switching packets between line cards. The IP/MPLS control-plane software, the brain of a router, resides in the control processor card. The phrase *IP/MPLS control plane* refers to the set of tasks performed by IP routing and MPLS signaling protocols. IP routing protocols are used to advertise network topology, exchange routing information, and calculate forwarding paths between routers within (intra) and between (inter) network routing domains. Examples of IP routing protocols include *Open Shortest Path First (OSPF)*, *Intermediate System-to-Intermediate System (IS-IS)*, and *Border Gateway Protocol (BGP)*. MPLS signaling protocols are used to establish, maintain, and release *label-switched paths (LSP)*. Examples of MPLS signaling protocols include BGP, *Label Distribution Protocol (LDP)*, and *Resource Reservation Protocol (RSVP)*. The IP control plane may also contain tunneling protocols such as *Layer 2 Tunneling Protocol (L2TP)* and *Generic Routing Encapsulation (GRE)*, but these protocols are not covered in this book.

Because redundant network elements add to the overall network cost, service providers typically employ different levels and types of fault tolerance in the edge and core network. For example, the core network is generally designed to protect against core router failures through mesh connectivity. This allows alternative paths to be quickly established and used in the face of a failure. In the core, additional routers and links are used to provide fault tolerance. In contrast, on the edge, often thousands of customers are connected through a single router, and the edge router usually represents a single point of failure. The edge router is what most service providers consider the most vulnerable point of their network after the core is protected. On the edge,

instead of using additional routers and links as in the core, redundancy within the edge router via redundant control processor cards, redundant line cards, and redundant links (such as SONET/SDH Automatic Protection Switching [APS]) are commonly used to provide fault tolerance.

Figure 1-2 *Functional Components of Router Architecture*



In summary, service (to a customer) downtime can result from failure of the access port, edge links, the edge router, backbone transport facilities, or the core routers. Generally, the core network offers a higher level of fault tolerance than the edge network. The edge router is an important network element because it routes traffic to/from multiple customers to the core network. Therefore, improving the availability of edge routers is extremely important. In short, service providers are looking for truly edge-to-edge reliability, and this includes all of the edge routers as well as the core routers.

Network and Service Outages

A service is the set of tasks performed by the network upon a request from the user such as a voice call, Internet access, e-mail, and so forth. A *service outage* is the users' inability to request a new service or to continue to use an existing service because the service is either no longer available or it is impaired. As discussed previously, availability of a network strongly depends on the frequency of service outages and the recovery time for each outage. A *network outage* is the loss of network resources, including routers, switches, and transport facilities, because of the following:

- Complete or partial failure of hardware and software components
- Power outages
- Scheduled maintenance such as software or hardware upgrades
- Operational errors such as configuration errors
- Acts of nature such as floods, tornadoes, and earthquakes

Planned and Unplanned Outages

Each network outage can be broadly categorized as either "unplanned" or "planned." An unplanned network outage occurs because of unforeseen failures of network elements. These failures include faults internal to a router's hardware/software components such as control-plane software crashes, line cards, link transceivers, and the power supply or faults external to the router such as fiber cuts, loss of power in a carrier facility, and so forth. A planned network outage occurs when a network element such as router is taken out of service because of scheduled events (for example, a software upgrade).

Main Causes of Network Outages

What are the main causes of network outages? As it turns out, several culprits contribute to network downtime. According to a University of Michigan one-year reliability study of IP core routers conducted in a regional IP service provider network, router interface downtime averaged about 955 minutes per year, which translates to an interface availability of only 0.998.³ As a reference point, a carrier-class router is expected to have a downtime of only 5.2 minutes per year. The same study indicated the following percentages of causes for total network downtime:

- 23 percent for router failure (software/hardware faults, denial-of-service attack)
- 32 percent for link failures (fiber cuts, network congestion)
- 36 percent for router maintenance (software and hardware upgrade, configuration errors)
- The remaining 9 percent for other miscellaneous reasons

According to another study, router software failures are the single biggest (25 percent) cause of all router outages.⁴ Moreover, within software-related outages, router control-plane failure is the biggest (60 percent) cause of software failures. The following section provides a brief overview of various node- and network-level fault-tolerance approaches that can help to improve network availability.

Design Strategies for Network Survivability

The reliability and availability of an IP/MPLS network can be examined from two interrelated viewpoints: service and network views. The *service view* deals with satisfying customer expectations such as availability of service and other *service-level agreements* (SLA). The *network view* deals with reducing network equipment and operation costs. Because the main task of a network is to provide user services, the reliability and availability requirements for the network are driven by the service view. An effective network design seeks to satisfy service reliability and availability objectives at the minimum network equipment (capex) and operational (opex) cost.

A packet-switched network consists of interconnected network elements, including routers, switches, and transport links. Network availability depends on the reliability and availability of its network elements. In particular, fault tolerance of router hardware and software components is crucial to deliver user services with negotiated SLAs. A carrier-class router is typically expected to satisfy requirements such as the following:

- No single hardware fault should result in a loss or degradation of user traffic or a loss of control-plane and management functions.
- System downtime should be less than 5.256 minutes per year.
- Line cards, switching fabric, and control processor cards should be redundant with capability to monitor standby cards.
- The control-plane software/hardware module should not be a single point of failure, and the service (forwarding plane) should not be disrupted due to failure of the control plane.
- The router should be capable of service recovery from link/node failures.

Generally, these carrier-class availability requirements are satisfied using a combination of node- and network-level fault-tolerance techniques, as described in the sections that follow.

Mitigating Node-Level Unplanned Hardware-Related Outages

One of the most effective techniques for reducing unplanned hardware-related downtime in a router is the use of redundant hardware components, including line cards, switching fabric,

control processor cards, and physical interfaces. Three types of redundancy schemes are commonly used for this purpose:

- **One-for-N (1:N)**—There is one standby component for every N active component.
- **One-for-one (1:1)**—There is a standby component for each active component.
- **One-plus-one (1+1)**—This is similar to the one-for-one scheme except that in the case of one-plus-one, traffic is transmitted simultaneously on both active and standby components. (Traffic is generally ignored on the standby.) An example of one-plus-one redundancy is the 1+1 SONET/SDH APS scheme that avoids loss of data traffic caused by link failure.

A detailed discussion of component redundancy architectures is beyond the scope of this book.

Mitigating Node-Level Unplanned Software-Related Outages

It is apparent that reliability and stability of router hardware and software are absolutely crucial for building reliable and available IP/MPLS networks. As discussed previously, routers use redundant switching fabric, control processor cards, line cards, and interfaces to achieve node-level hardware fault tolerance. Although most routers usually have adequate hardware-component redundancy coverage, the control-plane software still remains a weak link and a prime cause of router failures.

The two most important constituents of the router software are IP and MPLS control-plane protocols. The IP control-plane component consists of IP routing protocols such as OSPF, IS-IS, and BGP, which exchange network topology information and thus help build the IP forwarding state. The MPLS control-plane component is composed of signaling protocols such as LDP, RSVP-TE, and BGP. Label-switching routers (LSR) use information provided by IP/MPLS control-plane components to construct the MPLS forwarding state. The IP forwarding state is used to transfer IP packets from an incoming port of the router to an outgoing port using a destination IP address. In contrast, the MPLS forwarding state is used for moving packets from input to output ports based on label information.

IP and MPLS forwarding tables are collectively referred to as the *forwarding plane*. Because of the time-critical nature of packet-forwarding operations, the forwarding-plane functions are typically distributed on line cards to enhance forwarding performance. In contrast, control-plane tasks are relatively less time critical and therefore often reside on the central control processor card. Because control-plane protocols constitute router intelligence, the control processor serves as host to the router's brain. Because of the pivotal importance of the control-plane functions to the router operation, a control processor is normally protected against failure through 1:1 (active and standby) redundancy.

The existing control-plane software restart and switchover behavior in routers is disruptive and therefore undesirable. When a router detects a software/hardware failure in the active control processor, it switches over to the standby and, in this process, not only restarts its control software but also resets the forwarding plane in the line cards. The end result of this behavior means disruption of data forwarding and the accompanied service outage. Consider, for

example, the restart of an IP control-plane protocol such as OSPF or IS-IS. When OSPF or IS-IS restarts, the failing router's *interior gateway protocol* (IGP) neighbors detect this restart and originate LSAs or LSPs to omit links to the restarting router. Upon receiving new LSAs or LSPs, the nonrestarting routers recompute their paths to avoid the restarting router. This shows that the original IP control-plane restart behavior causes unnecessary disruption of traffic in the restarting router, generates extra IGP control traffic, and triggers costly *shortest path first* (SPF) recomputations in nonrestarting routers. Similarly, when the MPLS control plane restarts, LDPs withdraw labels that were advertised prior to this failure. Once again, this behavior results in disruption of the MPLS forwarding. In short, one can say that control-plane restart causes instability throughout the network.

This description clearly shows that the original IP/MPLS control-plane restart behavior is totally unacceptable, particularly when you consider the fact that service providers are deploying more and more IP/MPLS networks to deliver legacy services and customers are expecting a better or comparable level of reliability and availability. Therefore, disruption of the IP/MPLS forwarding plane must be reduced to an absolute minimum. The next section outlines some approaches to achieve this goal.

Reducing Downtime Related to Unplanned Control-Plane Restart

Several types of software redundancy schemes enable you to reduce router downtime resulting from unplanned control-plane failures. One such approach (similar to the 1:1 hardware redundancy scheme) is to instantiate two identical copies of control-plane software on active and standby control processors. The two instances execute independently without any inter-instance communication, and both instances send/receive identical control packets. For example, in the incoming direction, control packets are replicated and passed on to both instances. In the outgoing direction, control packets from the standby instance are discarded.

A second scheme (a variant of the previous approach) is to instantiate two identical copies of the control plane. The two instances execute in complete lock step using inter-instance communication. When a control packet is received, it is processed in identical fashion and at the exact same instant by the active and the standby instance.

A third approach is to instantiate two copies of the control plane on active and standby control processors. The active instance executes, whereas the inactive instance does not. However, the standby instance maintains partial state of the active instance by receiving state synchronization messages. For example, the active instance of an IP routing protocol establishes sessions, exchanges routing information with peers, and helps build and maintain routing/forwarding tables. In contrast, the inactive standby instance does not exchange routing information with external peers. After switchover, the standby instance takes over, reestablishes peer sessions, and resynchronizes its state information. Another variant of the third approach maintains complete state on the standby and can switch over without having to reestablish sessions from the point of view of the neighbors. However, this variant is less scalable because it requires preservation of the complete state.

Table 1-1 describes the advantages and disadvantages of each approach.

Table 1-1 *Advantages and Disadvantages of Software Redundancy Approaches*

Approach	Advantages	Disadvantages
Approach 1 Instantiate two identical copies of the control-plane software on active and standby control processors. The two instances execute independently.	Control-plane failure and resulting switchover hidden from neighbors No requirement for changes to IP/MPLS control-plane protocols	Extra processing burden for replicating control packets Necessity to start both instances at the same time Restriction of precluding software upgrades and downgrades
Approach 2 Instantiate two identical copies of the control-plane software on active and standby control processors. The two instances execute in lock step using inter-instance communication.	Fast recovery time No necessity for protocol changes	Design complexity of synchronizing state Requirement to run two instances synchronously Restriction of precluding software upgrades and downgrades
Approach 3 Instantiate two copies of the control plane on active and standby control processors. The standby instance maintains partial state.	Allows the restarting router to continue to forward across the control-plane recovery Allows for software upgrades	Necessity to reestablish sessions and recover the control-plane state information after the restart Needs protocol extensions

The main strength of the first approach is that the control-plane failure and resulting switchover is hidden from neighbors and therefore does not require any changes to IP/MPLS control-plane protocols. However, this approach has two big drawbacks: extra processing burden for replicating control packets, and the necessity to start both instances at the same time. The second drawback is very restrictive because it precludes software upgrades and downgrades. The key advantages of the second approach are fast recovery time and absence of necessity to make protocol changes. The main disadvantages of this approach are design complexity to synchronize state and the requirement to run two instances synchronously. The latter requirement implies that, like the first approach, the second approach does not allow software upgrades. The main disadvantage of the third approach is the necessity to reestablish sessions and recover control-plane state information after the restart. This requires IP/MPLS protocol extensions and support from neighbors in maintaining their forwarding state while the restarting router comes back.

The third approach, analogous to other two approaches, allows the restarting router to continue to forward across the control-plane recovery. However, unlike the other two schemes, the third approach allows software upgrades, which is a big plus toward achieving the carrier-class availability goals.

Cisco IOS architecture is likely to adopt the third approach and its variants to provide fault-tolerant control-plane software architecture on routers. For example, in the core and on the core side of the edge routers where scalability is extremely important, the third approach can be used because it requires preserving partial control-plane state. In contrast, on the customer side of edge routers where scalability is generally not much of an issue, a variant of the third approach (completely stateful) can be used.

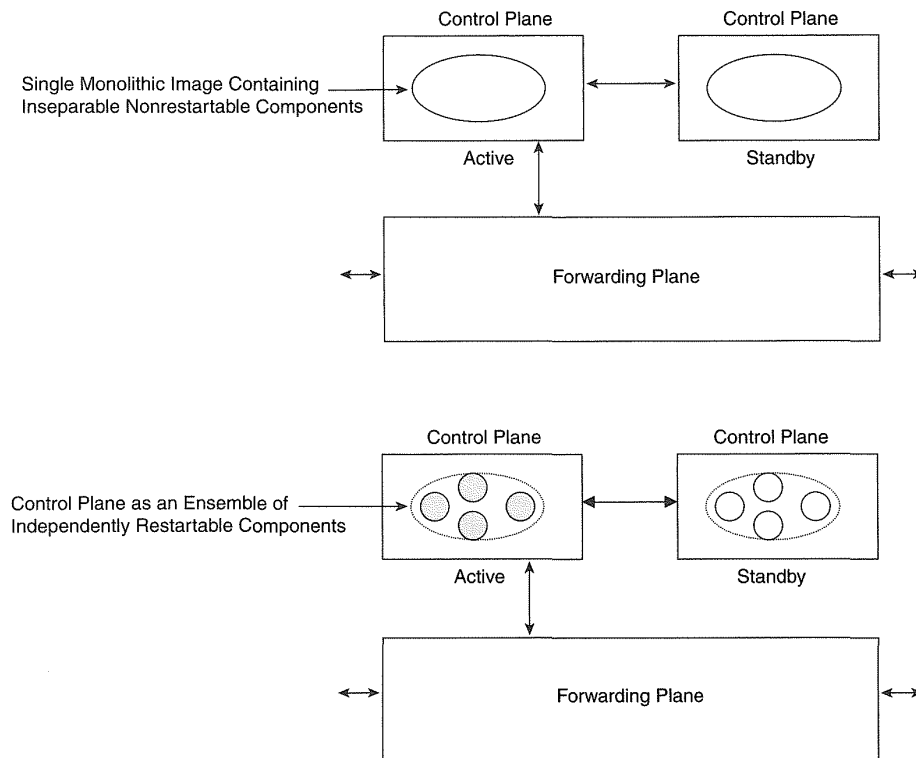
Stateful Switchover and Nonstop Forwarding

The combination of separation of control- and forwarding-plane components, 1:1 redundant control processors, and fault-tolerant control-plane software allows Cisco IOS architecture to make an automatic nondisruptive stateful control-plane switchover upon detection of software/hardware failures in the active control processor. In IOS the term *stateful switchover* (SSO) refers to the aforementioned control-plane redundancy framework that enables nondisruptive automatic SSO of the control plane upon detection of hardware/software failure. The term *nonstop forwarding* (NSF) refers to the capability of a router to continue to forward while its control plane recovers from a fault. NSF requires separation of control- and forwarding-plane functions.

Reducing Unplanned Downtime Using Component-Level Modularity and Restartability

The SSO discussion in the preceding section assumes that control-plane software executes as one or more inseparable components (or processes) that are sharing critical data structures. In that case, because various control-plane components are inseparable and incapable of restarting individually, failure in one component leads to failure of all other components and necessitates control-plane switchover. Therefore, nonrestartable components require stateful redundancy schemes and switchovers to recover from failures (see Figure 1-3).

A software component is said to be *restartable* if it is capable of recovering from fatal runtime errors. In a redundant system, a restartable component that can correctly recover from failures should not require a switchover. However, when a restartable component fails to restart correctly, it should cause a switchover to the standby to recover from failures.

Figure 1-3 *Nonrestartable and Restartable Control-Plane Components*

Hence, the component-level restartability complements and extends the SSO approach by providing additional fault-tolerance coverage. In the component-level restartability approach, following system initialization, a system management module instantiates all control-plane components and monitors their health. Upon detecting failure of a control-plane component, the system manager restarts the failed component without disturbing other components or requiring control-plane switchover. After having restarted, the process recovers its preserved state information and resumes normal operation.

It is worth noting that unlike SSO, the process-level restartability approach can be used to improve control-plane availability of routers with single as well as redundant control processors. In a nonredundant control processor scenario, the component restartability-based approach allows a router to recover from unplanned control-plane software component failures. In a redundant control processor case, a combination of SSO and component-level restartability helps improve the overall fault isolation, reliability, and availability of the control plane. In the latter case, for example, a router can recover from minor software component-level faults

without requiring switchover to the standby control processor and yet use SSO to recover from major software or hardware faults.

In summary, component restartability is not a remedy for all failure scenarios. For example, if critical data structures of the restartable component are damaged, that component will fail to restart. Complete switchover with redundant hardware components and no shared data (no shared memory) offers a higher level of fault tolerance. You should view these approaches as occurring within a spectrum of approaches to improve high availability, with their own attendant benefits and costs.

In the remainder of this book, it is assumed that the control-plane software runs as a single image containing inseparable nonrestartable components. Discussion of approaches for improving control-plane reliability and availability using component-level modularity and restartability is beyond the scope of this book.

Mitigating Node-Level Planned Outages

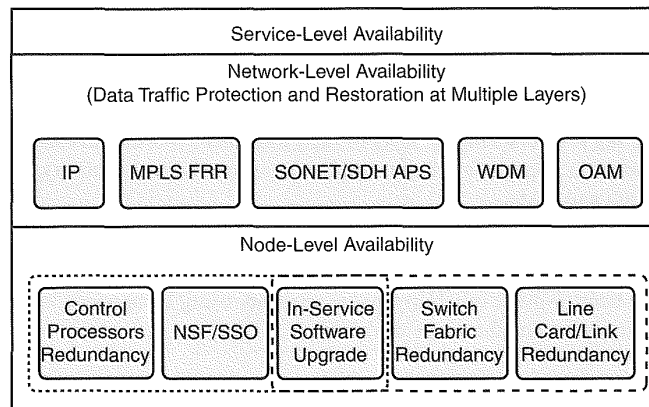
As discussed previously, planned events such as software upgrades are a big contributor to network downtime. To deliver carrier-class services, unplanned and planned outages must be reduced. The downtime due to planned events is reduced using an in-service software-upgrade mechanism that allows upgrading the router software without service disruption.

Mitigating Network Outages Against Link and Node Failures

So far this chapter has discussed strategies for reducing downtime caused by unplanned control-plane restart and planned router operations such as software upgrades. Another significant contribution to network downtime comes from link failures. The impact of transport path failures is mitigated by using multilayer protection/rerouting schemes such as SONET/SDH APS and emerging MPLS-based methods such as *FastReRoute* (FRR). With the success of MPLS deployments, the use of MPLS-based recovery schemes is also growing to provide LSP-level protection against link/node failures. Although SONET/SDH-based protection is widely deployed, protection at the lower transport layers is very coarse and can be very wasteful and expensive. In contrast, MPLS-based recovery can provide much finer granularity and presents an efficient, attractive, and complementary alternative to SONET/SDH-based protection.

Mitigating Network Outages via Effective Operation and Maintenance Mechanisms

As service providers move more and more revenue-generating services onto converged IP/MPLS networks, effective MPLS *operation and maintenance* (OAM) mechanisms become an absolute necessity to deliver carrier-class services. This is because service providers rely on robust OAM tools to quickly identify and remove network faults, reduce service downtime, and maintain a high level of network availability. A layered view of IP/MPLS availability architecture is depicted in Figure 1-4.

Figure 1-4 *Dependence of End-to-End IP/MPLS Service Availability on Node and Network Level Availability*

Improving Network Security via Fault-Tolerance Mechanisms

Network resources include routers, switches, hardware, software, data stored on line, data in transit over the network, and so forth. Network security refers to the set of measures taken to protect a resource against unauthorized access. For each resource, the key objectives of security are resource availability, data confidentiality (meaning that information is not made available or disclosed to unauthorized individuals or entities), and data integrity (meaning that information has not been modified in an unauthorized manner). Some exploits that might threaten an IP/MPLS network include attacks on control and forwarding planes, sniffing of data packets, denial-of-service (DoS) attacks, and so forth. In a DoS attack, an attacker seeks to disrupt or prevent the use of a service by its legitimate users. A DoS attack might appear in different forms such as taking network devices out of service by overwhelming the target devices with requests for service or modifying their normal behavior. A DoS attack in which the network is overwhelmed with requests for service is also known as a resource-exhaustion DoS attack. Resource-exhaustion DoS attacks can be mounted against any network resource such as forwarding plane, control plane (for example, control processor), link bandwidth, and so forth.

Because the goal of fault-tolerance mechanisms is to protect a system or network against different types of failures by improving its availability, fault-tolerance mechanisms may also be thought of as defensive techniques against malicious security threats. For example, separation of control plane and forwarding plane (as provided in the SSO/NSF framework) can be used to improve security against some attacks. This, for example, might help to limit DoS attacks against a control plane to that particular component only and might allow the forwarding-plane component to continue to function normally.

In general, to offer network services securely and reliably, security and fault-tolerance mechanisms must be built in to IP and MPLS networks. Examples of common defensive techniques against network security threats include data encryption, authentication, packet filtering, firewalls, separation of control and forwarding planes, intrusion detection, intrusion prevention, and so forth.⁵ A detailed discussion of network security mechanisms is beyond the scope of this book.

Scope of the Book

From the discussions in this chapter so far, you know that the design of carrier-class IP/MPLS networks involves reducing both unplanned and planned outages by using a variety of fault-tolerance techniques, including node-level hardware redundancy, control-plane software redundancy, MPLS-layer redundant LSPs, OAM mechanisms, and in-service software upgrades. In short, the reliability and availability of an IP/MPLS network encompasses a broad set of functional areas.

The main purpose of this book is to describe IP/MPLS control-plane fault-tolerance mechanisms that enable you to reduce downtime and improve network availability (by reducing unplanned IP/MPLS control-plane failures). Specifically, this book intends to cover three aspects of the control plane, as follows:

- IP/MPLS forwarding-plane NSF mechanisms that allow a router to continue to forward traffic while its control plane recovers from a failure
- IP/MPLS control-plane restart mechanisms that enable IP/MPLS control-plane components to restart and recover state without disrupting the forwarding plane
- Use of the previous two mechanisms to reduce downtime in the converged IP/MPLS backbone when using MPLS applications such as *traffic engineering* (TE), *Layer 2 VPNs* (L2VPNs), and *Layer 3 VPNs* (L3VPNs).

In the remainder of this book, it is assumed that the control-plane software executes as a single image containing inseparable nonrestartable components. A detailed discussion of process-level modularity and restartability is beyond the scope of this book.

Although for completeness sake fault-tolerance mechanisms such as MPLS FRR, MPLS OAM, and in-service software upgrades are briefly mentioned in a later chapter, a detailed discussion of these mechanisms is also beyond the scope of this book.

References

¹Heavy Reading Analysts, "2004 Survey of Carrier Attitudes Toward IP/MPLS Backbones and VPNs," *Heavy Reading Report*, Vol. 2, No. 4, January 2004.

²Network Strategy Partners, "Reliable IP Nodes: A Prerequisite to Profitable IP Services," White Paper, November 2002.

³ Ahuja, A., F. Jahanian, and C. Labovitz, "Experimental Study of Internet Stability and Wide-Area Backbone Failures," Proceeding of 29th International Symposium on Fault-Tolerant Computing, June 1999.

⁴ Heywood, P., and M. Reardon, "IP Reliability," *Light Reading Report*, March 2003.

⁵ Fang, L., "Security Framework for Provider Provisioned Virtual Private Networks," IETF work in progress, July 2004.

Improving the Survivability of IP and MPLS Networks

Chapters 4 through 9 described various *Internet Protocol (IP)* and *Multiprotocol Label Switching (MPLS)* control-plane restart mechanisms for improving the fault tolerance of different individual control-plane components. Generally, each IP/MPLS application involves multiple control-plane components. This chapter focuses on applications rather than individual components. Overall network availability has two aspects: node-level availability (internal to the system) and network-level availability (external to the system). Node-level availability refers to fault-tolerance mechanisms that are internal to the system and which typically are invisible to an external viewer. For example, control-plane *stateful switchover (SSO)* in conjunction with *graceful restart (GR)* mechanisms can be used to improve the node-level fault tolerance against unplanned control-plane failures. Network-level fault tolerance involves the interaction of a node with other nodes through external interfaces. Thus network-level availability encompasses link/node protection schemes at multiple layers, including *wavelength-division multiplexing (WDM)*, *Synchronous Optical Network Synchronous Digital Hierarchy (SONET/SDH)*, and MPLS. Because both node internal and external faults contribute to the total network downtime, fault tolerance at both levels is essential to have any meaningful end-to-end network availability. This chapter discusses how different IP and MPLS control-plane restart mechanisms (described in previous chapters) and node external protection schemes can be used collectively to improve the end-to-end availability of an MPLS-enabled IP network against different types of failures.

A service provider can offer to a customer two fundamentally different types of *virtual private network (VPN)* services: *Layer 3 virtual private networks (L3VPNs)* and *Layer 2 virtual private networks (L2VPNs)*. The VPN tunnel (a logical link between two devices such as *provider-edge (PE)* routers to carry customer traffic across the backbone) is a functional component that is common to both types of services. In MPLS, a tunnel corresponds to a *label-switched path (LSP)*.

In general, MPLS LSPs can be set up either along *Interior Gateway Protocol (IGP)* paths or explicit paths. The mechanisms to establish an LSP along an IGP path are described in Chapter 8, "MPLS Control Plane: Restarting LDP Gracefully." Although the main concepts of MPLS *traffic engineering (TE)* are introduced in the preceding chapter, that chapter does not discuss the mechanics of establishing MPLS-TE tunnels. This chapter draws on and extends the concepts of the preceding chapter and explains how you can use MPLS-TE mechanisms to set up TE tunnels. After describing the common building blocks such as

VPN tunnels, the chapter provides an overview of L3VPN and L2VPN applications and discusses VPN-specific mechanisms such as signaling of the tunnel demultiplexer field. This chapter also discusses ways to protect PE-PE tunnels against unplanned control-plane failures using restart mechanisms.

In addition to improving network fault tolerance against unplanned control-plane failures, it is also essential to reduce service outages that result from unplanned transport-plane faults such as link and node failures and planned events such as software upgrades. Although detailed discussion of fault-tolerance mechanisms other than control-plane restarts is beyond the scope of this book, for the sake of completeness this chapter briefly covers a few other mechanisms such as SONET/SDH protection, MPLS Fast ReRoute (FRR), MPLS *operation, administration, and maintenance* (OAM), *Bidirectional Forwarding Detection* (BFD), and in-service software upgrades.

Layer 2 and Layer 3 Services over MPLS Networks

To meet growing telecommunication and data communication demands, most enterprises and organizations find that network connectivity between different sites has become a necessity. Traditionally, service providers have offered connectivity between enterprise networks over private leased lines using *Frame Relay* (FR) or *Asynchronous Transfer Mode* (ATM) networks. More recently, with the migration of legacy Layer 2 infrastructure over a converged packet-switched network, these services are predominantly offered over an IP/MPLS backbone.

Provider-Provisioned Virtual Private Networks

The term *virtual private network* (VPN) refers to a private network offered over a service provider backbone infrastructure that is shared by multiple customers. From the perspective of an individual customer, the backbone appears to be dedicated to that customer.

The term *provider-provisioned virtual private network* (PPVPN) refers to a VPN in which the service provider participates in the management and provisioning of the VPN. A PPVPN can be further classified into L2VPNs and L3VPNs:¹

- In L2VPNs, *customer-edge* (CE) devices are interconnected using a mesh of point-to-point links (tunnels) over the shared backbone. As a result, customer addressing is not visible to devices in the service provider backbone.
- In L3VPNs, CE devices exchange routing information with the directly connected PE routers, and therefore some backbone devices know about customer addresses.

In any type of PE-based PPVPN, customer packets are carried in PE-PE tunnels.

VPN Tunnels

A VPN tunnel is a logical link between two devices such as PE routers that carries customer traffic across the backbone. One of the main reasons for using tunnels in VPN applications is to carry customer packets with nonunique addressing information between the VPN edge devices. Other important reasons for using tunnels include the need to isolate traffic from different VPN customers and to provide different *quality of service* (QoS) and security characteristics.

A tunnel that encapsulates one tunnel within another is known as a hierarchical tunnel. For example, L3VPNs and L2VPNs use hierarchical PE-PE tunnels to carry traffic from multiple VPN customers across the backbone between PE routers. In general, a tunnel is implemented by encapsulating packets within another header. More details on VPN tunneling can be found in the context of L2VPN and L3VPN discussion later in this chapter.

Tunnel Demultiplexing

Because each PE-PE tunnel can carry traffic from several customers, an individual customer's traffic within the tunnel is distinguished by using a tunnel demultiplexer field.

The format of the demultiplexer field depends on the underlying tunneling protocol and can be based on the following:

- IP header
- MPLS label
- Layer 2 Tunneling Protocol (L2TP) session ID^{2,3}
- Generic Routing Encapsulation (GRE) key field⁴

In the case of MPLS, a tunnel corresponds to a *label-switched path* (LSP) and a hierarchical tunnel corresponds to nesting multiple LSPs within a single LSP using label stack. (See Chapter 3, "MPLS Forwarding Plane: Achieving Nonstop Forwarding.") In this case, the demultiplexer field consists of another MPLS label known as a *VC label* or *inner label*.

In general, a hierarchical PE-PE tunnel might have two or more MPLS labels (label stack). Without loss of generality, in this chapter only the top and bottom labels are considered. The top label is used to forward labeled packets across the backbone and is meaningful to *label-switched routers* (LSRs) along the tunnel's path, including PEs and intermediate LSRs. In contrast, the bottom label is meaningful to the tunnel endpoints, namely, the PEs. In L3VPN and L2VPN applications, the bottom label allows traffic from multiple LSPs to be carried in a single tunnel across the backbone and provides a mechanism for distinguishing one nested LSP from others.

In the remainder of this chapter, the term *tunnel label* refers to the top label, and the term *VPN label* refers to the bottom label. The term *L3VPN label* refers to the bottom label for L3VPN applications. Similarly, the term *L2VPN label* refers to the bottom label for L2VPN application.

Signaling of the Tunnel Labels and VPN Labels

In MPLS, a tunnel can be established either along a hop-by-hop or explicitly routed path. In general, tunnel labels corresponding to IGP routes are distributed either via *Label Distribution Protocol* (LDP) or RSVP-TE. For example, as discussed in Chapter 8, LDP can be used to establish LSPs along IGP paths. Similarly, as discussed in Chapter 9, “MPLS Control Plane: Restarting RSVP-TE Gracefully,” RSVP-TE can be used to establish LSPs along explicit paths.

As far as signaling of the VPN labels is concerned, depending on L2VPN or L3VPN applications, different label distribution protocols can be used for this purpose. For example, in an RFC 2547-style L3VPN, L3VPN labels for customer addresses are distributed using *Border Gateway Protocol* (BGP). (See Chapter 7, “MPLS Control Plane: Restarting BGP with MPLS Gracefully.”) In contrast, L2VPN labels can be signaled using LDP (or BGP).

NOTE The L2VPN label-distribution option using BGP is not covered in this chapter.

NOTE A *service* is the set of functions or tasks performed by the network on request from the user such as voice call, e-mail, and Internet access. A *service outage* refers to user inability to request new services or to continue to use an existing service because the service is no longer possible or is impaired.

Service Attributes Related to Network Availability

The following list outlines the important attributes of a service that need to be considered from the perspective of network reliability and availability:

- Importance (mission critical, critical, noncritical)
- Failure impact (number of users affected)
- Outage duration
- Failure rate (frequency of the failure)⁵

For example, a mission-critical service might not tolerate lack of service greater than a few milliseconds, whereas for noncritical services, loss of service in the order of a few minutes might be acceptable. A network outage is the loss of network resources such as routers, switches, and links caused by complete or partial failures of hardware/software components in routers, link failures resulting from fiber or cable cuts, operational errors, or acts of nature such as earthquakes and floods. Depending on the location and severity of faults and the network design, loss of network resources might cause partial or complete service outage, reduce available bandwidth, or just reduce network protection (for example, reduce redundant paths) without having any effect on service.

Network Fault-Tolerance Techniques

For your convenience and to provide the proper context for the flow of concepts being discussed in this chapter, this section summarizes the main points from Chapter 1, "Understanding High Availability of IP and MPLS Networks." A network is said to be a fault-tolerant or survivable network if it can maintain or quickly restore an acceptable level of performance during network failures. The design of fault-tolerant networks requires the use of a variety of techniques, including the following:

- Redundant hardware components (for example, redundant switch fabrics, line cards, links, control processors, and so forth)
- Redundant software components (such as redundant control-plane software), SONET layer and MPLS layer redundant paths for protection against link and node failures
- Effective OAM mechanisms to rapidly detect and recover from faults
- In-service software upgrades to reduce downtime due to planned software upgrades

MPLS Traffic Engineering

The objective of this section is to provide an overview of MPLS-TE functional components and discuss various deployment scenarios. The present discussion draws heavily on the concepts discussed in Chapter 9.

TE mainly deals with improving the operational efficiency of a network by effective placement of data traffic flows on the underlying network infrastructure. Currently, several techniques are used in IP/MPLS networks for this purpose, including manipulation of IGP metrics and establishing LSPs along explicit paths. This chapter covers MPLS-TE techniques that allow the establishment of tunnels along explicit paths (either specified manually or computed dynamically using constraint-based routing). Chapter 9 covers concepts related to the RSVP-TE protocol and graceful restart.

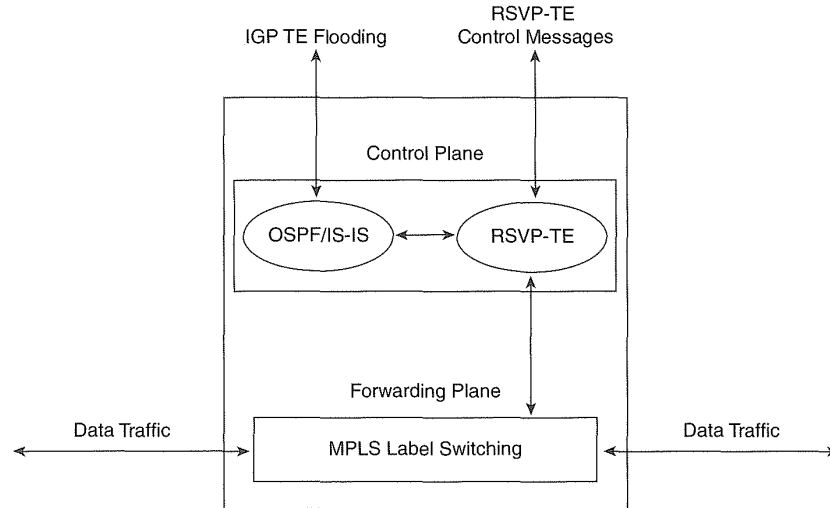
In conventional IGP protocols, the shortest paths correspond to routes with minimum additive scalar link cost (or link metric). Because the conventional IGP path-selection process does not consider link attributes such as available bandwidth, multiple shortest paths can converge on a link and cause it to be overutilized, while at the same time links along alternative but longer routes remain underutilized. *Constraint-based routing* refers to routing algorithms that compute their paths subject to a set of constraints such as bandwidth, delay, resource class, and so forth.

Because conventional IGP protocols are incapable of propagating link attributes such as available bandwidth and resource class, extensions have been made to IGP protocols that allow propagation of additional link attributes as opaque *link-state advertisements* (LSAs).^{6,7} Each router uses opaque LSAs that are propagated throughout the area to advertise the router's local link state. A collection of such opaque LSAs is known as an extended link-state database or TE topology database. Unlike a regular IGP topology database, the TE topology database contains additional information about link attributes that enables computation of *constrained shortest path first* (CSPF) routes.

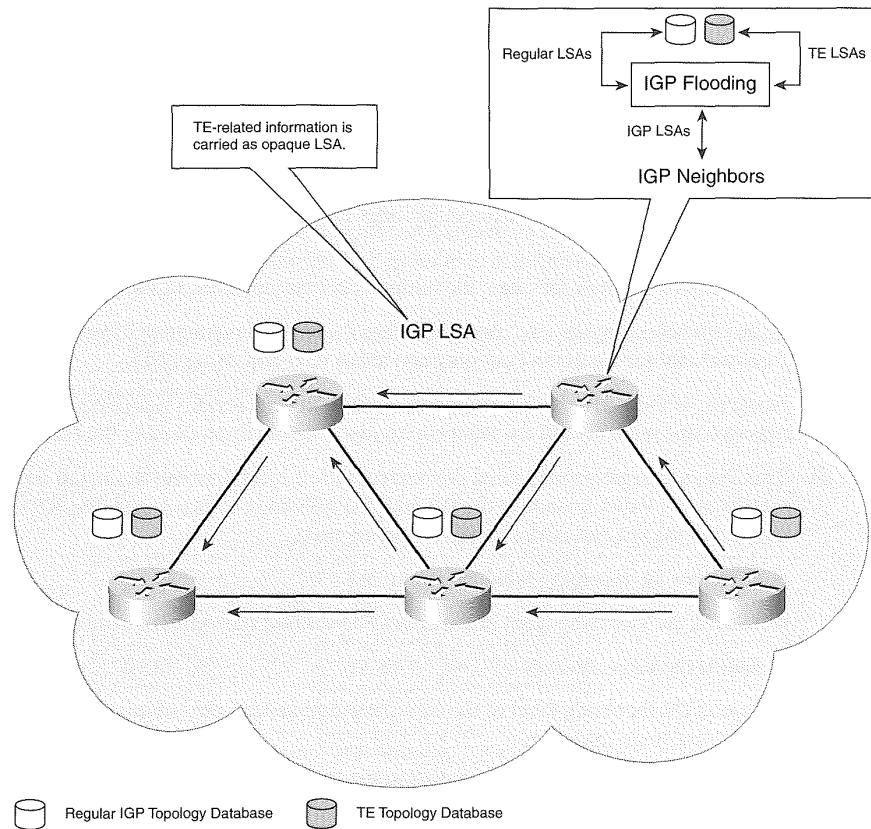
MPLS-TE Functional Modules

MPLS-TE uses enhanced IGP protocols in the IP control plane, RSVP-TE in the MPLS control plane, and MPLS label switching in the forwarding plane (see Figure 10-1). An MPLS-TE tunnel is unidirectional (head to tail) and has three parts: the head end, the midpoint, and the tail end. An MPLS-TE application contains several functional modules such as head-end control, link management, TE topology database, and path calculation.

Figure 10-1 *MPLS-TE Control-Plane and Forwarding-Plane Components*



The head-end control module is used to manage MPLS-TE tunnels that originate on the node. The link management module is primarily responsible for managing local link resources, including available bandwidth at each of eight priority levels, optional link attributes (coloring), and optional link administrative weight. The link management module provides the latest link information to the IGP for propagating as opaque LSAs. The IGP and TE topology database interactions are depicted in Figure 10-2. The MPLS-TE topology database is a link-state database that contains network topology information built using opaque LSAs from the IGP. The path calculation module exists on nodes that might function as the head-end for MPLS-TE tunnels and on midpoints when the loose *explicit route object* (ERO) needs to be expanded. The basic function of the path calculation module is to compute the explicit path using the TE topology database.

Figure 10-2 IGP and TE Topology Database Interactions

Establishment of an MPLS-TE Tunnel

The MPLS-TE tunnel establishment process begins with the configuration of tunnel parameters such as destination IP address, bandwidth, affinity, and so forth. The tunnel establishment process proceeds as follows:

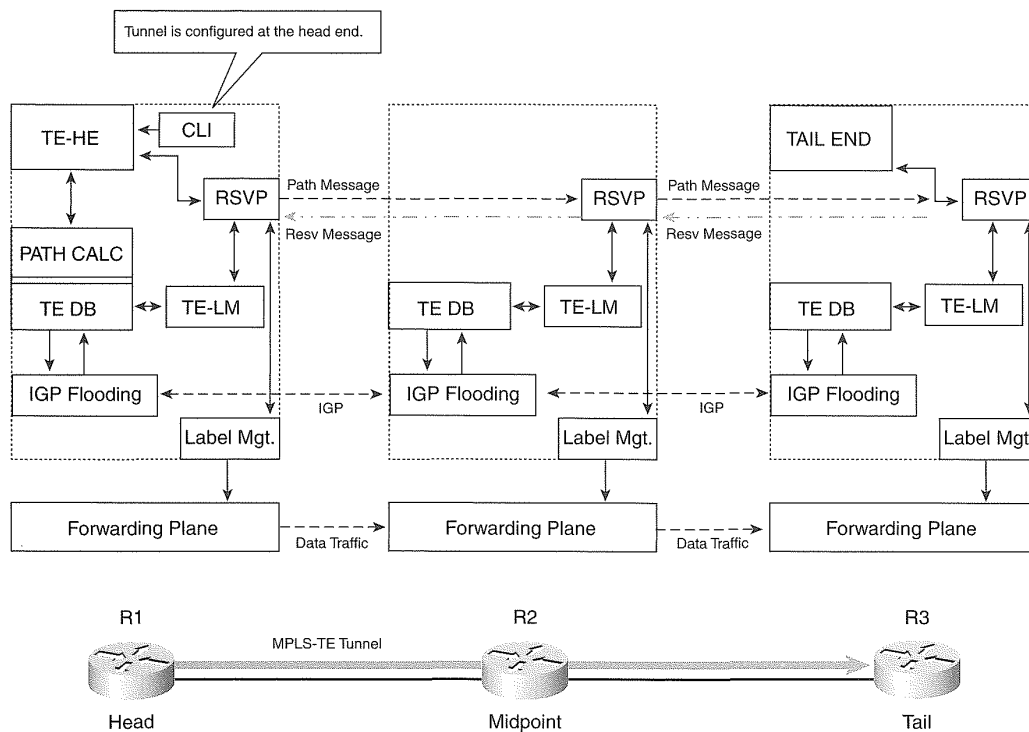
1. The head-end control passes tunnel parameters to the path calculation module.
2. The path calculation module then computes the path subject to specified constraints and returns the selected path as an ERO.
3. The head-end control hands over the ERO to RSVP-TE to establish the tunnel.

4. RSVP-TE performs a *connection admission control* (CAC) to check whether the requested bandwidth is available on the outgoing link in the appropriate-priority bandwidth pool. The link management module maintains up to eight priority bandwidth pools on each link.
5. If the CAC bandwidth check passes, the link management module sets aside the requested bandwidth from the free bandwidth pool to the allocated pool and informs the RSVP-TE module.
6. RSVP-TE builds the Path message, includes the ERO, and forwards the message to the next hop.
7. On receipt of the Path message, the midpoint router performs a CAC check.
8. If the CAC check passes, the router allocates the requested amount of bandwidth from the free pool and forwards the Path message to the next hop using ERO.
9. However, if the CAC check fails, the midpoint router generates a path error (PathErr) message upstream toward the head end to report errors in processing the Path message.
10. This Path message is propagated downstream itself until it arrives at the tail end. On receipt of the Path message at the tail end, the following events occur. The RSVP-TE module of the tail end performs the following tasks:
 - a. Allocates the incoming label for the tunnel
 - b. Installs the incoming label in the *Label Forwarding Information Base* (LFIB)
 - c. Builds the Resv message
 - d. Inserts the incoming label of the Resv message in the Label object and forwards the Resv message upstream
11. When the Resv message arrives at the upstream (midpoint) LSR, the midpoint checks whether the previously allocated bandwidth is still available.
12. If the bandwidth check passes, the midpoint performs the following tasks:
 - a. Reserves the bandwidth
 - b. Allocates an incoming label
 - c. Installs the remotely received outgoing and locally assigned label in the LFIB
 - d. Inserts the locally assigned incoming label in the Label object
 - e. Forwards the message to the upstream LSR

13. If the bandwidth check fails, the midpoint sends a reservation error (ResvErr) message downstream. The reception of a ResvErr message at a downstream node because of a CAC failure causes that node to create a local blockade state using a timer. When a blockade time expires, the associated blockade state is removed. (Refer to RFC 2205 for more information about the use of blockade states.)

Eventually, the Resv message arrives at the head-end LSR. On receipt of the Resv message, the head-end LSR checks whether the previously allocated bandwidth is still available. Assuming this is the case, the head end now reserves the bandwidth, installs the remote-received label in the LFIB, and informs the control module that the tunnel has successfully been established. The head-end control signals the IGP module to start forwarding data traffic onto the tunnel (see Figure 10-3).

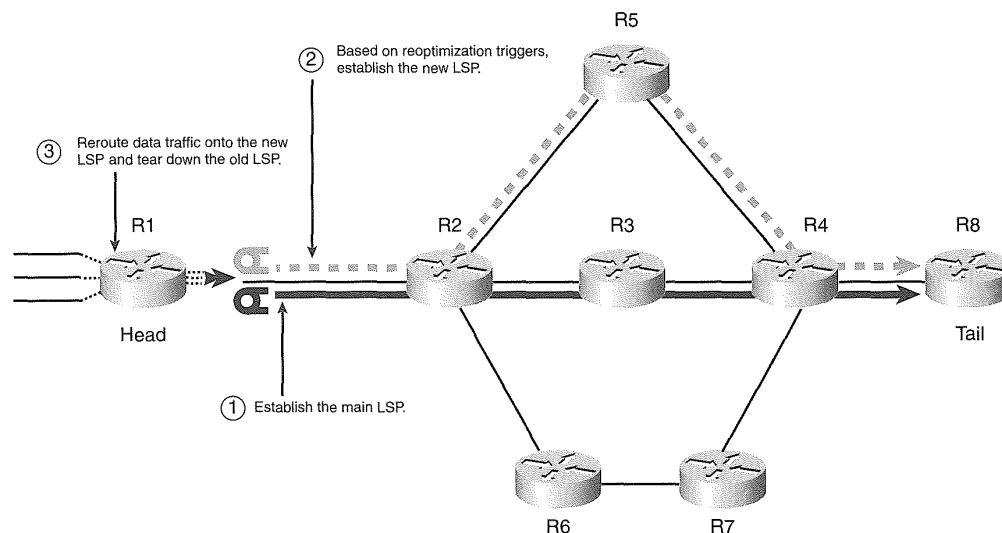
Figure 10-3 Interaction of MPLS-TE Functional Components During the Tunnel Establishment Process



MPLS-TE Tunnel Reoptimization

After a tunnel is established, topology changes or other events trigger path recalculation that might render the existing LSP path no longer optimal with respect to specified constraints. The events that could trigger path recalculation include the periodic expiration of the reoptimization timer and the forcing by an administrator of changes to the reoptimization or bandwidth of the tunnel. The objective of tunnel reoptimization is to find a better path in terms of lower IGP cost, if such a path exists, and to reroute the tunnel along the new path. To avoid disrupting traffic on the existing tunnel, another tunnel is first established along the new path before tearing down the old tunnel. This approach is commonly referred to as *make-before-break* (see Figure 10-4).

Figure 10-4 MPLS-TE Tunnel Reoptimization Using the Make-Before-Break Approach



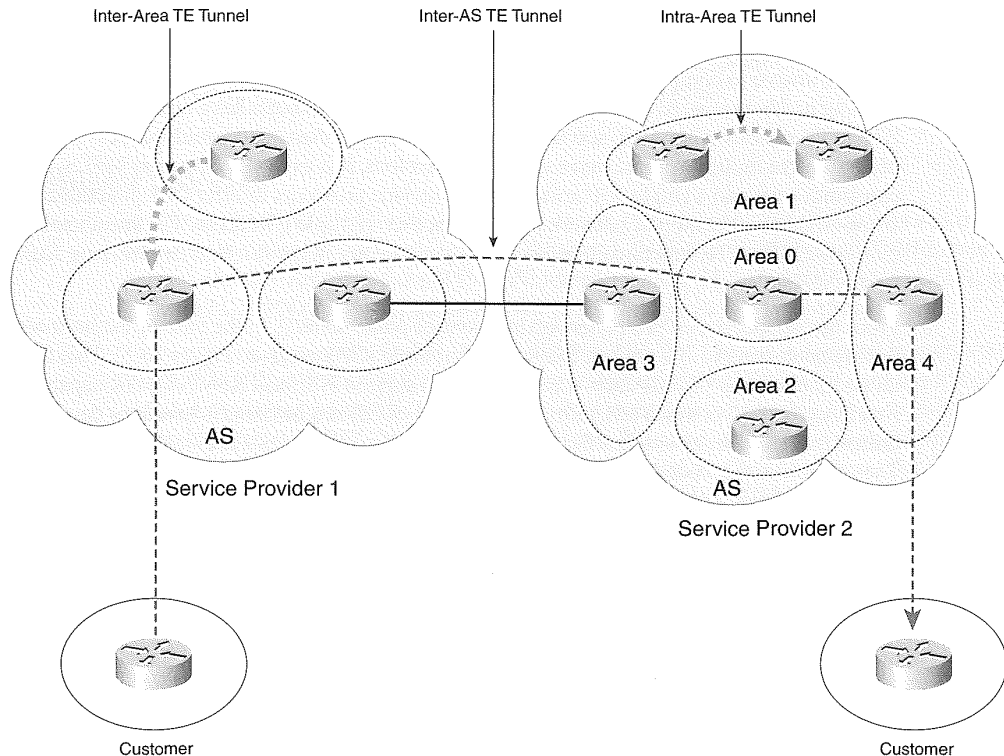
Protecting MPLS-TE Tunnels Against Control-Plane Failures

IGPs are used to distribute topology and routing information between routers within an *autonomous system* (AS). This fact has been previously discussed in Chapters 4 and 5, which are titled respectively, "Intradomain IP Control Plane: Restarting OSPF Gracefully," and "Intradomain IP Control Plane: Restarting IS-IS Gracefully." *Open Shortest Path First* (OSPF) and *Intermediate System-to-Intermediate System* (IS-IS) are two of the most widely deployed IGPs. For the purpose of routing scalability, each AS is partitioned into several smaller networks, known as OSPF areas or IS-IS levels. Within a given IGP area, routing is flat, which means each router within the area has the complete knowledge of the topology and routing information in the area. Across different IGP areas, however, routing is hierarchical, which

means that routers inside a particular area have a partial knowledge of the routing information of other areas.

With regard to MPLS-TE control-plane failures, you need to consider three deployment scenarios: intra-area, inter-area (or intra-AS), and inter-AS. (See Figure 10-5 for an illustration of each of the three cases.)

Figure 10-5 MPLS Intra-Area, Inter-Area, and Inter-AS Traffic Engineering



Intra-Area MPLS Traffic Engineering

This is the simplest case, in which both the head-end and tail-end LSRs of a TE LSP reside in the same IGP area, and thus TE LSP spans a single IGP area (see Figure 10-5). To avoid service outage on an intra-area TE tunnel that is caused by RSVP-TE control-plane failures, all nodes along the TE LSP path must support *nonstop forwarding* (NSF) and RSVP-TE GR capabilities. (For details on the RSVP-TE GR mechanism, see Chapter 9.)

Inter-Area or Intra-AS MPLS Traffic Engineering

Currently many network operators utilize MPLS-TE capabilities to deliver end-to-end Layer 2 and Layer 3 services with strict QoS guarantees. Using MPLS-TE optimizes network resources and provides fast recovery (in the order of milliseconds) against link and node failures. However, the existing set of MPLS-TE capabilities is confined to a single IGP area.

An undesirable side effect of the hierarchical routing (hierarchical routing that partitions the network into areas) is the loss of routing information for inter-area destinations. Only aggregated routing and topology information is propagated across an area boundary. Therefore, due to lack of knowledge about complete routing information, head-end LSRs in a particular area are unable to compute constrained-based optimal TE tunnels paths to the tail-end LSRs residing in other areas. (Constrained-based optimal TE tunnels paths have minimum end-to-end IGP or TE metrics.) However, service provider backbones often stretch multiple IGP areas and often traverse multiple autonomous systems. Therefore, it is essential to extend MPLS-TE capabilities beyond a single IGP area to enable service providers to offer QoS guarantees and allow efficient utilization of network resources across multiple IGP areas. The specification of inter-area MPLS-TE functionalities is currently underway in the *Internet Engineering Task Force* (IETF).⁸

From the perspective of protection against control-plane restarts, in the case of the inter-area deployment scenario, TE tunnels extend over multiple IGP areas (see Figure 10-5). This means all nodes along the TE LSP inter-area path should support NSF and RSVP-TE GR capabilities (see Chapter 9). It is worth reiterating that the goal of GR mechanisms is to reduce detrimental effects (such as loss of data traffic) due to control-plane failures that are internal to a node. These mechanisms do not address protection against failures external to the node, such as link and node failures along the tunnel path. To protect against link and node failures, other techniques such as SONET/SDH automatic protection switching and MPLS fast rerouting are used, as discussed later in this chapter.

Inter-AS MPLS Traffic Engineering

So far, this chapter has discussed using MPLS-TE functionalities within a single IGP area or across multiple IGP areas within a single AS. Many times, however, MPLS-TE tunnels need to traverse multiple autonomous systems that are part of one or more service providers' networks.⁹ This type of deployment scenario calls for inter-AS TE capabilities in order to optimize network resources and deliver negotiated service-level agreements (SLAs).

With respect to control-plane failures, in the case of inter-AS deployment scenarios, TE tunnels extend over multiple autonomous systems (see Figure 10-5). Therefore, all nodes along the TE LSP path must support NSF and RSVP-TE GR capabilities. Additional recovery methods such as fast rerouting may still be required to protect data traffic against link/node failures.

Layer 3 Virtual Private Networks

An L3VPN interconnects customer devices (routers or hosts) at multiple sites and allows them to communicate using IP addresses. L3VPNs can be broadly categorized as VPNs that are either CE or PE based.

CE-Based L3VPNs

In CE-based L3VPNs, CE devices are connected through CE-to-CE tunnels such as L2TP, IPSec, and MPLS. The service provider manages and provisions the CE devices. Devices in the backbone, however, do not know about customer IP addresses.

PE-Based L3VPNs

In PE-based L3VPNs, PE devices exchange routing information with CE devices and thus know about customer IP addresses. Like CE-based L3VPNs, with PE-based L3VPNs customer packets are also transported over PE-to-PE tunnels. Unlike CE-based L3VPNs, however, with PE-based L3VPNs, customer addresses are visible to devices in the backbone.

There are two approaches for PE-based L3VPNs: RFC 2457 based and *virtual router* (VR) based.¹¹ In VR-style L3VPNs, the PE router emulates a separate logical router, runs a separate instance of routing protocols, and maintains a separate forwarding table for each customer VPN. In contrast, in RFC 2547-style L3VPNs, the PE router maintains a separate forwarding table for each customer VPN, but runs only a single instance of routing protocols.^{7,10} The discussion in this chapter is restricted to the RFC 2547-style L3VPNs, in which the backbone consists of an MPLS-enabled IP network.

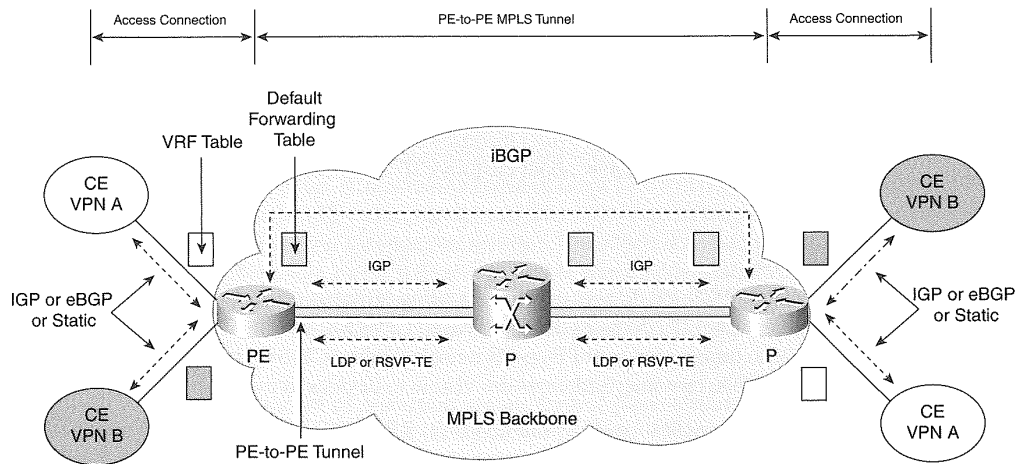
PE-Based L3VPN Reference Model

Figure 10-6 depicts the reference model for RFC 2547-style L3VPNs.

Different entities in the L3VPN reference model are briefly described herein. An access connection provides connectivity between a CE and a PE device. An access connection can be a physical (such as Ethernet interface) connection or logical (such as an IP tunnel, ATM virtual connection, or Frame Relay DLCI). The fact that a CE can connect to a PE device through a Layer 2 access connection (such as Ethernet or ATM) might seem to blur the differences between L2VPN and L3VPN services. However, you can easily make this distinction by recognizing that in the case of L3VPNs, when a CE device is connected to a PE device via a Layer 2 access connection, the PE still provides only an IP service (a Layer 3 service) to the customer. This means that in the case of L3VPNs, Layer 2 access connections must terminate at the PE devices. For example, in such cases, the PE extracts the customer's IP packets from any Layer 2 encapsulation and then sends them over PE-to-PE tunnels across the backbone. The

access connection that carries IP packets from a CE router to a PE is known as the incoming (ingress) access connection. The access connection that carries IP packets from a PE router to a CE is known as the outgoing (egress) access connection.

Figure 10-6 Reference Model for a Layer 3 PE-Based VPN with an MPLS Backbone



The service provider's backbone consists of interconnected PEs and *provider* (P) routers. The PE routers directly attach to CE devices, whereas P routers connect only to P and PE routers. A PE router may connect to CE routers belonging to different VPNs.

VPN Routing and Forwarding Tables

To provide per-VPN separation of routing information and to avoid erroneous forwarding of packets from one VPN to another, each PE router maintains a separate *VPN routing and forwarding* (VRF) table for each customer VPN. An access connection may be associated with one or more VRFs through configuration. If an IP packet arrives over an access connection that has been associated with a VRF, the packet is forwarded using the related VRF table.

Each PE router also maintains a default forwarding table that is used to forward packets from neighboring P or PE routers and packets arriving on access connections that have not been associated with any VRFs. PEs learn CE routes using IGP, external BGP (eBGP), or static routing. PE routers use *internal BGP* (iBGP) to distribute customer routes to each other. The customer routes are generally allocated from private address space. However, BGP requires IP addresses to be globally unique. Therefore, to disambiguate customer private addresses, a VPN-IPv4 address family is formed by concatenating a *route distinguisher* (RD) and the regular IPv4 address.

Within the service provider's network, the customer IP address is identified by a VPN-IPv4 address. Each VRF table is configured with a default RD, which is used for exporting a VPN-IPv4 route to BGP. In other words, PE translates customer addresses learned over PE-CE links into VPN-IPv4 addresses using a configured RD. The PE then exports VPN-IPv4 routes into BGP for distributing to other PEs. When there are multiple routes to a particular VPN-IPv4 address prefix, BGP chooses the best route based on the BGP decision process. (See Chapter 6, "Interdomain IP Control Plane: Restarting BGP Gracefully.") Note, however, that PE does not leak customer routes to the IGP in the backbone.

In summary, each VRF table is populated from two main sources: routes learned over PE-CE links, and VPN-IPv4 routes learned from another PE router via BGP. The distribution of routing information is controlled using BGP route filtering. Every VRF is associated with one or more *route target* (RT) attributes. An RT uniquely identifies a VPN or set of VPNs to which the route should be distributed. When a PE creates a VPN-IPv4 route, it associates one or more RT attributes (referred to as export route targets) with the route. These attributes are then carried as BGP extended-community RTs. Similarly, there is a set of RTs (referred to as the import route targets) that a PE router uses to determine whether a route received from another PE router via BGP could be imported in the VRF associated with a particular VPN site.

A route learned from a CE router over a particular access connection is eligible for inclusion in the associated VRF. However, a route learned from another PE router is eligible for inclusion in a particular VRF if and only if the RT attribute of the learned route matches with one of the import route targets of that VRF that has been configured. Unlike VRFs, routes learned via IGP in the backbone are used to populate the default routing table.

PE-to-PE Tunnel

As mentioned previously in this chapter, a VPN tunnel is a logical link between two PE routers and can be used to deliver customer data packets across the backbone. The tunnel can follow an IGP or an explicit route. In the case of MPLS, PE-to-PE tunnels are established using LDP or RSVP-TE. MPLS allows multiplexing (nesting) of LSPs within other LSPs. Because a single PE-to-PE tunnel can be used to carry traffic from several VPN customers, a particular unit of VPN traffic within the tunnel is distinguished using a demultiplexer field, which consists of another MPLS label (referred to as inner label). When a customer data packet arrives at a PE, the PE encapsulates it with the appropriate MPLS label (the L3VPN label). The labeled packet is further encapsulated with another MPLS label (known as a tunnel label). In the case of a hop-by-hop routed VPN tunnel, the tunnel label is distributed using LDP. For an explicitly routed PE-PE tunnel, the tunnel label is distributed via RSVP-TE. The following section describes the method for signaling L3VPN labels.

Distribution of L3VPN Labels

How are the L3VPN labels distributed? Each VPN route is assigned an MPLS label by the concerned PE router. When BGP distributes a VPN route, it also carries an MPLS label for that route. As discussed in Chapter 7, the MPLS label for a VPN-IPv4 address is carried in *Multiprotocol Reachable Network Layer Reachability Information* (MP_REACH_NLRI). Therefore, the PE pushes not just one but two labels (or label stack) to customer data packets. The tunnel label is used for forwarding the labeled packet across the backbone LSRs. The L3VPN label is used to select the outgoing access connection in the egress PE.

PEs can distribute VPN-IPv4 routes and the associated MPLS label using an iBGP connection between them, or they can have an iBGP connection to a *route reflector* (RR). As discussed in Chapter 7, the RR is used to avoid iBGP session mesh and yet ensure propagation of BGP routing information to all BGP speakers inside an AS. The RR transparently reflects the label information along with the VPN-IPv4 route. For a more detailed description of L3VPN operation, refer to the L3VPN architecture specification.¹¹

IPv6-Based L3VPN Services

To support IPv6-based L3VPN services, the RFC 2547b that is based on the method described in the previous section must be extended. For this purpose, the IETF has defined an IPv6 VPN address family and a corresponding route-distribution method.¹² A VPN-IPv6 address is a 24-byte quantity that begins with an 8-byte *route distinguisher* (RD) and ends with a 16-byte IPv6 address. Similar to the IPv4 case, a labeled IPv6 VPN route is encoded as MP_REACH_NLRI and distributed using BGP, but with different values for the *Address Family Identifier* (AFI) and *Subsequent Address Family Identifier* (SAFI) fields. For IPv6, the AFI field is set to 2, and the SAFI field is set to 128. Like the IPv4 case, with IPv6 VPNs, each PE maintains separate VRF tables for each IPv6 VPN.

Protecting L3VPN Service Against Control-Plane Failures

In the context of protection against control-plane failures, three deployment scenarios should be examined for L3VPNs, as follows:

- Single-AS backbone
- Multi-AS backbone
- Carrier supporting carriers (CSC)

Single-AS MPLS Backbone

With this, the simplest deployment scenario, two sites of a VPN attach to PE routers that are in the same AS (see Figure 10-6). In this case, an end-to-end data traffic path can be partitioned into the PE-CE access connection segment and into the PE-PE tunnel segment across the

backbone. The IP and MPLS control-plane components that can fail along the data traffic path include the following:

- IGP (or eBGP) on the PE-CE link
- LDP (or RSVP-TE) on PE and P routers
- iBGP on PE routers
- IGP on PE and P routers

To avoid traffic disruption caused by IP control-plane failure on the PE-CE link, both PE and CE routers must be able to preserve the IGP (or BGP) forwarding state across the restart and support relevant GR extensions (see Chapters 4, 5, and 6). Similarly, along the PE-PE segment, all LSRs must be able to preserve the MPLS forwarding state and IP forwarding state (when applicable) across the individual MPLS control-plane component restart (such as LDP, RSVP-TE, BGP) that created the corresponding label bindings. When different control-plane components are tightly coupled (for example, whenever IGP restarts, BGP would also restart and vice versa), to reduce overall detrimental effects of control-plane restart in the AS, it is essential that all restarting control-plane components support the pertinent GR capabilities. Other recovery techniques such as fast rerouting are needed for protection against link or node failures.

Multi-AS MPLS Backbone

In this deployment scenario, two sites of a VPN are connected to different autonomous systems (see Figure 10-7). The multi-AS backbone enables a service provider to span multiple autonomous systems and offer L3VPN service in different geographic areas. In this case, PE routers use iBGP to redistribute VPN-IPv4 routes and the associated MPLS labels to an *autonomous system border router* (ASBR) directly or through an RR of which an ASBR is a client. Between autonomous systems, labeled VPN-IPv4 routes are redistributed using eBGP. For example, the ASBR from the first AS uses eBGP to redistribute those labeled VPN-IPv4 routes to an ASBR in the second AS, which in turn distributes them to the PE routers in that AS, or to another ASBR.

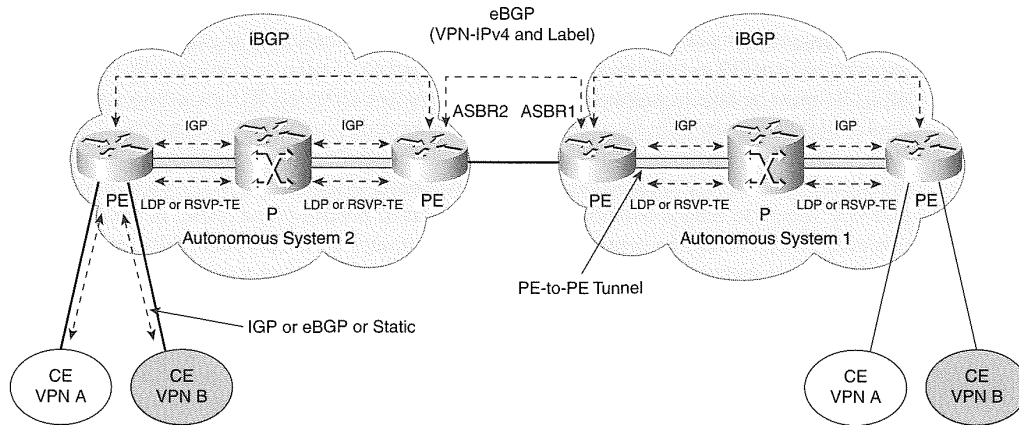
The use of eBGP to redistribute labeled VPN-IPv4 routes from AS to neighboring AS ensures loop-free exchange of routing information between autonomous systems. In this case, a data traffic path can be partitioned into the PE-CE attachment circuit segment, the PE-to-PE (or PE-ASBR) tunnel segment within the AS and ASBR-ASBR segment between two neighboring autonomous systems. The IP and MPLS control-plane components that can fail along the data traffic path include the following:

- IGP (or eBGP) on the PE-CE link
- LDP (or RSVP-TE) on PE and P routers

- iBGP on PE routers
- IGP on PE and P routers
- eBGP on ASBRs

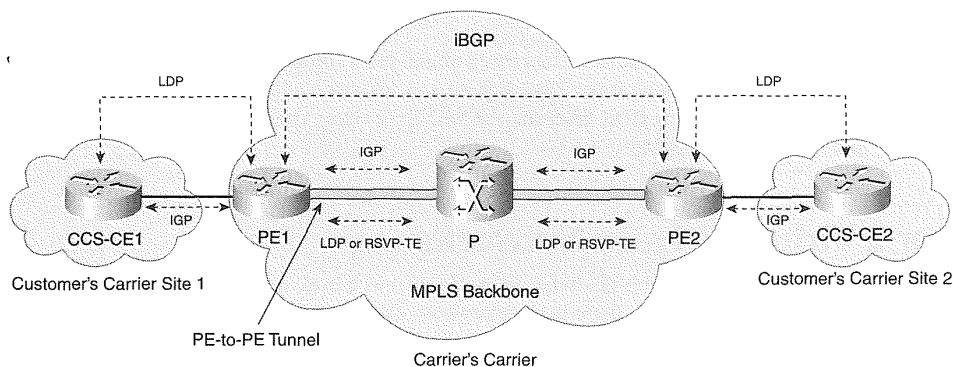
Therefore, all nodes along the data traffic path must be able to support the relevant control-plane GR capabilities.

Figure 10-7 Multi-AS Deployment Scenario for Layer 3 PE-Based VPN



Carrier Supporting Carrier (CSC)

A carrier's carrier is an L3VPN deployment scenario in which a backbone service provider (referred to as a *carrier's carrier*) offers VPN service to another service provider (referred to as a *customer carrier*) that in turn offers VPN services to its customers (see Figure 10-8). To offer carrier's carrier VPN service using RFC 2547 procedures, it is essential that the CE routers support MPLS. In this deployment scenario, LDP is used to distribute labels for routes between CE and PE routers. Thus, CE routers receive labels from the PE routers and send (receive) labeled packets to the PE routers.

Figure 10-8 Carrier's Carrier Deployment Scenario for a Layer 3 PE-Based VPN

As discussed previously, in a single-AS deployment scenario, packets between CE-PE routers are not labeled. As before, the PE routers use iBGP to redistribute labeled VPN-IPv4 routes between them (for example, when PE2 learns a route from CE2 via IGP). This is the process:

1. PE2 translates the learned IPv4 address to a VPN-IPv4 address by attaching the appropriate RD.
2. PE2 assigns a label (incoming label L2) and installs it in the MPLS LFIB.
3. PE2 then advertises the labeled VPN-IPv4 route in the BGP update message to its iBGP peer PE1.
4. PE2 sets the BGP next-hop attribute in the BGP update message to the IP address of PE2.
5. On receipt of the BGP update, PE1 performs the following tasks:
 - a. Installs the label received from PE2 as an outgoing inner label
 - b. Allocates a local label L1 for incoming traffic from CE1
 - c. Imports this route to the appropriate VRF table
 - d. Translates the VPN-IPv4 address back to an IPv4 address
 - e. Advertises this prefix to the CE1 via IGP

PE1 distributes the label-to-FEC mapping to CE1-PE1 using LDP. As usual, the outer label for the PE-PE tunnel is distributed either using LDP or RSVP-TE.

In the CSC deployment scenario, IP and MPLS control-plane components that can fail along the data traffic path include the following:

- IGP and LDP on the PE-CE link
- LDP (or RSVP-TE) on PE and P routers

- iBGP on PE routers
- IGP on PE and P routers

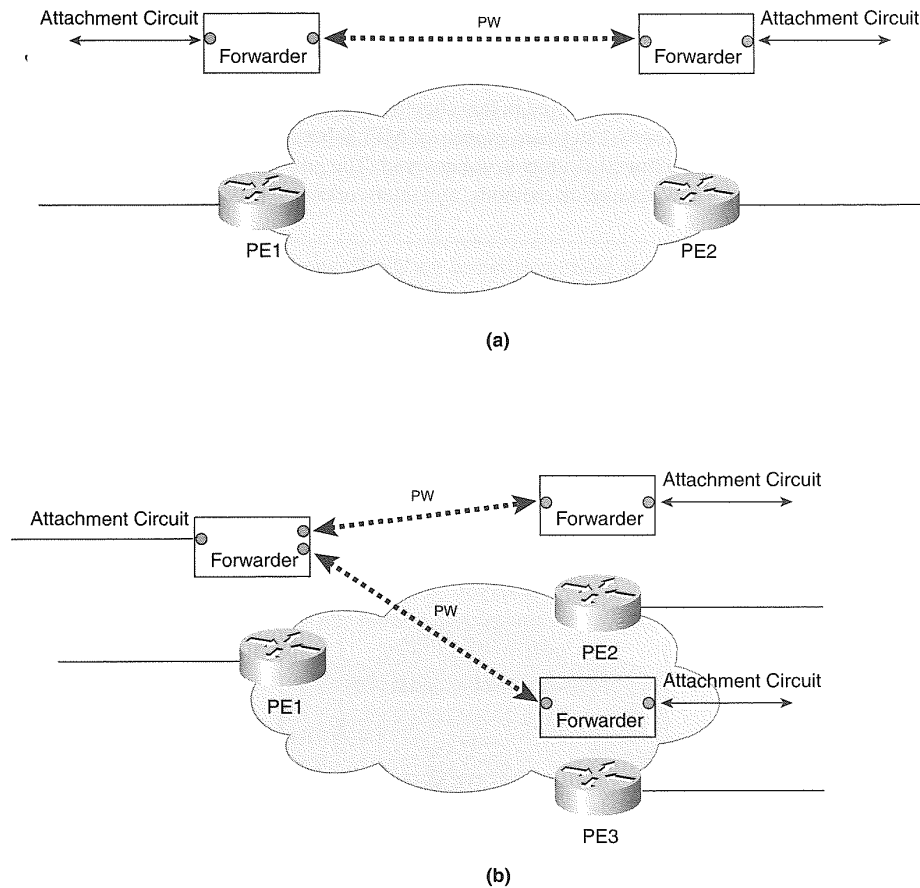
For example, PE2 maintains an incoming tunnel label (outer label distributed via LDP or RSVP-TE), an incoming inner label (which was allocated and distributed to PE1 via iBGP by PE2), and an outgoing LDP label (on the PE2-CE2 side.) Therefore, PE2 must be LDP restart- and BGP restart-capable.

Layer 2 Virtual Private Networks

An L2VPN interconnects customer devices (such as a host, a router, or a Layer 2 switch) at multiple sites and allows them to communicate using Layer 2 frames.¹³ An L2VPN service provider can offer several types of L2VPN services to its customers, such as point-to-point and point-to-multipoint. In any type of L2VPN service, a PE performs the following tasks:

- Receives a Layer 2 frame from a CE over an attachment circuit (which could be an FR DLCI, an ATM VPI/VCI, an Ethernet port, a VLAN, and so forth).
- Encapsulates the frame in a *pseudowire* (PW) protocol data unit. (A PW is a mechanism that emulates the essential attributes of a service, such as FR, ATM, Ethernet, Time Division Multiplexing [TDM], and SONET/SDH over IP and/or MPLS network. A protocol data unit [PDU] that contains all of the data and control information necessary to emulate the desired service is known as a PW-PDU.)
- Transports one or more PWs over a tunnel across the PSN to another PE, which demultiplexes the PW.
- Removes the encapsulation.
- Sends the frame in its native format to the destination CE.

In L2VPNs, the set of procedures used for making Layer 2 forwarding decisions is called a *forwarder*. Conceptually, you might think of a forwarder as bound to an attachment circuit on one side and a PW on the other side. Different types of L2VPNs contain different types of forwarders. For example, a point-to-point L2VPN has a forwarder that binds exactly one PW to exactly one attachment circuit (see Figure 10-9 [a]). In contrast, a point-to-multipoint L2VPN contains a forwarder that binds one attachment circuit to one or more PWs (see Figure 10-9 [b]).

Figure 10-9 L2VPNs Forwarder Architectures (a) Point-to-Point Case (b) Point-to-Multipoint Case

In the case of L2VPNs, PWs correspond to nested LSPs in the PE-to-PE tunnel in which an individual PW is identified based on the MPLS label. As in the case of L3VPNs, in L2VPNs PE-to-PE tunnels can be established using LDP or RSVP-TE. Thus PE-to-PE tunnels in L3VPNs and L2VPNs mainly differ in terms of the type of traffic carried within the nested LSPs and how VPN labels are signaled. For example, in L3VPNs, the PE-to-PE tunnels carry IP packets across the backbone. In contrast, with L2VPNs, PE-to-PE tunnels transport Layer 2 frames, such as FR, ATM, and Ethernet, across the backbone.

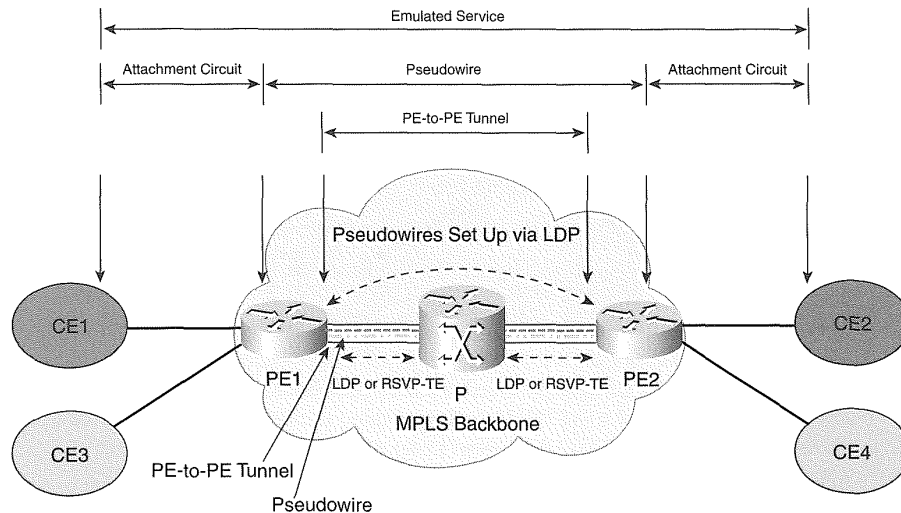
Protecting L2VPN Services Against Control-Plane Failures

With regard to protection against control-plane failures, two types of Layer 2 VPN services need to be considered: *virtual private wire service* (VPWS) and *virtual private LAN service* (VPLS). It is also possible to have a variant of the VPLS known as IP-only LAN-like Service (IPLS), but IPLS is not considered in this chapter. The discussion in this chapter is restricted to the control-plane-related functional aspects of MPLS-based L2VPN services. The control-plane restart of attachment circuits is not covered. The protocol details of VPWS and VPLS architecture, PW encapsulation methods, and interworking procedures are beyond the scope of this book.

Virtual Private Wire Service

Figure 10-10 shows the reference model for VPWS. A VPWS is a VPN service that provides Layer 2 point-to-point connectivity between two CE devices.

Figure 10-10 Reference Model for the Virtual Private Wire Service with an MPLS Backbone



In VPWS, CE devices (routers, bridges, switches, or hosts) are connected to PEs in the service provider network via attachment circuits. In the MPLS backbone, PEs are connected via PWs, which correspond to PE-to-PE LSPs that are established using LDP. In turn, PWs are multiplexed (nested) and transported over PE-to-PE tunnels. The PE performs a one-to-one mapping between the attachment circuit and PW. The transport of PWs that contain Layer 2 frames across the MPLS network involves three layers of encapsulations:

- The tunnel header to carry PWs from an ingress PE router to an egress PE router

- The demultiplexer to distinguish individual PWs within a single tunnel
- PW encapsulation (a 32-bit control word used to carry information about the enclosed Layer 2 frames, such as Layer 2 header bits, sequence number for in-order delivery of the frame, and so forth)

In the case of an MPLS backbone, the tunnel header and demultiplexer fields are based on MPLS labels. The tunnel labels are distributed using LDP or RSVP-TE, whereas distribution of demultiplexers (L2VPN labels) for PWs is typically accomplished via LDP.

NOTE

Note that in Cisco products VPWS are sometimes called an earlier name—*Any Transport over MPLS* (AToM).

In a PE-to-PE data traffic path, the IP/MPLS control-plane components that can fail (and therefore must support a relevant graceful restart mechanisms) include the following:

- LDP on PEs that is used to set up PWs
- LDP or RSVP-TE on PE and P routers that establishes the PE-PE MPLS tunnel
- IGP on PE and P routers that distributes routing information inside the backbone

Note that VPSW service might also be disrupted by control-plane failures on PE-CE links. However, this book does not cover control-plane restart procedures for Layer 2 attachment circuits.

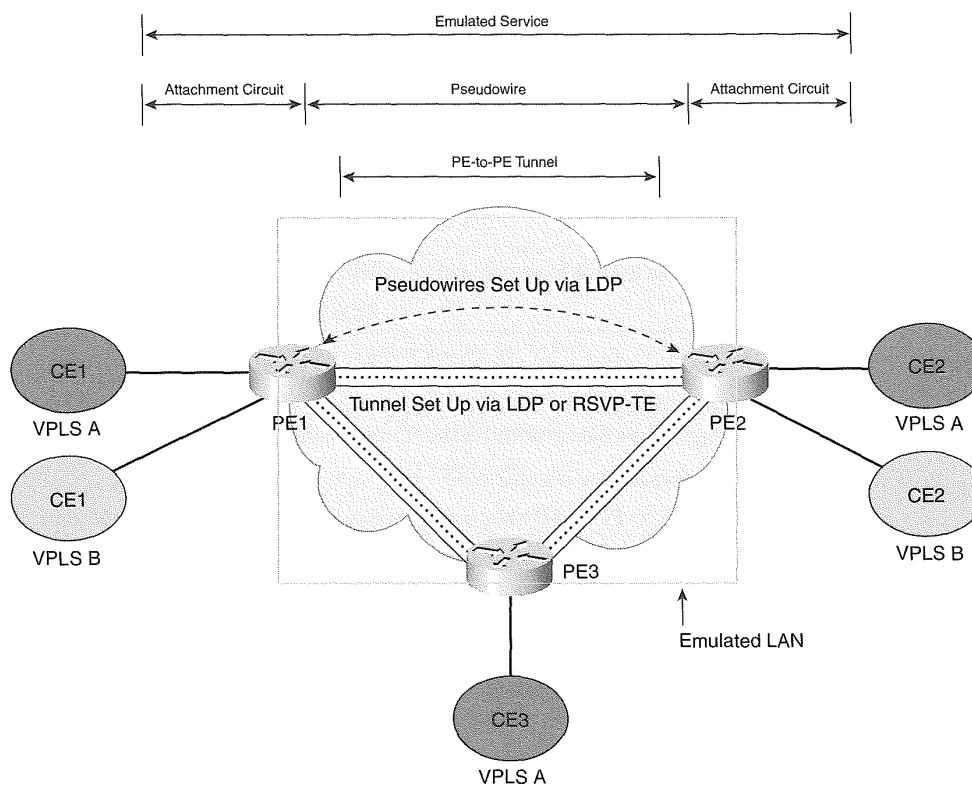
Virtual Private LAN Service

A VPLS is an L2VPN service that emulates the performance of a LAN across a packet-switched network and provides connectivity between several customer LAN segments. For users, sending data through an L2VPN seems the same as being connected via a single LAN. VPLS is similar to VPWS in that it makes its forwarding decisions without considering a Layer 3 header. However, VPLS is different from VPWS in the following two ways:

- First, VPLS PE forwards frames based on a frame's Layer 2 header.
- Second, VPLS allows frames from a single CE to be transmitted to multiple remote CEs.

In other words, VPLS is a point-to-multipoint service, whereas VPWS is a point-to-point service.

Figure 10-11 shows the reference model for VPLS. A CE device attaches to a PE device via an Ethernet port. PE devices are connected via point-to-point bidirectional PWs to carry frames between them. In the case of the MPLS backbone, a PW consists of a pair of point-to-point unidirectional LSPs (also referred to as *VC LSPs*). To a CE device, the set of PE devices interconnected via PWs appears as a single emulated LAN.

Figure 10-11 Reference Model for the Virtual Private LAN Service with an MPLS Backbone

To aggregate traffic in the backbone, tunnels are established between PEs. In the case of an MPLS network, as in L3VPN, tunnels are set up using LDP or RSVP-TE. Multiple PWs can be carried inside a single tunnel. To distinguish a particular PW inside the tunnel, a unique demultiplexer value is assigned to each PW. In general, for a point-to-point PW, the egress PE of the tunnel selects the demultiplexer value and distributes the value using signaling protocols to the PE that is the ingress of the tunnel.

In the case of an MPLS backbone, the demultiplexer value for a PW is based on a label, which is signaled over an LDP session using LDP extensions defined for PW signaling. To set up a full mesh of point-to-point PWs, all PEs in a VPLS must have an LDP session between them. After an LDP session has been established between two PEs, they can exchange L2VPN labels (VC labels) for all PWs between them over this session. A PW is considered operational when both incoming and outgoing VC LSPs have been established. When the PWs are point to multipoint, they can be set up using either a point-to-point signaling protocol such as LDP or a

point-to-multipoint signaling protocol such as BGP. Further discussion on point-to-multipoint PWs is beyond the scope of this book.

From the perspective of data-plane operation, when a PE receives a frame over an attachment circuit, it examines the MAC destination address and forwards the address to another local attachment circuit or over a PW to another PE. This second PE in turn forwards the frame out on another remote attachment circuit.

A conventional LAN provides broadcast and unicast services. This requires that each VPLS PE be able to perform the following tasks:

- Dynamically learn MAC addresses on physical ports and virtual ports (PWs)
- Associate learned addresses with the appropriate physical and virtual ports
- Replicate packets on physical and virtual ports
- Flood packets with unknown unicast destinations

When a PE learns a new MAC address on an inbound VC LSP, it associates the learned MAC address with the outbound VC LSP. Similarly, as a PE learns a MAC address over an inbound physical port, it associates that address with the corresponding outbound physical port. For further details, refer to the VPLS architecture specification.^{13,14}

In the case of VPLS, IP/MPLS control-plane components along the PE-PE data path are identical to the VPWS except that PWs may be signaled using LDP or BGP. Hence, to avoid VPLS service disruption because of control-plane failures in the backbone, all nodes along the data path must support relevant IP and MPLS control-plane restart procedures (see Chapters 4 through 9 for details).

Network Fault Tolerance and MPLS-Based Recovery

To this point, discussion has focused on ways to protect service outages against control-plane failures. Although discussion of data-plane protection against network failures is not the main focus of this chapter, for the sake of completeness it would be useful to examine schemes that are commonly used to improve network survivability in the face of node and link failures.

Generally network outages are caused by failures of network elements such as links and nodes. *Network survivability* refers to the capability of a network to continue to provide an acceptable level of service in the presence of network failures. Design of survivable networks involves trade-offs between several conflicting objectives such as recovery speed, cost, and complexity. Network operators are generally interested in providing the fastest recovery mechanisms at the lowest feasible costs.

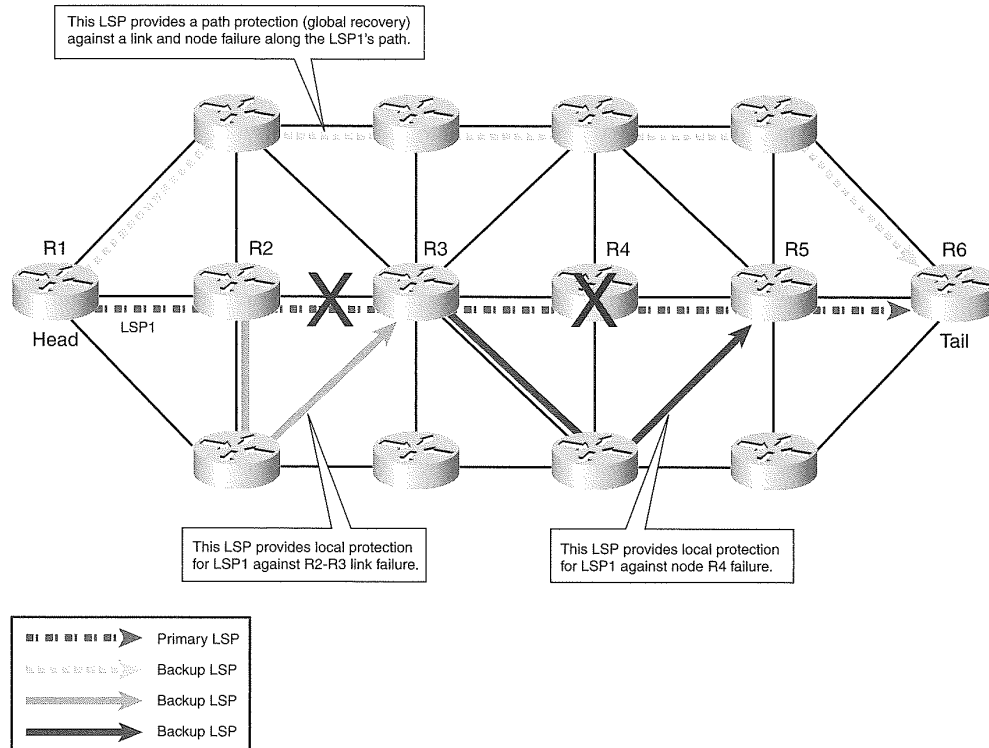
To satisfy these conflicting design objectives, survivable network architectures consider recovery at multiple layers, including WDM, SONET, MPLS, and the IP layer. At a particular layer, there are two main types of recovery approaches: protection and restoration (or rerouting).^{15,16}

Protection and Restoration

In the case of protection, a backup path is pre-established for traffic on the primary path, and when a fault is detected the protected traffic is switched over to the backup path. In contrast to protection, when using restoration a backup path to reroute traffic on the primary path is established on demand after detection of a network failure.

A protection or restoration recovery scheme can be further categorized as local or global repair, based on the scope of recovery. A local recovery scheme protects a primary path segment in the vicinity of the fault against a single link or node failure, whereas a global recovery scheme protects an end-to-end primary path against any link or node failure (see Figure 10-12). Unlike global recovery, a local recovery scheme provides a backup path closest to the point of failure and thereby avoids extra delay by propagating failure notification to the upstream nodes to reroute traffic onto the backup path. Avoiding delay is highly desirable to reduce traffic disruption during failure. In general, recovery at a lower layer is faster than at a higher layer, but is costly. For example, recovery at the SONET layer is faster than at the MPLS layer, which in turn is faster than recovery at the IP layer. Moreover, local recovery is faster than global recovery, and at any given layer protection is faster than restoration, but costly.

Figure 10-12 *MPLS-Based Local and Global Recovery Schemes*



In L2VPNs, increasing the level of network availability with packet-loss duration not exceeding tens of milliseconds during network failures is required. This increase is necessary because emulated circuits can be carrying delay-sensitive data such as *voice over IP* (VoIP) and loss-sensitive data such as stock exchange data services. It is required to provide QoS comparable to the native emulated circuits. Therefore, to satisfy the negotiated SLAs, customer data traffic must be protected against link and node failures in the backbone and access network. The following sections describe recovery schemes at different layers.

Optical Layer Protection

Wave-division multiplexing (WDM) technology enables the aggregation of multiple channels (or wavelengths) onto a single fiber channel. Each WDM channel can carry traffic from multiple higher layers such as SONET, MPLS, and IP. The protection at the optical layer provides the ability to switch data traffic from one wavelength to a completely different wavelength.

In WDM networks, optical layer protection can be deployed in point-to-point and ring configurations. Figure 10-13 shows SONET equipment making use of optical layer point-to-point link protection. For example, if for some reason the working fiber is cut, the switch mechanism directs the traffic onto the protection fiber. Generally, handling of fiber cuts is faster at the optical layer than the SONET layer. This is because a single switch at the optical layer can restore all wavelengths on the failed link instead of restoring individual SONET links.

Figure 10-14 shows an example of the optical layer protection in a ring configuration.

Figure 10-13 Optical Layer Protection in Point-to-Point Architecture

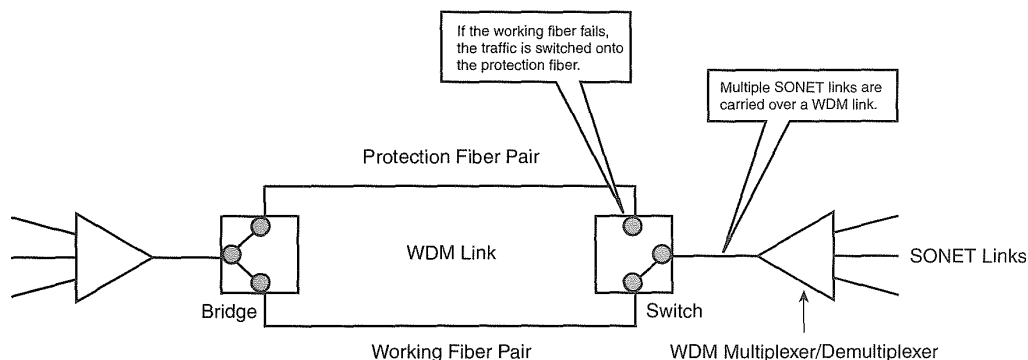
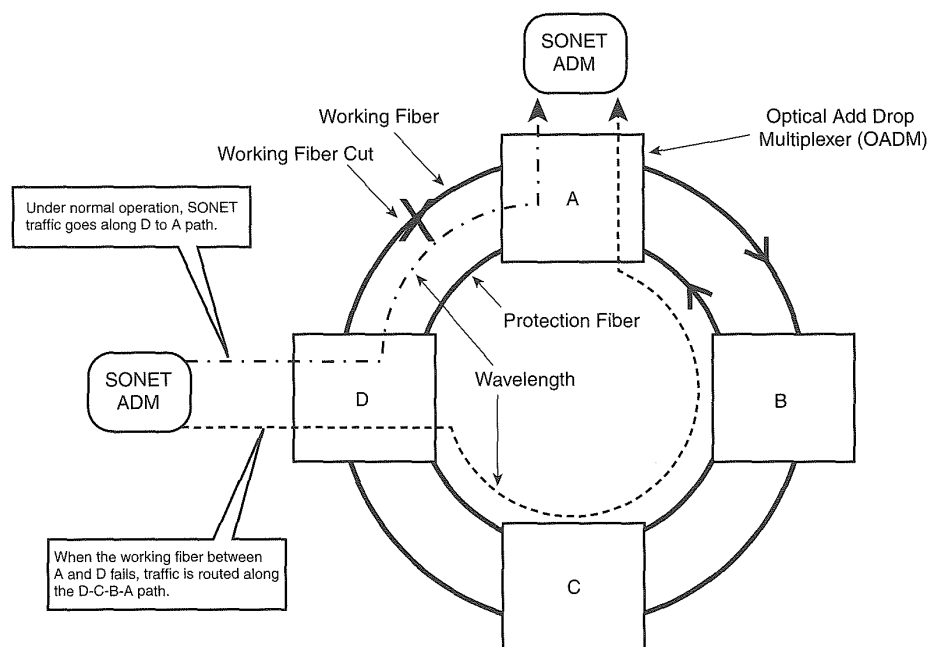


Figure 10-14 *Optical Layer Protection in Ring Architecture*

SONET/SDH Layer Protection

The SONET *Automatic Protection Switch* (APS) is a mechanism that protects SONET equipment against link failures (such as fiber cuts and transceiver failures). Like the optical layer, the SONET/SDH layer protection can also be used in point-to-point and ring architectures. After a failure has been detected, SONET APS makes it possible to switch from the working line to the protection line within 50 milliseconds.

IP Layer Restoration

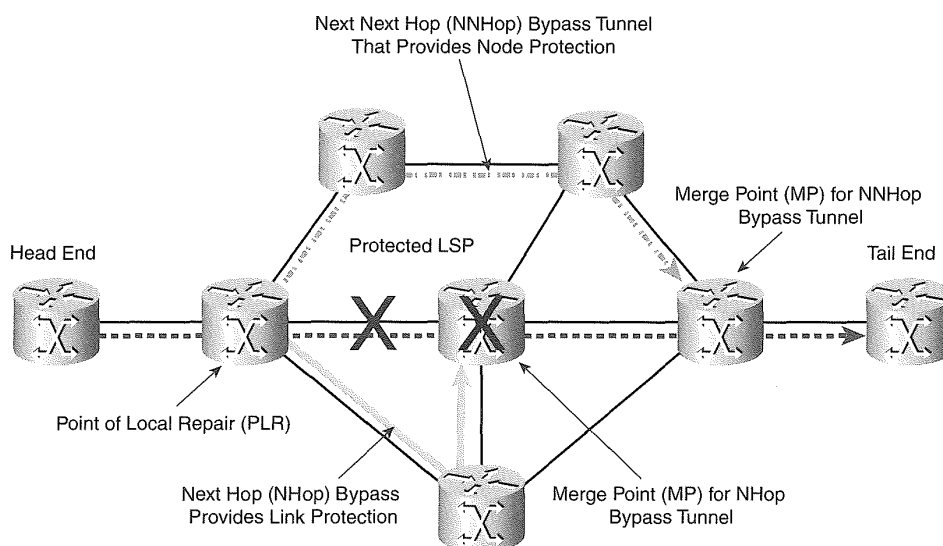
You can also use the IP layer to reroute traffic around link and node failures. For example, an IGP maintains a synchronized link-state database that describes the network topology. Using this topology information, an IGP calculates a shortest-path tree. When a topology change is detected, a new route is quickly calculated, and the data traffic is rerouted along the new route. Depending on the routing protocol and the setting of various time parameters, it might take a long time (in the order of minutes) to reroute traffic around the failure. Although IP layer restoration might be acceptable as a recovery method for best-effort data traffic, IP layer rerouting typically is not suitable for time-sensitive and data-loss-sensitive applications (such as VoIP), which require recovery times in the order of tens of milliseconds.

MPLS Layer Protection—Fast ReRoute

Currently WDM and SONET/SDH-based recovery mechanisms are widely used in existing networks. However, deployments of higher-layer recovery schemes such as MPLS-based local and path protection are also increasing rapidly with the emergence of converged packet-switched networks. Generally a higher-layer recovery scheme is slower, but provides a more granular and cost-effective solution than a lower-layer scheme such as SONET/SDH protection. One such MPLS-based local protection scheme that limits packet loss in the order of tens of milliseconds during network failure is known as *Fast ReRoute* (FRR).

MPLS FRR provides a mechanism to set up backup label-switched paths and quickly reroute traffic from protected TE LSPs onto the backup tunnels on detection of local link and node failures.¹⁷ Because the backup LSPs (referred to as *bypass tunnels*) are established before the failure and provide local protection, MPLS FRR can reroute traffic within tens of milliseconds (see Figure 10-15 for a summary of terminology). When the node immediately upstream of the failure point (known as the *point of local repair* [PLR]) detects a network element failure (based on a Layer 2 mechanism or using IGP/RSVP Hellos), it quickly reroutes the set of protected TE LSPs onto the appropriate bypass tunnel around the failed network elements. An LSP is said to be protected at a particular node if it has one or more associated backup tunnels originating at that node. MPLS FRR provides not only a backup path, but also a backup path with equivalent QoS (such as bandwidth) to the primary path.¹⁸ In short, an MPLS FRR-based local protection scheme provides a natural, practical, and effective solution against link or node failures.

Figure 10-15 MPLS Fast ReRoute Commonly Used Terms



The two approaches of MPLS FRR are referred to as bypass and detour. Table 10-1 compares these approaches.

Table 10-1 Comparison of the Bypass and Detour Approaches of MPLS FRR

	Bypass Approach	Detour Approach
Link and node failures	Provides protection against link and node failures	Provides protection against link and node failures
LSP protection	Protects multiple LSPs by using a single bypass tunnel; offers a facility (or one-for-n) LSP protection	Requires a separate backup LSP for each primary LSP that needs to be protected; in other words, detour provides a one-for-one LSP protection
Label stacks	Employs label stack to nest multiple LSPs	Does not use label stack

As discussed in Chapter 9, RSVP has a soft state model, which means that the RSVP state is periodically refreshed. If the state is not refreshed within an associated refresh interval, it is removed. Following a link or node failure, if a node in the proximity of the failure were to remove the RSVP-TE state for the LSPs for which the node had not received refresh messages from the peer, this would cause traffic disruption on those LSPs and defeat the purpose of rerouting such LSPs over the bypass tunnel (see Figure 10-16). The bypass scheme provides an elegant solution for this problem (see Figure 10-17).

Figure 10-16 Undesirable Effects of Not Refreshing RSVP-TE Messages for Protected LSPs After the Link or Node Failure

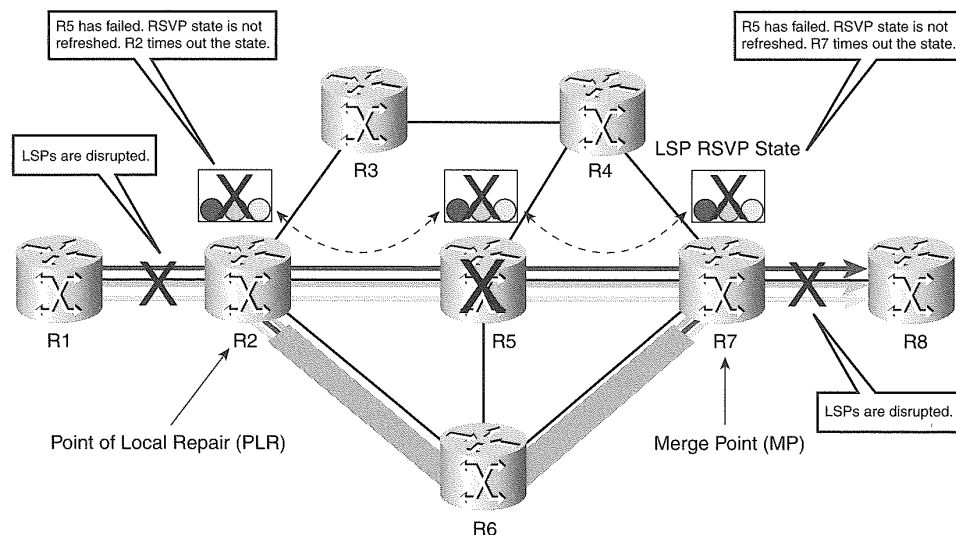
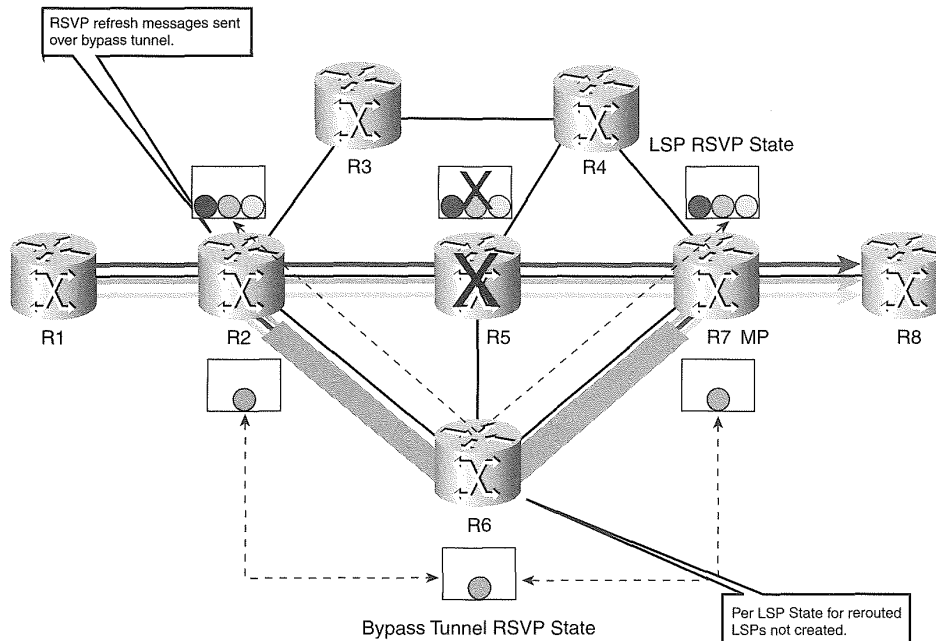
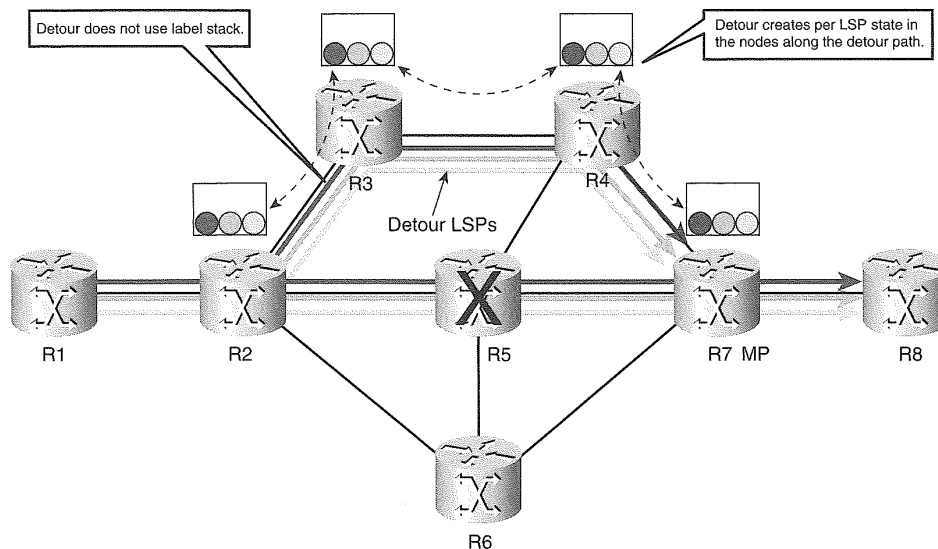


Figure 10-17 Bypass Handling of RSVP-TE Messages After Link or Node Failure

To maintain LSP RSVP-TE state across link or node failure between PLR and *merge point* (MP) nodes, RSVP refresh messages are also sent over the bypass tunnel. Because of label stacking, RSVP-TE messages are not visible to any node along the bypass tunnel. As a result, even when several LSPs are being rerouted over the bypass tunnel, nodes along the bypass tunnel do not see RSVP messages and therefore do not create any RSVP state for the rerouted LSPs. From the point of view of scalability, this feature of the bypass approach makes it an effective solution for local protection. Detour, in contrast, does not use a label stack and as a result creates and refreshes per LSP state in all nodes along the backup path (see Figure 10-18).

Figure 10-18 *Detour One-for-One LSP Protection Solution*

Protecting Bypass Tunnels Against Control-Plane Failures

A bypass tunnel is identical to any other regular TE tunnel in that it is established using RSVP-TE. However, the bypass tunnel is different from a regular TE tunnel in that it carries data traffic only when one or more protected LSPs are rerouted onto it following a link or node failure. Therefore, from the control-plane restart perspective, two cases deserve consideration. The effect of RSVP-TE control-plane restart on the bypass tunnel is harmless while no traffic is being rerouted over the tunnel. This is because even if the bypass tunnel state is lost because of a control-plane restart, it can be reestablished without effecting user traffic. A large number of LSPs could be protected over a single bypass tunnel. Therefore, in terms of service outage, the effect of a control-plane restart can be devastating when the bypass tunnel is carrying protected LSPs' traffic. If protected LSPs' traffic is being carried, to mitigate the negative impact of a control-plane restart, all nodes along the bypass tunnel's path must be capable of RSVP-TE GR and NSF.

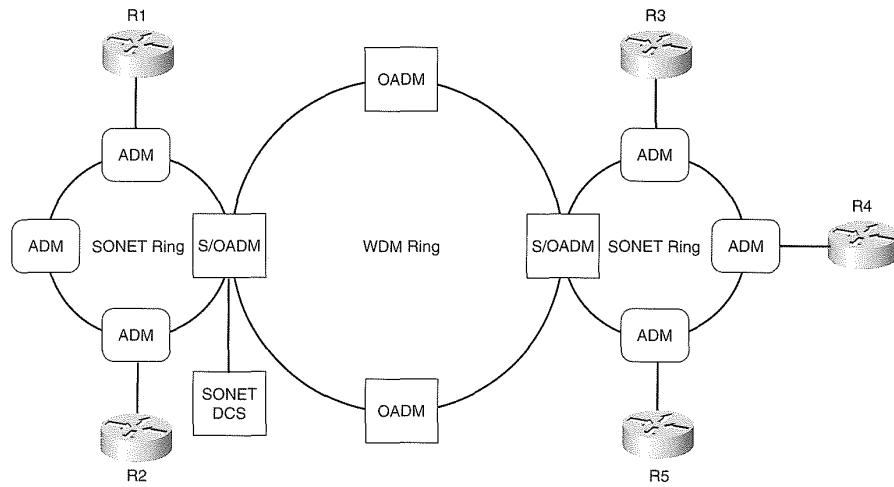
With reference to intra-area, inter-area, and inter-AS MPLS-TE deployment scenarios, the existing MPLS FRR mechanisms can be used to provide local protection for a TE LSP against link or node failure, regardless of whether the LSP is an intra-area, inter-area, or inter-AS TE LSP. However, local repair notification procedures to trigger rerouting of the TE LSP might be somewhat different in the case of inter-area and inter-AS TE LSP.¹⁹

Interactions Between Different Protection Layers

A network might use different types of protection architectures because of cost and recovery speed requirements. Generally, different layers are better suited for handling different types of faults. For example, fiber cuts can be typically handled more efficiently at the optical layer, whereas equipment failures (such as laser failure) can be handled at the SONET layer. MPLS-based recovery methods allow faster recovery than the IP layer rerouting. Although recovery at the MPLS layer may be slower than the SONET layer, it allows more granular options and more bandwidth-efficient recovery options than SONET and optical layer protection. The physical topology of a network with multiple protection layers is shown in Figure 10-19 (a). The virtual topology as seen by the MPLS/IP layer devices such as routers is shown in Figure 10-19 (b).

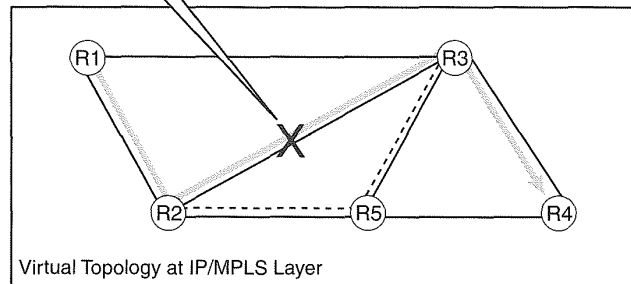
In multilayer survivable network architectures, a client/server relationship exists between different protection layers (see Figure 10-20). For example, circuit-switched *time-division multiplexed* (TDM) services and IP/MPLS layer services are transported over SONET. Similarly, SONET is transported over the optical layer. This means the failure of a single WDM link can cause simultaneous failure of several optical channels (wavelengths), which can make recovery impossible at the higher layers (SONET, MPLS, IP).

Therefore, interworking between different layers becomes crucial to improve the end-to-end service availability during network failures. In other words, it is highly desirable to provide some coordination between protection mechanisms at different layers. You can achieve this coordination between different layers automatically by ensuring that a server layer recovers from a fault before its client layer detects that failure. For example, the SONET layer takes about 2.3 microseconds to detect a link failure. This means that as long as the optical layer can restore traffic within this time, the SONET layer will not even notice this failure. In general, it is difficult to achieve recovery times in the order of microseconds. You can also achieve this coordination between different layers by introducing an additional delay (hold-off time) at the client layer before it attempts to recover to allow adequate time for the server layer to complete its recovery.

Figure 10-19 Multilayer Network Protection Architecture (a) Physical Topology (b) Virtual Topology

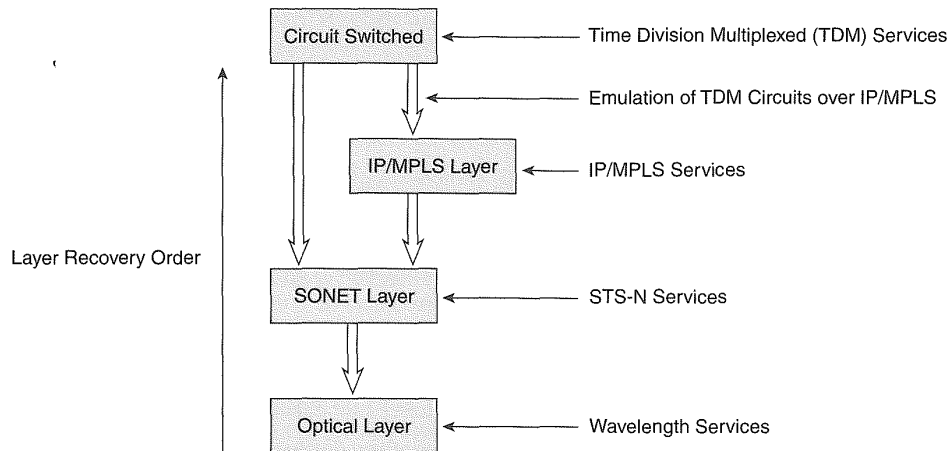
(a)

If the link between R2-R3 fails, at the MPLS layer traffic is rerouted around the failed link along the R2-R5-R3 path.



(b)

Figure 10-20 Layered Recovery Architecture



Network Fault Tolerance and MPLS OAM Mechanisms

Operation, administration, maintenance, and provisioning (OAM&P) encompass a broad set of management functional areas, including fault management, configuration management, accounting management, and security management.²⁰ A fault (or defect) is an interruption of the capability of a network connection caused by a network element such as link and node failures. Discussion of functions in this section is restricted to fault management and performance management:

- **Fault management**—A set of functions including fault detection, fault verification, fault isolation, and fault recovery
- **Performance management**—A set of functions for gathering and analyzing network QoS (such as packet delay, delay variation, and loss) and availability performance parameters for the purpose of measuring the quality of network performance against negotiated SLAs or to help with planning and provisioning of the network

From the perspective of end customers, network availability is measured by the frequency of service outages and the recovery time for each outage. In reference to L2VPN and L3VPN services, network availability depends on the fault tolerance of the PE-to-PE tunnel, pseudowire, and VPLS instance. Therefore, capability to quickly detect and recover from data-plane and control-plane faults is critical to reduce service-outage duration. In short, OAM mechanisms play a pivotal role in improving the network availability, and network providers rely heavily on these tools for swift identification and removal of network faults.

With the increased acceptance of MPLS as the technology of choice to deliver legacy Layer 2 and emerging Layer 3 services over the converged packet-switched networks, effective MPLS OAM capabilities become critical to manage, troubleshoot, and monitor SLAs. Standardization efforts are underway to specify a suite of MPLS OAM mechanisms to facilitate network operators in identifying defects, verifying network connectivity, and measuring network performance.

Examples of MPLS OAM mechanisms include LSP-Ping, Y.1711, pseudowire *virtual circuit connection verification* (VCCV), and VPLS OAM:^{21–24}

- **LSP-Ping**—A tool to detect data-plane faults in LSPs and isolate faults. LSP-Ping offers two modes of operation called ping and traceroute. The ping mode is used to check basic connectivity of an LSP, whereas the traceroute mode is used to verify the forwarding path of data packets.
- **Y.1711**—A tool that provides fault management functions such as the following:
 - Connectivity verification (CV) to detect LSP connectivity
 - Forward defect indication (FDI) to inform downstream nodes about an upstream defect
 - Backward defect indication (BDI) to inform upstream nodes about a downstream defect.
- **VCCV**—Uses MPLS LSP-Ping to exchange sets of control messages between PEs to verify connectivity of the pseudo wire. VCCV defines two modes of operation: diagnostic mode and fault-detection mode. The diagnostic mode is used to verify the connectivity. The fault-detection mode is used to emulate attachment circuit-specific fault-detection mechanisms.
- **VPLS OAM**—Mainly consists of fault management and performance management functions. VPLS is an end-to-end service, which can span different types of access and core networks. For example, the access network might be a bridged network, whereas the core network might be an MPLS network. This means the VPLS OAM mechanisms must work end to end, independent of the underlying transport technologies such as Ethernet, IP, and MPLS.

Bidirectional Forwarding Detection

BFD provides a fast mechanism to detect faults in the bidirectional forwarding-plane path between pairs of directly connected or non-directly connected systems.²⁵ You can use BFD to provide fault detection on different types of forwarding paths such as direct physical links, MPLS LSPs, and tunnels. As you will learn shortly, you can use BFD to provide an efficient fault-detection component for many OAM-related functions.

Motivations

For many applications, the fault-detection times (in the order of seconds) of the existing mechanisms (such as routing protocol Hellos) are too slow and can cause a great deal of data loss during network failures. Moreover, there are certain media types, such as Ethernet, that currently do not have mechanisms for detecting interface or forwarding plane failures rapidly. To summarize, the main goals of the BFD are to provide the following:

- An efficient and fast fault-detection mechanism (with low processing overhead and with subsecond granularity)
- A common liveness-detection framework that is media and protocol independent and that allows a wide range of fault-detection times

How Does BFD Improve Network Availability?

As discussed in Chapter 1, “Understanding High Availability of IP and MPLS Networks,” availability can be improved by reducing mean time to repair (MTTR). MTTR includes time required for failure detection, fault diagnosis, and actual repair. Thus, quick fault detection is a key for improving network availability.

How Does BFD Improve Network Convergence?

The time duration beginning with an event resulting in a topology or route change (such as link failure, router failure, and next hop change) until all routers have the consistent view of the network is known as the convergence period. Because the routers may have inconsistent topology and routing information during convergence period, there is increased likelihood of data loss or service outage. Therefore, it is extremely important to speed up the network convergence. It turns out that slow detection of neighbor failures is one of the main reasons for the delayed IGP convergence.²⁶ Thus, the BFD fast failure detection mechanism helps to improve network convergence.

BFD Protocol Mechanics

In some respects, the BFD protocol resembles the Hello mechanism of a routing protocol. For example, two systems that support the BFD protocol periodically transmit BFD control packets over each forwarding path between them. If a system does not receive BFD control packets from a neighbor on a particular path for a while, the system assumes that some component in that particular bidirectional path to the neighbor has failed.

Before two systems can exchange BFD control packets, they need to establish a separate BFD session for each communication path and data protocol in use between them. There are two modes of BFD operation: asynchronous mode and demand mode.

- **Asynchronous mode**—In this mode, two systems periodically transmit BFD control packets to each other; if a system does not receive a successive number of those packets, that system declares the BFD session to be down.
- **Demand mode**—In this mode, after the two systems have established a BFD session, they stop sending BFD control packets to each other. Afterward, when either system needs to verify connectivity explicitly, that system sends a short sequence of BFD control packets and then the BFD protocol quiesces.

In addition to these two modes, BFD also defines an Echo function that allows a local system to transmit a sequence of Echo packets to a remote system (see Figure 10-21). The remote system is expected to loop back these Echo packets through its forwarding path. If the transmitting system does not receive back a number of Echo packets in succession, it declares the BFD session to be down. The Echo function can be used in conjunction with either asynchronous or demand mode.

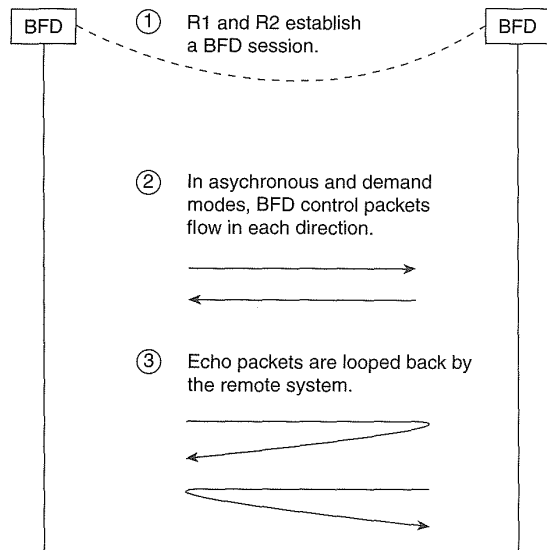
In general, different situations might require different BFD modes. For example, demand mode might prove advantageous in situations in which you want to reduce the processing overhead of periodic control packets. For example, demand mode might be advantageous for a system with a large number of BFD sessions. The advantage of the echo function is that it enables the transmitting system to verify the forwarding path through a neighbor with reduced jitter in the round-trip time. However, the echo function has the drawback of requiring twice as many packets to achieve a particular fault-detection time as the pure asynchronous mode requires.

BFD Applications

For different applications, BFD may be operating at multiple layers such as physical links, virtual circuits, and LSPs. You can implement BFD in the forwarding plane using a client/server model in which BFD provides services such as indication of session status (up/down) to its clients. BFD can also be implemented in the control plane; however, this may make it difficult to detect certain types of failures. In summary, many applications can take advantage of the BFD fast fault-detection mechanism to improve network availability. Some examples of such applications are described in the sections that follow.

Using BFD for Detecting IGP Neighbor Liveness

The Hello protocols of IGPs, such as OSPF and IS-IS, do not allow fast detection of an adjacency failure. For example, it can take OSPF and IS-IS on the order of a few seconds to detect an adjacency failure. Thus, the BFD fast detection mechanism can be used to overcome limitations of IGP Hello protocols. BFD usage for IGP neighbor liveness not only allows faster IGP convergence but also improves IGP adjacency scalability by allowing larger values for Hello timers to reduce processing associated with periodic IGP Hellos.

Figure 10-21 BFD Modes of Operation

Using BFD for LSP Data-Plane Fault Detection and Control-Plane Verification

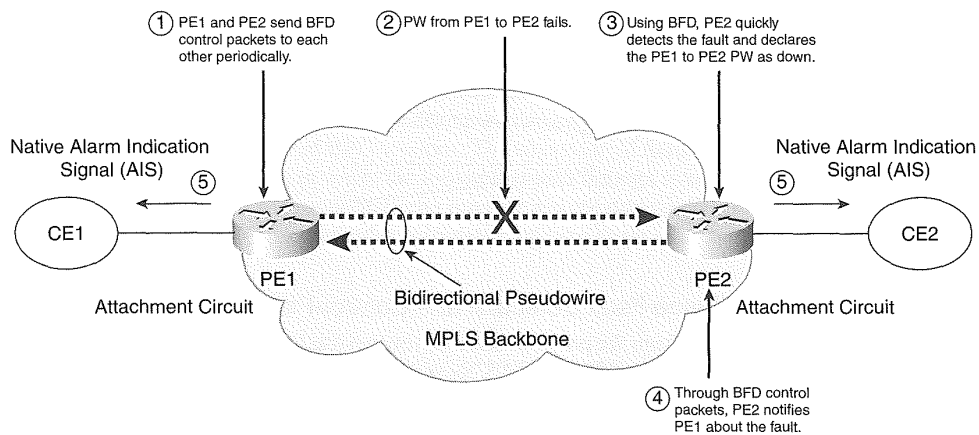
As you already know from the previous section, you can use LSP-Ping to detect an MPLS LSP data-plane failure and to verify the MPLS control plane against the data plane. In contrast, although you can use BFD for fast detection of a data-plane failure, it does not provide a mechanism to verify the MPLS control plane against the data plane. Because LSP-Ping includes extensive control-plane verification, it is computationally more onerous than BFD for detecting MPLS LSP data-plane faults. This is because BFD is designed as a lightweight means to detect forwarding-plane faults quickly, and BFD uses a fixed-packet format that makes BFD conducive to hardware- or firmware-based implementations. Therefore, it is advantageous to take the efficient and fast fault-detection capabilities of the BFD and combine them with the control-plane verification capabilities of the existing LSP-Ping mechanism to provide a scalable solution for LSP fault detection.²⁷

To use BFD in conjunction with LSP-Ping for fault detection on an LSP, a BFD session is established between the ingress and egress LSR of that LSP. The ingress LSR encapsulates the BFD control packets in the MPLS label stack that correspond to LSP being verified. The labeled BFD control packets take the same path as the LSP being verified and are processed by the control plane of the egress LSR.

Using BFD for PW Fault Detection

Because PWs can be carrying data of delay-sensitive and loss-sensitive emulated services, fast failure detection might be highly desirable. One way to achieve this goal is to use the BFD mechanisms along with the VCCV fault-detection mode. For an illustration of BFD used for PW fault detection, consider the network diagram shown in Figure 10-22. Using the BFD mechanism, PE1 sends BFD control packets periodically to PE2. When PE2 stops receiving control packets for a defined period of time, PE2 declares that direction of the PW is down and notifies PE1 via control packets. Depending on the type of emulated service (such as ATM, FR, or TDM), the PEs might also need to send a native *alarm indication signal* (AIS) over the related attachment circuits to inform the endpoints about the fault condition.

Figure 10-22 BFD-Based PW Fault Detection

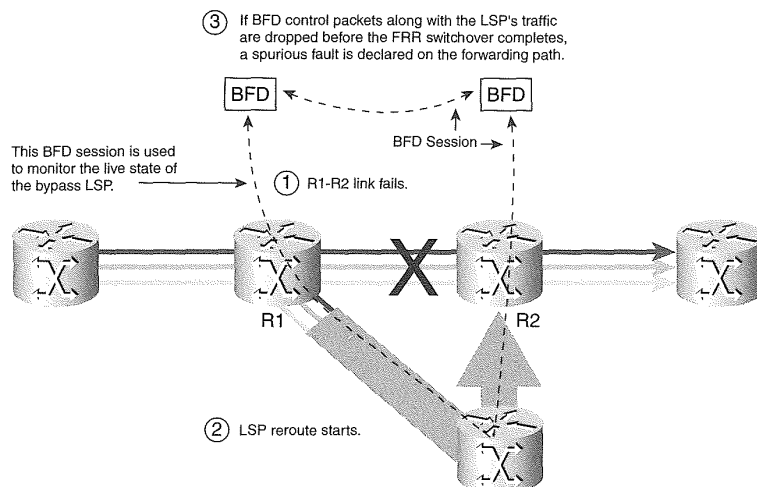


Using BFD for MPLS FRR Fault Detection

As discussed previously, in the event of a link or node failure data traffic from multiple protected LSPs is rerouted onto the bypass tunnel. The bypass tunnel might be carrying traffic from a large number of LSPs, and you can use BFD for fast detection of bypass LSP failures. Consider a scenario in which a number of LSPs are protected against the SONET link failure using a bypass LSP (see Figure 10-23). Assume there is a BFD session to detect failures on the bypass LSP and an additional BFD session to verify the forwarding path between R1 and R2. If the link between R1 and R2 fails, based on SONET alarms, R1 detects the link failure and attempts to reroute traffic from all protected LSPs onto the bypass LSP quickly. Although

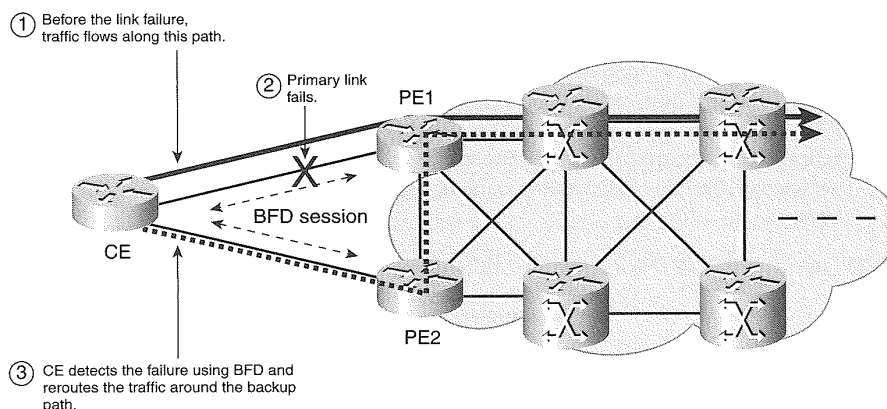
MPLS FRR switchover can be accomplished rapidly, nevertheless it takes a finite amount of time (in the order of tens of milliseconds) to complete. During this time, data traffic on the protected LSPs is dropped. This means that following an MPLS FRR switchover due to a link or node failure, BFD control packets will be lost until the switchover is complete. Therefore, if the fault-detection interval of the BFD session between R1 and R2 is smaller than the MPLS FRR switchover time, a fault is declared for an LSP even though the LSP is in the process of being locally repaired. To avoid spurious BFD fault detections following MPLS FRR switchover, the BFD fault-detection interval should be greater than the MPLS FRR switchover time.

Figure 10-23 Interactions of BFD and MPLS FRR Mechanisms



Using BFD for Fault Detection in the Access Network

To provide high-availability services such as VoIP, a CE device may be dual homed to guard against a PE-CE link failure. In this scenario, there are two forwarding paths: the primary and the backup. During normal operation, traffic flows on the primary path. In the event of a primary path failure, traffic is rerouted over the backup path. Thus, a BFD session can be used to monitor the liveness of each forwarding path (see Figure 10-24). If the primary forwarding path fails, the BFD can quickly be used to detect the link failure and reroute the traffic onto the backup path.

Figure 10-24 *BFD Use in the Access Network*

BFD Interactions with the IP and MPLS Control-Plane Graceful Restart Mechanisms

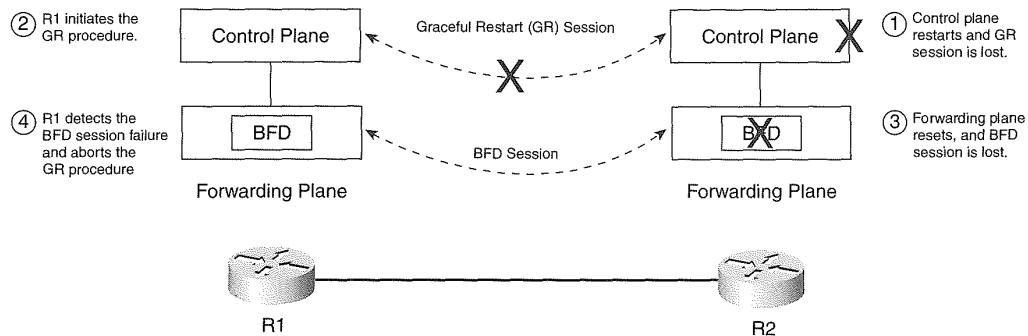
As discussed in the previous chapters, to mitigate the detrimental effects to forwarding that might result from IP/MPLS control-plane restarts, the restarting router and its neighbors must support a certain set of control- and forwarding-plane capabilities. In particular, the restarting router must support not only the appropriate graceful restart extensions, but also be able to preserve its IP/MPLS forwarding state. In summary, the control-plane graceful restart mechanisms depend on the existence of a separate forwarding plane that can continue to function across the control-plane restarts.

Although most graceful restart mechanisms provide a way out of unsuccessful restart attempts by reverting to normal restart procedures, it might take a while (in the order of seconds) before a router determines whether a given graceful restart attempt with a neighbor is viable. For example, suppose a router had indicated its capability to preserve the forwarding state, but actually could not preserve it across the restart. The neighbors honoring the pre-restart capability exchange that are unaware of the actual situation, would continue to forward data traffic through the restarting router which would be black holed. In fact, in the event of an unsuccessful graceful restart attempt, data traffic could be black holed a little longer (it could be black holed for the time it takes to detect a session failure plus an additional waiting period for session reestablishment) than it would be in the case of normal restart (only time to detect session failure).

The BFD control packet contains a bit called the Control Plane Independent bit. By setting the C bit to 1, a system announces independence of forwarding- and control-plane components. In other words, if the C bit is set to 1, it indicates that the transmitting system's BFD is

implemented in the forwarding plane and can continue to function across the control-plane restart. This means the information in the C bit can be used to determine whether it is appropriate to continue with a graceful restart procedure or abort it. For example, consider two graceful restart-capable routers R1 and R2 with a BFD session and a graceful restart session between them (see Figure 10-25). Assume both routers have expressed their capability to preserve a forwarding state across the restart and both have been transmitting BFD control packets with the C bit set to 1. Suppose R2 restarts and loses its forwarding state across the restart. In this situation, if a BFD session fails while graceful restart is in progress, R1 should abort the graceful restart procedure.

Figure 10-25 Interactions of BFD and Control-Plane Graceful Restart Mechanisms



Network Fault Tolerance and In-Service Software Upgrades

In general, router outages can be caused by unplanned or planned events. An unplanned outage occurs because of unforeseen faults in hardware or software components. An MPLS control-plane restart is an example of an unplanned outage. In contrast, a planned outage occurs when a router is taken out of service for scheduled maintenance, such as during a software or hardware upgrade. Using redundant hardware and software components helps to mitigate the unplanned outages. As described previously, a combination of redundant control plane, IP/MPLS control-plane protocol enhancements, and NSF capability enable the nondisruptive *stateful switchover* (SSO) from an active to standby control processor. This switchover helps to reduce the downtime caused by unplanned control-plane failures.

To this point in the book, discussions have focused on mechanisms that assist in avoiding service disruption caused by unplanned control-plane failures. However, as mentioned in Chapter 1, operations related to planned software upgrades are the single largest source of router downtime. Therefore, carrier-grade services require reduction of not only unplanned outages, but also planned outages. The in-service software upgrade (also known as *hitless software upgrade*) mechanism enables you to reduce router downtime during planned software upgrade/downgrade operations.

The in-service software upgrade mechanism involves upgrading the standby control processor with a new version of software, synchronization of state information between the active control processor running with the old version of software and the standby control processor running with a new version of software, and making a nondisruptive SSO to the standby. The in-service software upgrade procedure entails a deliberate switchover (in contrast to automatic switchover, as in the case of unplanned control-plane failures) across which forwarding must not be disrupted. Therefore, the in-service software upgrade procedure relies on the SSO and NSF capabilities, which are used for unplanned control failures. Further details on in-service software upgrade are beyond the scope of this book.

Summary

This chapter described how you can use various IP and MPLS control-plane restart mechanisms to improve network fault tolerance against unplanned control-plane failures. The design of survivable networks has multiple aspects. To provide a broader view of network fault tolerance, this chapter also briefly touched on MPLS layer recovery schemes against network failures, the relationship of MPLS OAM to network availability, and reduction of unplanned outages through the use of the in-service software upgrade.

References

- ¹ Andersson, L., and T. Madsen, "PPVPN Terminology," IETF Work in progress, March 2004.
- ² RFC 2661, "Layer Two Tunneling Protocol (L2TP)," W. Townsley et al., August 1999.
- ³ Goyret, I., J. Lau, and M. Townsley, "Layer Two Tunneling Protocol (Version 3)," IETF Draft, Work in progress, June 2004.
- ³ RFC 2784, "Generic Routing Encapsulation (GRE)," D. Farinacci et al., March 2000.
- ⁴ Committee T1-Telecommunications, "Technical Report on a Reliability/Availability Framework for IP-Based Networks and Services," Working Group T1A1.2 (Network Survivability Performance), Technical Report No.70, September 2001.
- ⁵ RFC 3630, "Traffic Engineering (TE) Extensions to OSPF Version 2," D. Katz, K. Kompella, and D. Yeung, September 2003.
- ⁶ RFC 3784, "Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)," H. Smit and T. Li, June 2004.
- ⁷ RFC 2547, "BGP/MPLS VPNs," E. Rosen, et al., March 1999.
- ⁸ Le Roux, J. L., et al. "Requirements for Inter-Area MPLS Traffic Engineering," IETF work in progress, March 2004.

- ⁹ Vasseur, J. P. and R. Zhang, "MPLS Inter-AS Traffic Engineering Requirements," IETF work in progress, January 2004.
- ¹⁰ RFC 2547, "BGP/MPLS IP VPNs," E. Rosen, et al., September 2003.
- ¹¹ Callon, R. and M. Suzuki, "A Framework for Layer 3 Provider Provisioned Virtual Private Networks," IETF Draft, Work in progress, October 2003.
- ¹² Clercq, J. D., et al., "BGP-MPLS VPN Extension for IPv6 VPN," IETF Draft, Work in progress, June 2004.
- ¹³ Andersson, L., and E. Rosen, "L2VPN Framework," IETF Draft, Work in progress, September 2003.
- ¹⁴ Lasserre, M., et al. "Virtual Private LAN Services over MPLS," IETF Draft, Work in progress, November 2003.
- ¹⁵ Wu, T-H., "Emerging Technologies for Fiber Network Survivability," *IEEE Communications Magazine*, February 1995.
- ¹⁶ RFC 3469, "Framework for Multi-Protocol Label Switching (MPLS)-based Recovery," F. Hellstrand and V. Sharma, February 2003.
- ¹⁷ Pan, P., et al., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels," IETF Draft, Work in progress, August 2004.
- ¹⁸ Vasseur, J. P., et al., "MPLS Traffic Engineering Fast Reroute: Backup Tunnel Path Computation For Bandwidth Protection," IETF Draft, Work in progress, October 2002.
- ¹⁹ Ayyangar, A., and J. P. Vasseur, "Inter-Area and Inter-AS MPLS Traffic Engineering," IETF Draft, Work in progress, February 2004.
- ²⁰ ITU-T Recommendation M.3400, "TMN Management Functions," February 2000.
- ²¹ Pan, P. et al., "Detecting Data Plane Liveness in MPLS," IETF Draft, Work in progress, February 2004.
- ²² ITU-T Recommendation Y.1711, "OAM Mechanism for MPLS Networks," November 2002.
- ²³ Aggarwal, R., and T. D. Nadeau, "Pseudo Wire (PW) Virtual Circuit Connection Verification (VCCV)," IETF Draft, Work in progress, February 2004.
- ²⁴ Mohan, D., et al., "VPLS OAM Requirements and Framework," IETF Draft, Work in progress, July 2004.
- ²⁵ Katz, D., and D. Ward, "Bidirectional Forwarding Detection," IETF Draft, Work in progress, July 2004.
- ²⁶ Alaettinoglu, C., et al., "Towards Milli-Second IGP Convergence," IETF Draft, Work in progress, November 2000.
- ²⁷ Aggarwal, R., et al., "BFD For MPLS LSPs," IETF Draft, Work in progress, July 2004.