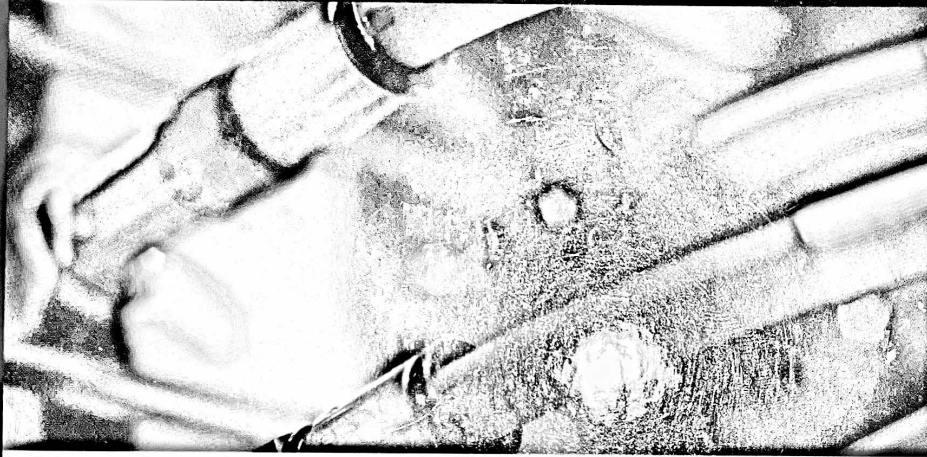




Definitive MPLS Network Designs

Field-proven MPLS designs covering MPLS VPNs,
pseudowire, QoS, traffic engineering, IPv6, network
recovery, and multicast

ciscopress.com

Jim Guichard, CCIE® No. 2069
François Le Faucheur
Jean-Philippe Vasseur

Definitive MPLS Network Designs

Jim Guichard

François Le Faucheur

Jean-Philippe Vasseur

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 and 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 March 2005

Library of Congress Cataloging-in-Publication Number: 2003116497

ISBN: 1-58705-186-9

Warning and Disclaimer

This book is designed to provide information about MPLS network designs. Every effort has been made to make this book as complete and accurate as possible, but no warranty or fitness is implied.

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

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

Trademark Acknowledgments

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

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 of the professional technical community.

Reader feedback is a natural continuation of this process. If you have any comments about 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 be sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

- Layer 3 MPLS VPN backbone traffic engineered with MPLS TE**—This scenario is slightly more complicated. In addition to the labels usually required by MPLS VPN (the LDP and BGP labels), an additional label corresponding to the TE LSP is pushed onto the label stack as the traffic is forwarded onto the TE LSP by the TE LSP headend router. In Figure 2-14, an IP packet is sent by the CE router. Upon receiving the IP packet, the PE router determines the VRF and pushes a two-label stack. This label stack corresponds to the BGP label provided by the PE router, which advertises the destination prefix for the VPN in question plus the LDP label to reach the destination PE router. The packet is then label-switched until it reaches P1. At this point, the headend router P1 pushes an additional label corresponding to the TE LSP. It is worth mentioning that the LDP label is swapped and the remote LDP peer P3 provides the corresponding label. Why is an LDP session required between the headend and the tail-end of the TE LSP (P1 and P3 in this example)? As you saw in the previous example, a PHP operation is performed by the penultimate hop of the TE LSP. Hence, without an LSP session between the headend and tail-end of the TE LSP, the TE LSP tail-end router (P3 in this example) would receive a packet with an unknown BGP label instead of receiving an LDP label and would just drop the packet (or forward it using what it believes to be the correct forwarding entry, thus causing a traffic black hole).

NOTE

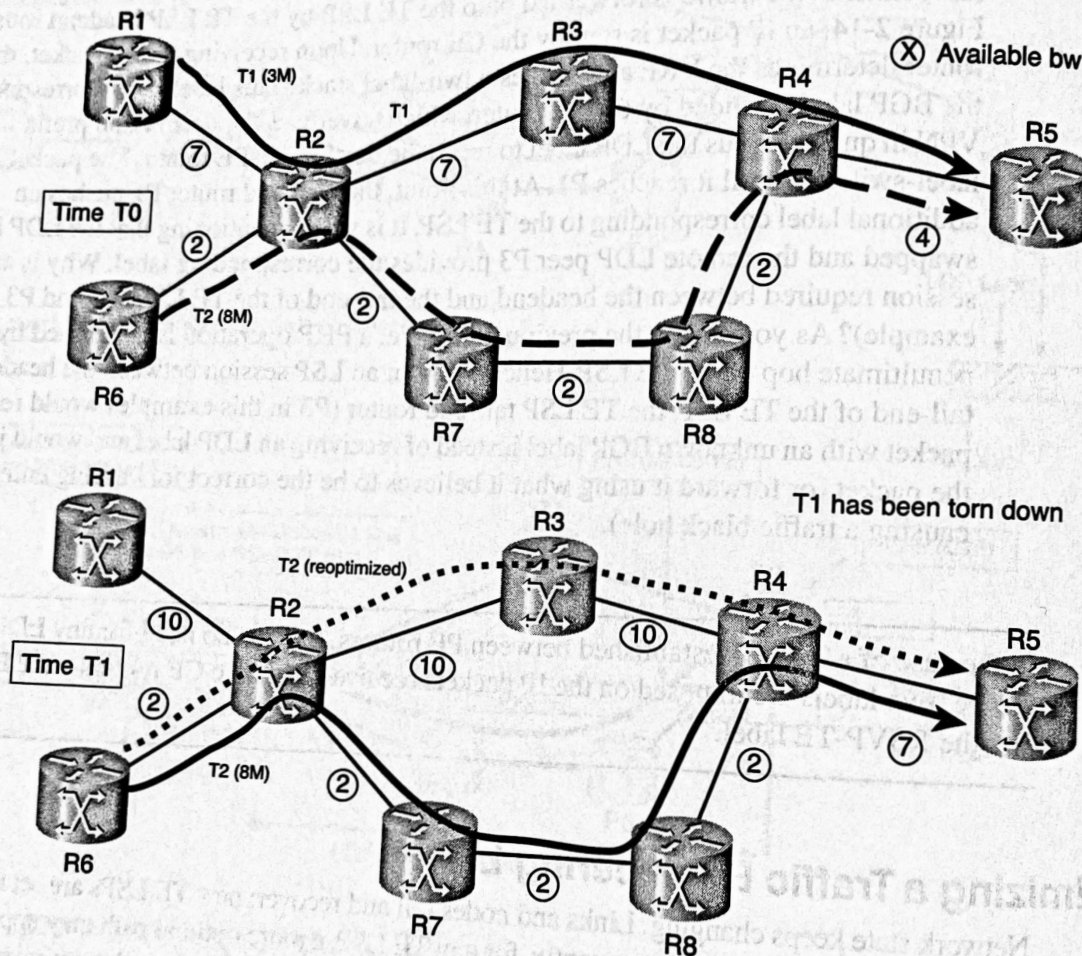
In the case of a TE LSP established between PE routers, there is no need for any LDP session. Hence, two labels are imposed on the IP packets received from the CE routers—the BGP label and the RSVP-TE label.

Reoptimizing a Traffic Engineering LSP

Network state keeps changing. Links and nodes fail and recover; new TE LSPs are set up while others may be torn down. Consequently, for any TE LSP, a more optimal path may appear in the network. It is then highly desirable to detect the existence of such a path and reoptimize a TE LSP along a better path when it becomes available. Consider the example shown in Figure 2-15.

Figure 2-15 Reoptimizing a TE LSP

All links have a IGP metric of 1 and an initial reservable bandwidth of 10M (except link R4-R5 which has 15M).



In Figure 2-15, all the links have a cost of 1 and an initial reservable bandwidth of 10 Mbps, except the link R4-R5, which has 15 Mbps of reservable bandwidth. Suppose that the first TE LSP to be signaled is T1 between R1 and R5 (for 3 Mbps), which follows the shortest path, obeying the bandwidth constraint, of R1-R2-R3-R4-R5. Immediately following this event, R6 signals the TE LSP T2 (for 8 Mbps), which follows the shortest path offering 8 Mbps of bandwidth, which is R6-R2-R7-R8-R4-R5. Hence, at time t0, both T1 and T2 are up and running. At time t1, the TE LSP T1 is torn down, which frees up 3 Mbps of bandwidth along the path R1-R2-R3-R4-R5. When a reoptimization evaluation process is triggered by R6, it determines that a more optimal (shortest) path exists between R6 and R5 (the path R6-R2-R3-

R4–R5), and it reroutes the TE LSP along this more optimal path. This illustrates the concept of reoptimization. There are several important aspects to highlight:

- **Nondisruptive reoptimization**—A key property of MPLS TE is its ability to reoptimize a TE LSP along a more optimal path *without any traffic disruption*. Hence, the headend router should not tear down the “old” LSP and then reestablish it along a more optimal path. Instead, the headend router first establishes the new TE LSP (which belongs to the same session as the old TE LSP). As soon as that new TE LSP is successfully set up, the old TE LSP is torn down. This is known as the *make-before-break* mechanism, a property that was not really available (implemented) with other Layer 2 protocols such as ATM.
- **Avoid double booking**—If you carefully observe Figure 2-15, you will notice that after T1 has been torn down, the available bandwidth on the link R4–R5 is 7 Mbps, which is not enough to accommodate the second occurrence/LSP of T2 (along the more optimal path). So how can R6 set up the LSP along the path R6–R2–R3–R4–R5 without first tearing down the TE LSP, because the make-before-break procedure appears to have both LSPs active at the same time? The answer is that T2 is reoptimized with the option of sharing the bandwidth with the old LSP of T2. In other words, when the new LSP of T2 is signaled, the router R4 sees that it shares the bandwidth with the old LSP (thanks to RSVP’s Share Explicit option). Hence, no double booking occurs (R4’s call admission control does not count the bandwidth twice). Similarly, when computing the path for the reoptimized TE LSP, the headend (R6 in this example) knows that the available bandwidth to reoptimize T2 is the current network reservation state plus the bandwidth that will be freed up by the current LSP for T2.

An important aspect of any MPLS TE network design relates to the reoptimization triggers. How often and when should a TE LSP be evaluated for reoptimization? A Cisco router has several reoptimization triggers:

- **Manual reoptimization**—A command is issued on the headend router that triggers a reoptimization evaluation for each TE LSP the router is the headend for. If a better path is found, the TE LSP is reoptimized according to the make-before-break procedure.
- **Timer-based reoptimization**—A reoptimization timer is configured on the headend router. Upon expiration, the headend tries to reoptimize each TE LSP it is the headend for.
- **Event-driven reoptimization**—In some cases, it may be desirable to trigger a reoptimization upon the occurrence of a particular event, such as the restoration of a link in the network. Indeed, if a link is restored in the network, some TE LSPs may benefit from that new link to follow a more optimal path. When configured, each time a new link is advertised as newly operational, this triggers the evaluation of any potential reoptimization. It is worth noting that this should be handled with some care to avoid network instability. Let’s consider the case of an unstable link. This would have the effect of constantly attracting new TE LSPs, which would then fail and be reoptimized again, creating some very undesirable network instability and consequently traffic disruption. This explains why such a trigger should always be used in conjunction with a dampening mechanism at the interface or IGP level to avoid network instabilities in case of unstable links or routers.