CHAPTER 3

Interexchange Carrier Design Study

USCom is a fictitious nationwide data and long-distance voice service provider in the U.S. that provides connectivity between local exchanges in different geographic regions. It also facilitates inter-Local Access and Transport Area (LATA) services (as described in the Federal Communications Commission [FCC] Telecommunications Act of 1996), as well as a complete portfolio of data services. USCom may be classified as an Interexchange Carrier (IXC) that owns its fiber and transmission facilities as well as a Layer 2 switching infrastructure (ATM and Frame Relay) spanning its service footprint.

NOTE A LATA in the U.S. determines where a Local Exchange Carrier (LEC) can transmit traffic and where an IXC is required to carry traffic between LATAs. A state may have several LATAs. A few LATAs cross state boundaries.

This chapter discusses the current USCom MPLS network design, its evolution, and how USCom characteristics and objectives influenced the corresponding design decisions that were made.

USCom's Network Environment

USCom has been offering Internet access for many years to other service providers (wholesale), Enterprises, and small/medium business customers. It currently has an installed base of more than 35,000 Internet ports. These Internet ports are supported on 350 Internet edge routers (called Internet access provider edge [PE] routers) located in their 100 Points of Presence (POPs) that are situated across the country. Internet connectivity is obtained via transit providers, private peering sessions, and connections in major cities to various Network Access Points (NAPs).

USCom has also had great success with its Layer 3 Multiprotocol Label Switching (MPLS) VPN service (which is based on the architecture described in [2547bis]) since its inception in 2002. Acceptance of the service has grown throughout USCom's customer base. Currently some 12,500 VPN ports are installed across the country, and this number is growing considerably on a monthly basis. The customer-managed customer edge (CE)

routers are connected via 255 Layer 3 MPLS VPN PE routers hosted in USCom's various POPs. Note that PE routers are dedicated to either the Internet or Layer 3 MPLS VPN access. Given the success of this offering, USCom plans to add 6000 customer access links per annum, although based on the current trend this figure is considered conservative. Total traffic volume, which includes both Internet and VPN, is expected to grow at approximately 30 percent per annum.

References Used in this Book

Throughout this book you will see references to outside resources. These are provided in case you want to delve more deeply into a subject. Such references will appear in a bracketed code, such as [L2VPN]. If you want to know more about this resource, look up the code in this book's appendix and you can find out specific information about the resource.

USCom owns fiber across the country and is running a long-distance optical core based on dense wavelength division multiplexing (DWDM) technology. This translates to availability of raw high-speed links (OC-48 (2.488 Gbps) and OC-192 (10 Gbps)) for provider router (P router) and PE router interconnection, at relatively low cost and provisioning time. USCom can activate additional capacity by enabling additional wavelengths (lambdas) in a relatively short time frame. USCom takes advantage of this to enforce an overengineering policy for core router links.

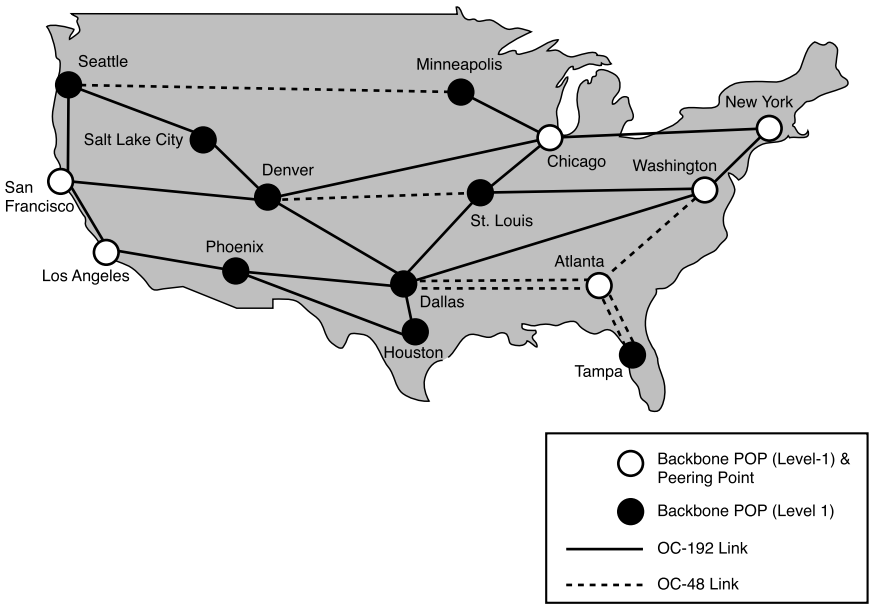
The high-speed core links are provided to routers as native lambdas straight from the DWDM equipment without any intermediate SONET Add/Drop Multiplexer (ADM). (Note that SONET framing is in use between the routers and the DWDM equipment.) These links do not benefit from any protection at the optical level. Some links interconnecting P routers and PE routers are provided through a SONET infrastructure overlaid over the optical infrastructure. The SONET links are protected by means of SONET protection provided by Bidirectional Line Switch Rings (BLSRs) with four fibers, also called BLSR/4. (See [NET-RECOV] for more details on SONET-SDH recovery mechanisms.)

Intra-POP connectivity is achieved via Packet over SONET (PoS) or switched Gigabit Ethernet. Because of the relatively low cost of switched Gigabit Ethernet technology and the negligible cost of fibers within a premises, USCom also maintains an overengineered intra-POP capacity.

Access from CE router to PE router for both Internet and Layer 3 MPLS VPN connectivity is provided via Frame Relay, ATM, leased line, or SONET. Each of these physical (or logical) links is dedicated to a single CE router. These links involve a significant cost that typically precludes simple overengineering and mandates tight dimensioning. Access speeds range from 64 kbps to OC-48.

The USCom nationwide backbone POP topology, interconnected through OC-48 and OC-192 links, is illustrated in Figure 3-1.

Figure 3-1 *USCom Nationwide Topology*

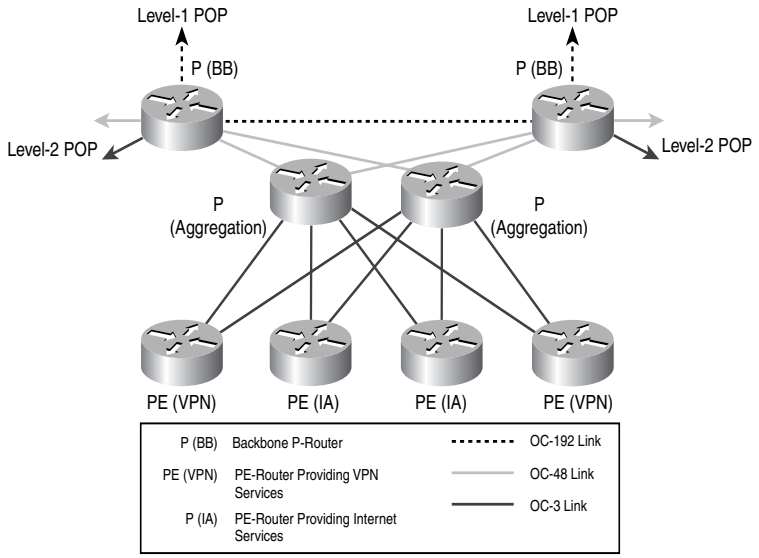


The USCom network is structured into three levels of POPs. Each POP is classified as either a backbone (Level 1), medium (Level 2), or small (Level 3) facility. The level depends on the density of the customer access and combined traffic throughput requirements. All routers are operated as a single autonomous system, with American Registry for Internet Numbers (ARIN) assigned AS number 32765. USCom has been assigned the 23/8 IP address space. The company uses this for its internal infrastructure as well as customer allocation.

Level 1 POPs are the backbone POPs (as shown in Figure 3-1) comprising the high-capacity backbone P routers dedicated to long-distance transit and interconnection of lower-level POPs to this long-distance transit backbone. PE routers providing Internet and Layer 3 MPLS VPN services from these major locations are also deployed, as well as some additional P routers acting as an aggregation layer inside the POP for these PE routers. Aggregation P routers reduce the number of IGP adjacencies that have to be maintained by the backbone P routers to two, because each core P router has to peer with only two aggregation P routers (in addition to the other core P routers in the backbone) instead of with all the PE routers in the POP (whose number can be fairly high, and growing, in a Level 1 POP).

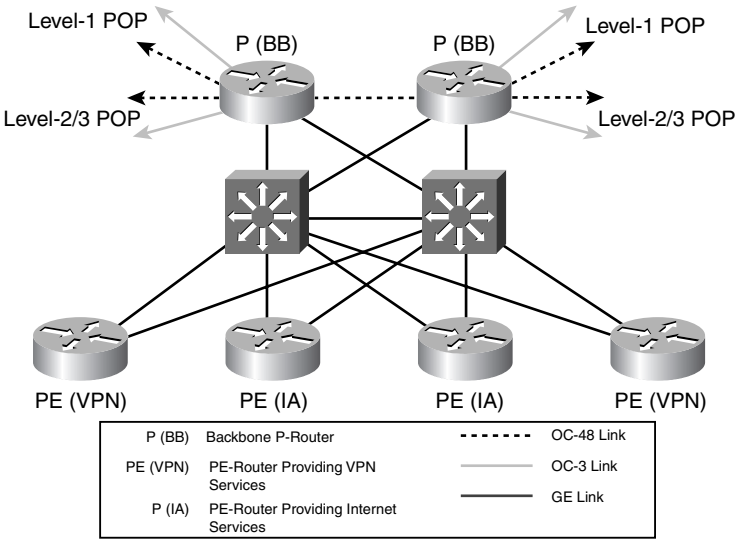
Each Level 1 POP has two backbone P routers that interconnect via OC-48, dual OC-48, or OC-192 links to the rest of the backbone network. They also interconnect with lower-level POPs using either OC-3 (155.52 Mbps) or OC-48 links. Each backbone P router is connected to both local aggregation P routers via a point-to-point OC-48 link. Each PE router (and there may be several) is connected to both aggregation P routers via OC-3 PoS links. There are currently 15 Level 1 POPs, the structure of which is illustrated in Figure 3-2.

Figure 3-2 *USCom Level 1 POP Design*



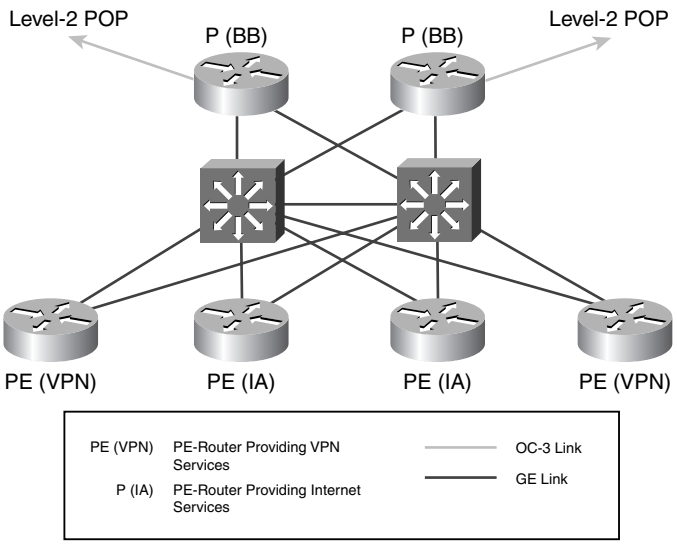
The Level 2 POPs are composed of P routers that connect to the Level 1 POPs, or another Level 2 POP, via OC-3 or OC-48 links, and the PE routers in medium access locations. Each PE router is connected to both backbone P routers via redundant switched Gigabit Ethernet (using two separate Gigabit Ethernet switches). There are currently 25 Level 2 POPs, the structure of which is illustrated in Figure 3-3.

Figure 3-3 *USCom Level 2 POP Design*



The Level 3 POPs are composed of PE routers in remote locations and P routers that connect to Level 2 POPs via OC-3 links. There are currently 60 Level 3 POPs, the structure of which is illustrated in Figure 3-4.

Figure 3-4 *USCom Level 3 POP Design*



Several years ago, USCom deployed a SONET network providing OC-3 links. These links are protected at the SONET layer by the protection mechanisms provided by four-fiber BLSRs. These allow recovery from any link failure, with some special conditions specified by the SONET standard, within 60 ms. USCom satisfies all the conditions, including ring distance limited to 1200 km, less than 16 SONET stations, and ring in idle state before protection. Figure 3-5 shows the protected OC-3 links provided by the four-fiber BLSRs and used between Level 1 and Level 2/3 POPs. Because these links are protected and stable, USCom decided to use them in the core network without any changes.

NOTE The use of SONET protection covers only the case of a link failure within the SONET network but not an IP router interface failure (sometimes considered a link failure) or a router failure. On the other hand, USCom considers router interface failures and router failures rare enough that they are acceptable and do not the use of additional recovery mechanisms such as Automatic Protection Switching (APS).

Figure 3-5 *Protected OC-3 Links Provided by Four-Fiber BLSRs*

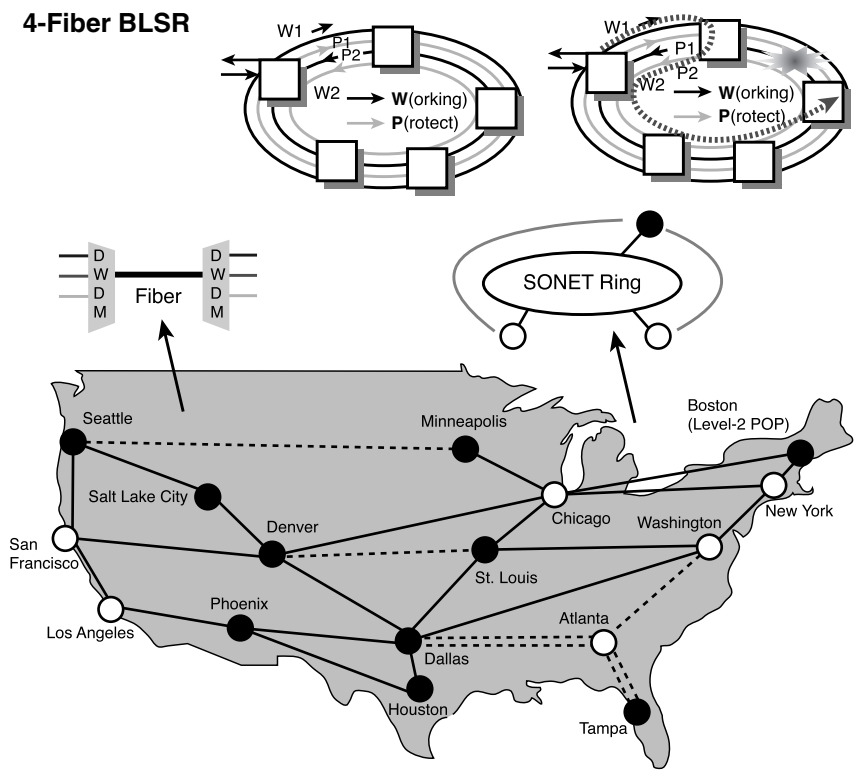


Figure 3-5 also shows that the USCom optical network uses DWDM technology, allowing the multiplexing of tens of light paths over a single fiber. Note that USCom has deployed Coarse Wave Division Multiplexing (CWDM) equipment in some metro areas, offering a lower degree (4) of multiplexing. The DWDM equipment lets the company provide 1+1 optical protection. Such a protection scheme relies on specialized optical equipment performing traffic bridging along the primary and secondary light paths, each of which follows diverse paths. Upon a link failure, such as a fiber cut or optical equipment failure, the receiving side quickly detects the failure and switches the traffic received from the primary light path to the secondary. This type of mechanism, usually qualified as "single-ended," is undoubtedly efficient because it does not require any extra signaling mechanisms or coordination between the sender and receiver (just the receiving side performs the switching function). Hence, the rerouting time is very fast (a few milliseconds). Moreover, a strictly equivalent quality of service (QoS) is guaranteed upon a network element failure because the secondary path is identical to the primary path (although it might be longer to be diverse from the primary path). On the other hand, this requires dedicating half of the fiber capacity for backup recovery. Furthermore, such a protection scheme implies that additional optical equipment needs to be purchased.

Hence, USCom decided to use all the network bandwidth to route the primary traffic and rely on some upper-layer protection mechanisms (see the section "Network Recovery Design") to offer equivalent rerouting time at significantly lower costs. All the light paths provided to the IP/MPLS layer for inter-Level 1 links and Level 1-to-Level 2 links therefore are unprotected. This is perfectly in line with the previously described core network overengineering strategy adopted by USCom.

Although DWDM offers the ability to provide high bandwidth in a very cost-effective fashion, it has a downside. Multiple links share some common resources and equipment whose failure may impact several links. This is called Shared Risk Link Group (SRLG), and the production design should take it into account.

Putting all this information together, you can see from Figure 3-6 how connectivity is typically achieved from a Level 3 to a Level 2 to a Level 1 POP.

Figure 3-6 *Inter-POP Connectivity Within the USCom Network*

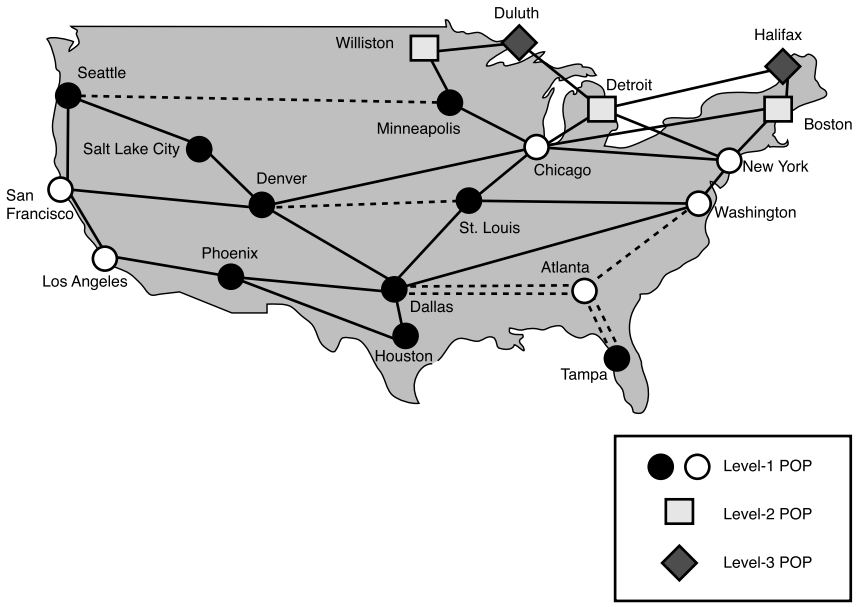


Table 3-1 summarizes the various types of links used in the USCom network, along with their main characteristics and localization.

Table 3-1 *Link Types and Characteristics in the USCom Backbone*

Link Type	Speed	Protection	Localization
OC-192 DWDM	10 Gbps	None	Level 1 POP-Level 1 POP
OC-48 DWDM	2.5 Gbps	None	Level 1 POP-Level 1 POP Level 1 POP-Level 2 POP
OC-48 SONET	2.5 Gbps	SONET protection	Level 1 POP-Level 2 POP Level 2 POP-Level 2 POP
OC-3 SONET	155 Mbps	SONET protection	Level 2 POP-Level 3 POP
Gigabit Ethernet	1 Gbps	None	Intra-Level 2 POP Intra-Level 3 POP

During the past several years, USCom has gathered various network failure statistics; they are summarized in Table 3-2. These statistics have been used to assess USCom's design requirements for its backbone network.

Table 3-2 *Link Failure Statistics Within the USCom Network*

Failure Type	Link/Router Type	Occurrence	Duration
Link failure	OC-3 SONET links	On average once a day in the network	From a few seconds to several days (fiber cut)
Link failure	OC-48 and OC-192 links	Unknown	Unknown
Router interface failure	Edge+core	Negligible	A few hours
Router failure (such as power supply, router software failure with traffic impact)	Edge+core	Once every two months	Variable
Router reboot (planned failure)	Edge (IA and VPN PE routers)	Once every six months	10 minutes
Router reboot (planned failure)	Core	Once a year	10 minutes

USCom's Network Design Objectives

USCom's objectives for its network design include the following considerations:

- Ensure that the Layer 3 MPLS VPN design can cope with current scale requirements as well as the predicted growth of this service over the coming years.

- Enrich the Layer 3 MPLS VPN service with quality of service commitments, allowing it to be marketed as the service of choice for enterprises that want to converge their data/voice/video networks onto a single multimedia intranet.
- Offer high-availability commitments to VPN users without additional capital expenses.

The following sections review the design elected by USCom, as well as the corresponding rationale behind the routing, Layer 3 MPLS VPN, QoS, traffic engineering, and network recovery. A final section points out a number of lessons that can be drawn from the USCom design.

Routing and Backbone Label Forwarding Design

All networks, whether they span whole continents or just a group of geographic regions, present design challenges that must be addressed by the network architects. Some issues are easier to tackle than others, and certain services present unique challenges. This section reviews how USCom decided to deploy its internal and external IP routing, and also how it decided to organize its Layer 3 MPLS VPN service.

We have established that USCom operates a national backbone infrastructure that spans the continental U.S. This network must support a number of different services, including Internet access and Layer 3 VPN service. During the initial Layer 3 VPN deployment, USCom decided to deploy MPLS technology to support the architecture specified in [2547bis]. This architecture provides a network-based VPN service. It was discussed in detail in Chapter 1, "Technology Primer: Layer 3 VPN, Multicast VPNs, IPv6, and Psuedowire."

Having deployed MPLS for this service, USCom also felt that it was the right technology to support fast rerouting (FRR) capability (which you'll read about in the "Network Recovery Design for Link Failures" section). Clearly, the network will need to support even more new services in the future, so USCom's selection of MPLS as its primary technology allows the company to support existing and future service requirements.

Label Distribution Protocol (LDP) is used within the backbone to allow label switching from one edge of the USCom network to the other. However, at this point in time, only the Layer 3 VPN traffic is label-switched, leaving the Internet traffic to be forwarded by normal IP forwarding procedures. The rationale behind the decision to separate VPN forwarding from standard IP forwarding was driven primarily by the desire to continue operating the Internet network in the exact same way as before Layer 3 VPN services were introduced. This avoided any changes in configuration, monitoring, troubleshooting, or any other operational procedures that were in place for Internet traffic. In addition to this, a number of technical challenges exist if Internet traffic is label-switched, including how the existing IP tools (such as NetFlow) might behave, and how network events such as denial of service (DoS) attacks can be tracked and resolved. Chapter 5, "Global Service Provider Design Study," shows how these issues can be overcome and the USCom plan to introduce these new technologies in the future.

From an internal routing perspective, USCom runs Intermediate System–to-Intermediate System (IS-IS) as its Interior Gateway Protocol (IGP), which carries the loopback interface addresses of the PE routers (IP and VPN PE routers) as well as internal link addresses. The number of internal routes is approximately 3000. USCom does not expect to have more than 1000 routers in the IS-IS routing domain within the next two years. Hence, the IS-IS network is a flat Level 2 network that avoids having to manage the complexity of multiple levels of hierarchy.

USCom measured that the flooding activity on the existing network was perfectly reasonable. The Shortest Path First (SPF) computation time was calculated on the order of 100 ms (usually closer to 60 ms), not including the routing table updates. If at some point in the future the number of IS-IS routers has to be drastically increased because of the activation of IS-IS on various edge devices such as the ADSL or Dial access routers, USCom might consider splitting the network into multiple levels (each POP would be the Level 1 hierarchy). This would be necessary to also preserve the network convergence times. (A detailed analysis of these aspects appears in [NET-RECOV].)

Separation of Internet and Layer 3 MPLS VPN Services

From a forwarding perspective, Layer 3 VPN traffic is separated from Internet traffic, where VPN traffic is label-switched across the USCom network and Internet traffic is IP-routed/forwarded. The PE routers serving VPN and Internet customers are also separate. This is primarily because the Internet service has been deployed for a number of years and USCom wanted to deploy the new Layer 3 MPLS VPN service as a separate project, without concern that it might affect the existing customer base.

The backbone network infrastructure is addressed from the 23.49.0.0/16 block. This includes all P routers, PE routers (whether Internet or Layer 3 VPN), and any other equipment within the USCom network. The P router and core-facing interfaces on the Internet and Layer 3 VPN PE routers take their addresses from the 23.49.0.0/21 range (providing IP addresses 23.49.0.1–23.49.7.254).

The Internet PE routers and IPv4 route reflectors (RRs) take their loopback interface addresses from the 23.49.8.0/22 range (providing IP addresses 23.49.8.1–23.49.11.254).

The Layer 3 MPLS VPN PE routers and VPNv4 RRs (used for the MPLS VPN service) take their loopback interface addresses from the 23.49.16.0/22 range (providing IP addresses 23.49.16.1–23.49.19.254). This block is large enough to address 1022 devices. If the service increases above this amount, the 23.49.20.0/22 range is made available.

Each Layer 3 MPLS VPN PE router has a loopback interface configured; it is used as the source address for all Multiprotocol BGP (MP-BGP) peering sessions. Likewise, each Internet access PE router has a loopback interface assigned; it is used as the source address for all IPv4 BGP-4 peering sessions.

USCom also evaluated using one of the private IP address blocks from the [PRIVATE] range for its internal infrastructure. The use of private addresses provides some protection from the Internet because it is not a routable address space. Therefore, the internal USCom network would theoretically be hidden from the outside. However, locally attached customers could still access the network—for example, by sending traffic via a default route to USCom. Therefore, the advantages of using private address space are mitigated. Also, a future acquisition of another company might present some integration challenges, so the use of private addresses for the design was rejected.

Because the Internet PE routers and Layer 3 MPLS VPN PE routers are separate, and because USCom chose to forward only VPN traffic through label switching, forwarding separation needs to occur at the LDP level. The default behavior of the LDP protocol when executing in frame-based mode is to create and distribute label bindings for every IGP learned or local (static or connected) prefix. This is unnecessary in the USCom network because only the VPN traffic is to be label-switched, and all Internet traffic is to be routed and will never need any of the allocated label space. Therefore, only the MPLS VPN PE router loopback interface addresses (255 currently) require label bindings, because they are the only destinations to which traffic is forwarded through label switching. Example 3-1 shows how LDP filtering is achieved.

Example 3-1 *Filtering Label Binding for PE Router Loopback Interfaces*

```
no tag-switching advertise-tags
tag-switching advertise-tags for ldp-pe-filter
!
ip access-list standard ldp-pe-filter
 ! Main IP VPN PE-router loopbacks
 permit 23.49.16.0 0.0.3.255
 ! Reserved IP VPN PE-router loopback block
 permit 23.49.20.0 0.0.3.255
```

NOTE Example 3-1 also shows that USCom has added the 23.49.20.0/22 address range to the filter. This range currently is not used in the deployed network. It is held in reserve in case the existing 23.49.16.0/22 address block becomes exhausted. Rather than updating the filter on all routers in the future, USCom chose to permit this block from Day 1 of the design.

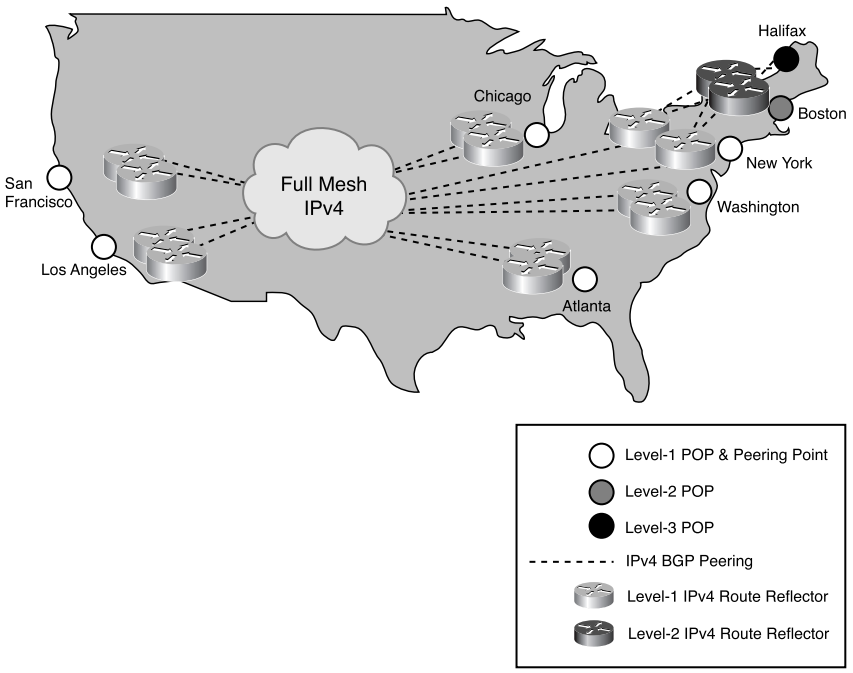
NOTE It is worth mentioning that LDP filtering needs to be activated on all the routers in the network, including the P routers, not just the PE routers. This prevents label space from being allocated unnecessarily throughout the network. It also prevents the forwarding of Internet traffic using label switching.

In the future, USCom may also decide to label-switch its Internet traffic. This may be achieved by either removing the LDP filtering (the configuration of which is shown in Example 3-1) or updating the LDP filter to include the Internet PE router loopback interface addresses.

Internet Service Route Reflection Deployment

The USCom RR design for Internet service is fairly typical. It follows the network's physical topology (for loop avoidance), as shown in Figure 3-7. (Only core POPs with external peering points are shown in the figure even though the design is relevant to all Level 1 POPs.) Each Level 1 POP has two Internet RRs (the backbone P routers). All Internet PE routers peer locally and are clients of these devices. All Level 1 POP RRs are fully meshed at the BGP-4 level. The aggregation P routers are also clients of these RRs.

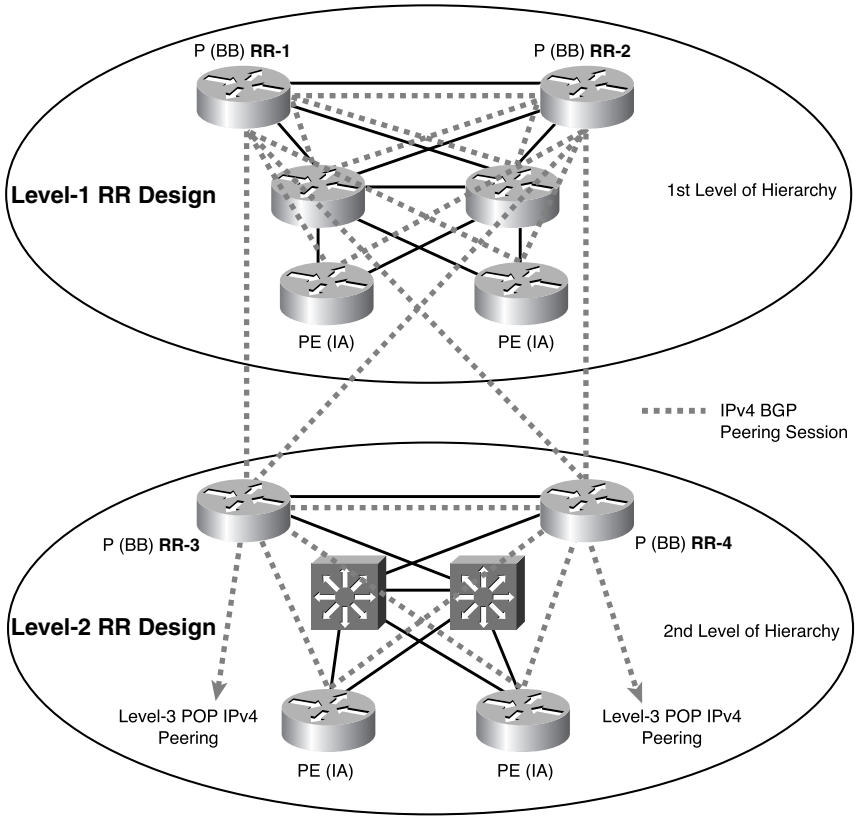
Figure 3-7 *Placement of IPv4 Route Reflectors for Internet Service*



A second level of RR hierarchy is deployed between the Level 1 and Level 2 POPs. Each Level 2 POP has two RRs (which again are the exiting backbone P routers); these are clients of their nearest Level 1 POP RRs. Every Internet PE router within a Level 2 POP is a client of the local Level 2 RRs. Each Level 3 POP Internet PE router and backbone P router peers with its nearest Level 2 POP RRs (once again following the network's physical topology).

Figure 3-8 shows how the IPv4 BGP peerings are arranged between different levels of POPs and the placement of the RRs within those POPs. Note that this figure provides the typical topology, although in some cases the Level 2 POP RRs may peer with different Level 1 POPs. (In other words, one RR peers with a different Level 1 POP than the other RR within the Level 2 POP.) This depends on the RR's geographic location in the overall topology.

Figure 3-8 *IPv4 POP-to-POP BGP Route Reflection*



The global IPv4 BGP table currently contains approximately 155,000 Internet routes.

Layer 3 MPLS VPN Service Design Overview

USCom's Layer 3 MPLS VPN service is designed to target the growing number of organizations that are outsourcing their information technology to a third-party service provider. It also targets organizations that want to move away from a traditional overlay Layer 2 VPN model (such as Frame Relay or ATM). This trend is primarily driven by cost reductions

for the end customer and the ability to receive additional services in a scalable manner, such as QoS and multicast. In many cases a hub-and-spoke topology, which is common if the environment is Frame Relay-based, is no longer sufficient to meet the end users' application requirements. The ability to use an infrastructure that inherently provides any-to-any connectivity is very attractive from an availability, scale, and service deployment perspective. The initial service offering addresses only the "unmanaged" market, where USCom provides network connectivity for the end user but the end user maintains control over their own routing. However, the design positions USCom to offer "managed" service, where it manages the end-user equipment, in the future.

USCom uses all 100 POPs to provide its Layer 3 MPLS VPN service. Customer access is via Frame Relay, ATM, leased line, and PoS. Access speeds range from *n* * DS0 (64 kbps) to OC-3 (low- to medium-speed) and from OC-3 to OC-48 (high-speed).

The current network deployment has 255 PE routers; this may be considered a dense deployment in the U.S. These are spread around all 100 POPs, with an average of two in each Level 3 POP, three in each Level 2 POP, and six in each Level 1 POP. However, PE routers are deployed based on customer demand at each location; therefore, the average numbers do not necessarily correspond to the actual deployed topology. For example, 30 of the Level 3 POPs currently have only one PE router deployed rather than two. On the other hand, the New York POP, which is a Level 1 facility, has ten PE routers, which is more than the average.

USCom initially defined two different types of customers who may access the national Layer 3 MPLS VPN service, as described in the following list. Note that Internet connectivity is considered a separate service and therefore is not bundled with the VPN service:

- **VPN intranet**—This customer requires connectivity between internal sites for the creation of an intranet. No extranet connectivity is provided. However, if the evolution of the USCom network introduces any central services (such as web hosting, firewalls, and so on), the customer is eligible for connectivity to these services.
- **VPN extranet**—This customer requires connectivity between internal and external partner sites for the creation of an extranet.

USCom breaks its VPNs into three categories—small, medium, and large—as described in Table 3-3. These categories are based on the customer's size as measured by the number of sites in the VPN. Current statistics show that 500 VPNs are deployed, with a combined total of 12,500 VPN sites, representing ten large VPNs, 200 medium VPNs, and 290 small VPNs. The VPN sites represent 62,500 total VPNv4 routes in the network.

Table 3-3 *IP VPN Categories*

VPN Category	Number of Sites	Percentage of Total Sites	Number of Prefixes in VPN	Percentage of Total Customers
Small VPN	2 to 10	15 %	Ones to tens	58%
Medium VPN	11 to 200	45%	Tens to hundreds	40%
Large VPN	201 to thousands	40%	Hundreds to thousands	2%

As you can see, although the majority of VPN customers fall within the small VPN category, they represent only 15 percent of the total number of sites. Only 2 percent of customers fall within the large VPN category, but they represent 40 percent of the total number of sites.

PE Router Basic Engineering Guidelines

Configuring a new Layer 3 MPLS VPN customer requires a set of engineering guidelines that is flexible and easy to implement from a centralized management system. A number of common attributes need to be configured for each new customer. These are outlined in Chapter 1 and can be summarized as follows:

- Definition and configuration of the Virtual Routing/Forwarding instance (VRF)
- Definition and configuration of the Route Distinguisher (RD)
- Routing protocol context and/or static routing configuration
- Import/export policies
- Interaction between the backbone control plane and the VRF
- Configuration and association of customer-facing router interfaces with previously defined VRFs
- Quality of service (QoS) policies

The values chosen for each VRF attribute, as well as specifics of the routing protocol context, vary from VPN customer to customer. However, because a number of default attributes can be assumed, the provisioning system needs only to provide a template that can accept different values on a per-customer basis. Most commercially available provisioning systems today have default templates for configuring these attributes.

In most cases, the PE-CE links used for the Layer 3 MPLS VPN service have their IP addresses allocated from the USCom IP address space. However, on an exception basis, customer address space is sometimes used. The decision to use customer address space is primarily driven by the customer's size and topology and whether the customer's address space can be summarized into convenient blocks. If customer address space is used, USCom requires that it be a globally assigned block and not from the private range so as to avoid any potential address range clash with other VPN customers.

In all cases, USCom uses the Interface Group MIB (see [IF-MIB]) to monitor the status of the physical PE-CE links. This is achieved by polling the *IfOperStatus .ifIndex* object, the details of which are specified in [IF-MIB].

USCom will not provision more than 15 percent of the total access links of any given customer onto a single PE router. This will help prevent a large percentage of the VPN from losing connectivity in the event of a PE router hardware failure or planned maintenance. Although USCom has not reached the scaling limits on any of its PE routers, it has decided to apply an upper boundary to the total number of Layer 3 MPLS VPN customer accesses per PE router. This is driven by a number of factors, including the type of access device (such as the router's

size and capability), traffic throughput requirements, additional service requirements (such as QoS), and so on. Table 3-4 provides an overview of USCom's engineering rules in this space for two of its router platforms. USCom will stop adding customers on each platform if any of the hard limits is reached.

Table 3-4 *PE Router Sizing Rules*

Engineering Parameter Limits	Platform 1 Limits	Platform 2 Limits
USCom IGP routes	3000	3000
Number of iBGP/eBGP peers	350	2000
Number of VRFs	200	1000
Total VRF routes	60,000	300,000
Average number of sites per VRF	3	3
Average number of routes per VPN	300	300
Total CE connections per PE	600	3000

Table 3-4 also shows that because the typical average number of sites per VRF per platform is 3, this translates into a total number of CE connections per PE of 600 and 3000.

VRF Naming Convention

When choosing a name for a given VPN VRF, it is important to remember that the network operations staff will use the name to troubleshoot connectivity problems for the VPN. Several naming conventions might be adopted. USCom chose to use a representation of the name followed by an abbreviation of the customer name, starting with a VRF name of V101 and incrementing it by 1 for each new VPN deployed. This allocation scheme is shown in Table 3-5.

Table 3-5 *VPN Name Allocation Scheme*

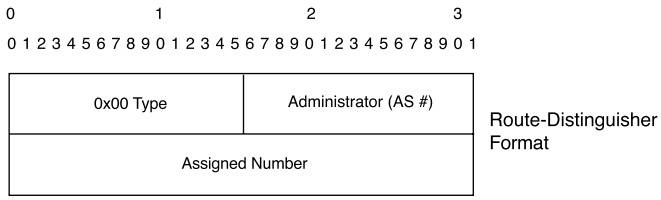
Customer Name	VRF Name
U.S. Post Office	V101:USPO
SoccerOnline International	V102:SoccerOnline
BigBank of Massachusetts	V103:BigBank
<Next customer>	V104 and so on

Route Distinguisher Allocation

The *route distinguisher* (*RD*), as described in [2547bis], is an 8-byte entity that lets the MPLS VPN architecture uniquely identify a particular route within the operator's backbone network. The structure of the RD depends on the type specified in the first 2 bytes of the attribute.

USCom chose to use its autonomous system number plus a uniquely defined number specific to a given VPN customer. Figure 3-9 shows the format of the RD chosen by USCom.

Figure 3-9 *Route Distinguisher Format*



In theory, several schemes are available when choosing an RD allocation method. The main ones can be summarized as follows:

- **Use a unique RD for every VPN**—A unique RD for each VPN is the easiest option to deploy because every PE router uses the same value for a given VPN customer. However, deploying this scheme prevents the operator, which has VPNv4 route RRs in its topology (which is typically the case; USCom has such a topology), from offering load-balancing services to customers who are dual-homed to the Layer 3 MPLS VPN service. This is because the VPNv4 routes cannot be guaranteed to be unique. Therefore, certain paths may be unavailable to the PE routers that will perform the load balancing, because the RRs will advertise only the "best" path.
- **Use a unique RD for each VPN on a per-PE basis**—A unique RD may be allocated to each VPN at each PE router, although the value may be different between PE routers for a given VPN. In this case the operator can provide load-balancing services when RRs are deployed. This is because the VPNv4 routes can be guaranteed to be unique within the MPLS backbone. Note that such a scheme requires a little more memory space to store the additional VPNv4 routes.
- **Use a unique RD for each VPN on a per-interface/per-PE router basis**—A unique RD may be allocated for each VRF on a per-interface basis. The advantage is that a particular site within a VPN can be identified based on the RD value of any route originated by that site. However, other methods are available to achieve the same aim, such as use of the *site of origin* (*SoO*) attribute, which is much less resource-consuming. The format of this attribute can be found in [EXTCOM].

USCom chose to use a unique RD per VRF (the second option), because it required load-balancing services for a number of VPN customers with dual-homed CE routers. Although this scheme requires additional memory at the PE routers, the ability to provide load balancing when RRs are deployed was necessary to address USCom customer requirements. The range of RDs available is 32765:1 through 32765:4,294,967,295, which is way beyond what USCom will ever require.

NOTE Traffic load balancing and its implications for the service provider backbone network are discussed in more detail in the section "Load Balancing Support" in Chapter 4, "National Telco Design Study."

Route Target Allocation for Import/Export Policy

Selecting a route target for each VPN is necessary to specify that VPN's specific import/export policies. [EXTCOM] specifies three main formats that may be used for the route target extended-community attribute.

USCom chose to use the two-octet AS format with its own AS number, 32765, as the ASN portion of the community. Use of any customer AS numbers was rejected in the design because the possibility of conflicting numbers was apparent if any VPN customers were using private AS numbers from the [64512–65535] range.

Route target values 32765:[1–100] were reserved for future use, so values 32765:101 through 32765:65535 are available for VPN customer allocation. This fits nicely with USCom's VRF naming convention, in which it maps the VRF name to the number in the route target. For example, BigBank of Massachusetts, whose VRF name is v103:BigBank, uses a default route target value of 32765:103.

NOTE Some VPN customers may require the use of more than one route target per VRF. An example is a topology in which the spoke sites require connectivity to a central service. This type of topology is often called central services or hub and spoke. The mapping of the VRF name with the route target cannot be used in this case.

Basic PE Router Configuration Template

Example 3-2 provides the basic PE router configuration template used by USCom.

Example 3-2 *PE Router Configuration Template*

```
hostname USCom.cityname.PErouter-number
!
ip vrf vpn-name
 rd 32765:1-4294967295
 route-target export 32765:101-65535
 route-target import 32765:101-65535
!
interface Loopback0
 description ** interface used for BGP peering **
 ip address 23.49.16.0/22 range address and network mask
!
```

PE Router Control-Plane Requirements

One of the most significant challenges for any Layer 3 network-based VPN service is distributing customer-specific routing information between edge routers and achieving this in a scalable manner. As the service grows, more and more VPN routes need to be advertised using the backbone control-plane infrastructure. The amount of information could become significant as the service becomes more and more successful.

Although USCom's future expansion projections for its Layer 3 MPLS VPN service do not indicate any kind of saturation point in terms of routing information capacity, it is clear that over time, as the service matures, the design of the backbone control-plane infrastructure will be critical.

The current network deployment has 255 PE routers providing Layer 3 MPLS VPN services. With this number of PE routers, and the requirement to carry an ever-expanding VPNv4 address space, USCom chose to deploy VPNv4-specific RRs (the details of which are discussed in section "VPNv4 Route Reflector Deployment Specifics") to help scale the distribution of routes. RRs help scale the network infrastructure in a number of ways. You will see in other chapters that additional functionality may be added to further increase this scaling. However, USCom chose to use RRs primarily to ease the network's operational complexity as the number of MP-BGP TCP sessions required by the PE routers into the backbone could be reduced to two (one to each RR), as opposed to every other PE router in the network.

Each PE router is required to maintain at least two MP-BGP peering sessions into the USCom backbone network. These sessions will be used to exchange VPNv4 prefix information with other PE routers via the VPNv4 RRs. Two sessions are necessary for redundancy in case an RR fails or connectivity to that RR becomes unavailable.

PE Router Path MTU Discovery

In Cisco IOS, by default, all PE routers have Path MTU Discovery [see PMTU] disabled. This means that the default TCP Maximum Segment Size (MSS) is used for all TCP sessions. This default normally is based on the outgoing interface MTU size minus the IP header/options and TCP header/options (for a total of 40 bytes). For example, for an Ethernet interface with an MTU of 1500 bytes, the MSS is calculated as 1460.

BGP on Cisco routers uses a default MSS value of 536 bytes regardless of the outgoing interface type. The problem with this small value is that BGP signaling information sent across a given BGP session needs to be segmented into a much higher number of packets, substantially increasing convergence times. However, [PMTU] provides a mechanism in which the PE router can discover the optimum MSS to use for its BGP sessions, and therefore reduce the number of

messages generated. USCom enables [PMTU] on all its VPN PE routers, and VPNv4 RRs, using the configuration shown in Example 3-3.

Example 3-3 *PE Router PMTU Configuration Template*

```
hostname USCom.cityname.PErouter-number
!
ip tcp path-mtu-discovery
```

VPNv4 Route Reflector Deployment Specifics

Two tools are available to assist in the scaling of the TCP sessions required to support the VPNv4 address family for Layer 3 MPLS VPN service—*confederations* and *route reflectors*. USCom chose to deploy RRs in its Layer 3 MPLS VPN design; these are completely separate from the RRs used for its Internet service. This separation provides improved convergence times, as well as scalability in terms of CPU and hardware memory requirements. USCom did not have a requirement to deploy confederations, because its service requirements did not necessitate multiple sub-autonomous systems to split the MP-BGP topology.

While reviewing the needs of the Layer 3 MPLS VPN service from a control-plane perspective, it was clear that the rules for RR deployment were different from those followed for the Internet service, because the VPN traffic would be label-switched rather than routed. The primary difference between label switching and IP forwarding within the backbone network is that label switching allows the RRs to be deployed outside the packet-forwarding path, because the forwarding decision for a given packet is made at the edge of the network rather than on a hop-by-hop basis. This paradigm is a little different from the typical Internet design used for IPv4 route distribution, in which the common practice is to place the RRs so that they follow the network's physical connectivity. This type of design avoids any forwarding loops that could be caused by bad route placement. Figure 3-10 shows what can happen if these rules are not followed when forwarding IP traffic natively instead of via label switching.

Figure 3-10 *Forwarding Loop with Incorrectly Designed IPv4 Route Reflection*

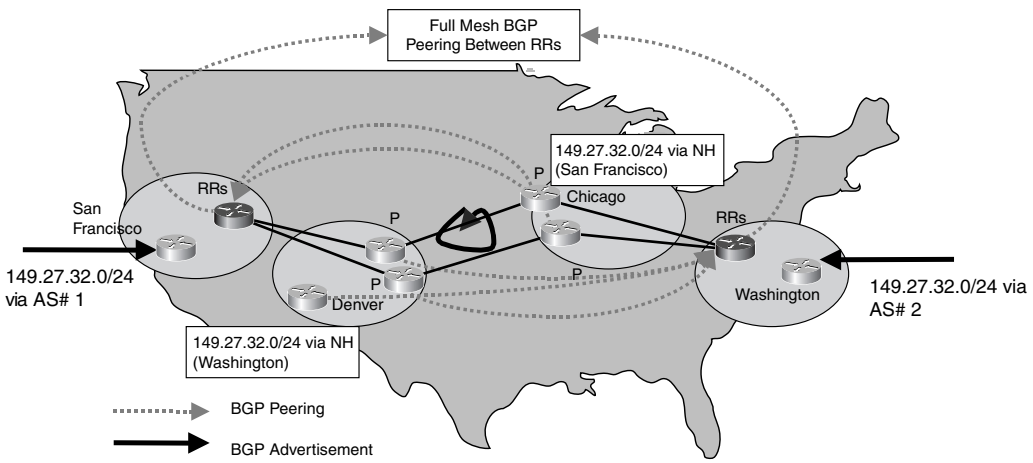


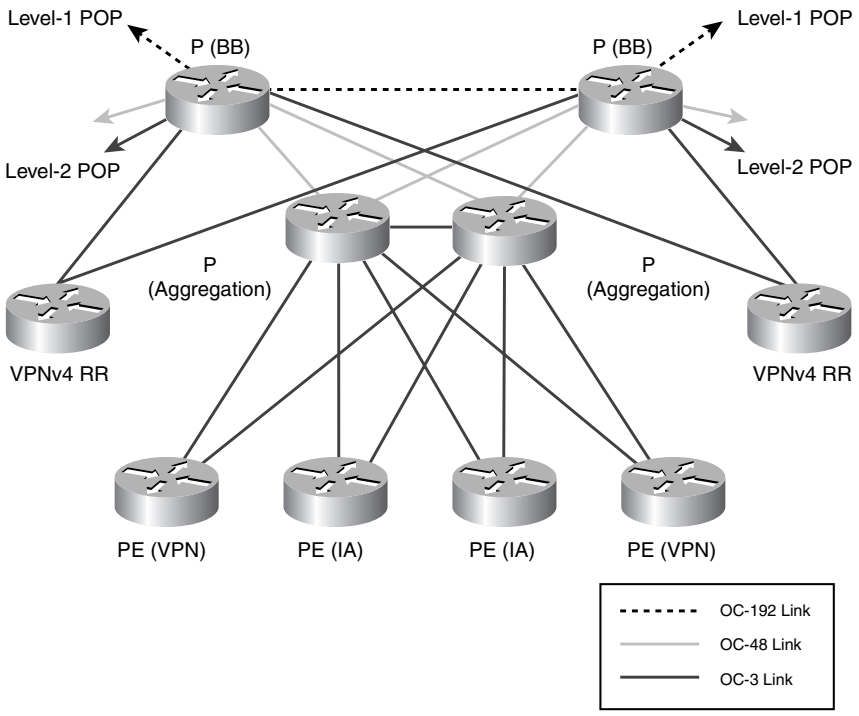
Figure 3-10 shows that the Denver POP believes that 149.27.32.0/24 can be reached via a Washington next-hop address, whereas the Chicago POP believes it can be reached via a San Francisco next-hop address. This is clearly a bad design, because both POPs should peer with their geographically closest RRs. In this case, packets loop between Chicago and Denver.

This issue is eliminated when packet forwarding is achieved through label switching (or IP tunneling), because the packets' original destination IP address is no longer examined in the network core.

Deployment Location for VPNv4 Route Reflectors

The Internet RR topology described previously is not well suited to the Layer 3 MPLS VPN service because the topology assumes that all RRs will carry the same set of routes. This may not be necessary, or even desirable, for the VPN service, because not all PE routers will need the same set of routes. Hence, it would not scale as well as a partitioned VPN RR design, which may become necessary for a large-scale VPN topology. Another drawback of this Internet topology for the VPN service is that it introduces a number of BGP hops that increase the convergence delay for routing updates. This may be detrimental to the Layer 3 MPLS VPN SLA, and therefore the topology of the VPNv4 RRs is a little different, as shown in Figure 3-11.

Figure 3-11 *VPNv4 Route Reflector Deployment*



The VPNv4 RRs are deployed only within Level 1 POPs and are connected directly to both core backbone P routers via OC-3 links. The topology does not follow the network's physical path; this is unnecessary because of the deployment of LDP in the backbone. Level 2 and Level 3 POPs do not house any RRs but instead peer directly with their local Level 1 POP. All VPNv4 RRs peer within a full mesh.

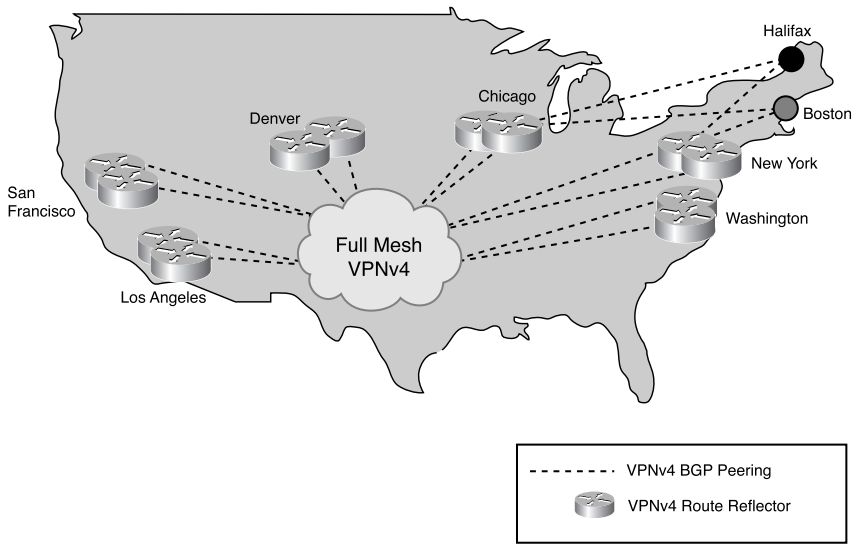
The initial deployment has VPNv4 RRs in six locations: San Francisco, Los Angeles, Denver, Chicago, New York, and Washington. Each PE router has peering sessions with a pair of RRs that are within its local regional vicinity. For example, a PE router in Boston may peer with an RR in New York and another in Chicago. A maximum of 200 peering sessions has been defined within the engineering deployment guidelines for the RRs. Although the currently deployed hardware could support more than this number, USCom has validated only up to 200 peering sessions within its labs. Because the current network has 255 existing VPN PE routers, and each PE router peers to local RRs based on geography, no RR within the topology has close to this maximum number of peering sessions.

USCom takes advantage of *update groups,* which are enabled by default in the level of Cisco IOS it is running on its routers. Therefore, USCom can dynamically build groups of MP-BGP peering partners that have the same outbound policy. The update group does not consider the extended communities used by the PE routers for import/export policy; therefore, all the PE routers belong to the same group. This provides the ability to build one MP-BGP update (instead of one per PE router) and to replicate it to all members of the update group. This functionality provides improved performance at the RRs. Each RR, just like the PE routers, also uses [PMTU].

The design of the control plane must provide the ability for all Layer 3 MPLS VPN PE routers to learn routes from the centralized VPNv4 RRs. Ideally each PE router should peer based on geography as much as possible, and to different Level 1 POPs. This is useful so that a particular Level 2/3 POP does not lose all routing information in the event of a catastrophic failure within a given Level 1 POP, such as a complete power outage.

Figure 3-12 illustrates the topology of the VPNv4 RRs. It shows that the Boston Level 2 POP peers to both the Chicago and New York Level 1 POPs to provide geographic redundancy. All PE routers within a Level 1 POP (for example, the New York POP) peer to their local RRs, because a local power failure would mean that they would not be able to maintain a peering session with another Level 1 POP because all connectivity would be lost. Therefore, there is little point in following the same design rule as the Level 2/3 POPs.

Figure 3-12 *Physical Topology of VPNv4 Route Reflection*



Preventing Input Drops at the VPNv4 Route Reflectors

Each RR has an interface-level queue, referred to in Cisco IOS as the *input hold queue,* that may not by default be large enough to prevent input drops at the interface. Dropping TCP packets reduces the protocol's efficiency and causes retransmissions to occur. This behavior can slow down the convergence of MP-BGP at the VPNv4 RRs. For this reason, USCom tunes the queue value using the following algorithm:

> Input hold queue = (TCP window size / mss) * number of MP-BGP peers
> where TCP window size is the TCP window size for the MP-BGP session, mss is the TCP maximum segment size, and number of MP-BGP peers is the number of route reflector clients.
> The window size (sndwnd) and mss (max segment size) values can be found using the **show ip bgp neighbor** command in Cisco IOS.

PE Router and Route Reflector VPNv4 MP-BGP Peering Template

USCom uses the template shown in Example 3-4 for the VPNv4 MP-BGP configuration of the PE routers and RRs.

Example 3-4 *PE Router and Route Reflector VPNv4 BGP Configuration Template*

```
! PE-router configuration
hostname USCom.cityname.PErouter-number
!
ip tcp path-mtu-discovery
```

continues

Example 3-4 *PE Router and Route Reflector VPNv4 BGP Configuration Template (Continued)*

```
!
interface Loopback0
 description ** interface used for BGP peering **
 ip address 23.49.16.0/22
!
router bgp 32765
 no bgp default ipv4-unicast
 neighbor 23.49.16.0/22 address-for-1st-RR remote-as 32765
 neighbor 23.49.16.0/22 address-for-1st-RR update-source Loopback0
 neighbor 23.49.16.0/22 address-for-1st-RR remote-as 32765
 neighbor 23.49.16.0/22 address-for-1st-RR update-source Loopback0
 ..
 !
 address-family vpnv4
 neighbor 23.49.16.0/22 address-for-1st-RR activate
 neighbor 23.49.16.0/22 address-for-1st-RR send-community extended
 neighbor 23.49.16.0/22 address-for-1st-RR activate
 neighbor 23.49.16.0/22 address-for-1st-RR send-community extended
 ..
 exit-address-family

! VPNv4 Route Reflector configuration
hostname USCom.cityname.RRrouter-number
!
interface Loopback0
 description ** interface used for BGP peering to RR-clients **
 ip address 23.49.16.0/22 range address and network mask
!
router bgp 32765
 neighbor 23.49.16.0/22 address-1st-PE-router remote-as 32765
 neighbor 23.49.16.0/22 address-1st-PE-router update-source Loopback0
 ..
 !
 address-family vpnv4
 neighbor 23.49.16.0/22 address-1st-PE-router activate
 neighbor 23.49.16.0/22 address-1st-PE-router route-reflector-client
 neighbor 23.49.16.0/22 address-1st-PE-router send-community extended
 ..
 exit-address-family
```

PE-CE Routing Protocol Design

As discussed in Chapter 1, various routing protocols (and static routing) are available for connectivity between the CE routers and PE routers. When assessing the requirements for the Layer 3 MPLS VPN service design, the set of routing protocols to offer was evaluated. The initial deployment of the service included static routing and BGP-4 support only on the PE-CE links. However, RIPv2 was added fairly shortly afterwards to provide service to customers who

were unable to run BGP-4, such as those with PE-CE links backed up via ISDN to a Network Access Server (NAS).

USCom avoids using RIPv2 as much as possible because of its periodic update behavior and the implications of this on the PE routers' CPU cycles. For customers who require RIPv2, USCom configures **flash-update-threshold 30** to prevent Flash updates from being sent before the regular periodic updates. Flash updates send new routing information as soon as something changes in the customer topology and therefore can increase CPU requirements substantially during customer routing instability events. Also, USCom imposes the use of BGP-4 for dual-attached sites to avoid having to configure RIP tagging for loop prevention.

Static Routing Design Considerations

VPN sites that are single-homed to the USCom network may use static routing. However, this depends on the number of routes (a low number is mandatory, usually no greater than 5) and whether these routes are likely to change on a regular basis. Static routing is particularly suitable if route summarization is easily achievable for the set of routes that can be reached for a particular VPN site. In the majority of cases, only a few routes can be accessed via a single-homed site, such as a local /24 LAN segment, so static routing is adequate.

Clearly static routing does not provide any dynamic rerouting capability. Although static routing provides good stability while requiring minimal router resources, USCom actively encourages its larger Enterprise customers to run a dynamic routing protocol. The overhead of managing static routing in this case is considerable, especially at the central sites, where route summarization is often impossible.

In many cases, even if the customer has only a single connection to the Layer 3 MPLS VPN service, if the customer takes Internet service from somewhere else within the site, whether from USCom or some other Internet service provider, it is likely that the customer will follow a default route toward the Internet exit point. This means that the CE router needs to have *all* the relevant static routes from the VPN pointing toward the PE router. An appropriate addressing scheme that allows some summarization simplifies the configuration exercise but nevertheless is prone to errors and typically is avoided.

For stability reasons, USCom prefers to configure the static routes with the **permanent** keyword. This prevents the static routes from being withdrawn in MP-BGP in the event that a PE-CE link flaps or fails. The downside of this design decision is that traffic continues to be attracted toward the failed link, even if the PE router is unable to forward traffic from other sites across the link. However, because the customer site is single-homed, the added backbone stability is preferred over the suboptimal (unnecessary) packet forwarding.

Current statistics show that approximately 40 percent of USCom's PE-CE connections use static routing.

PE-CE BGP Routing Design Considerations

50 percent of VPN PE-CE connections use external BGP (eBGP). This is the protocol of choice for USCom, because it is used to dealing with this protocol (with experience from the Internet service), and it can easily add policy on a per-VPN basis. Some end users are already familiar with the BGP protocol and have been running it within their network before migrating to the VPN service, although this is normally restricted to large Enterprises. Also, many of these end users already subscribe to an Internet service and therefore are familiar with how the protocol is used. Therefore, standardizing on BGP is an obvious choice.

To protect the PE routers, every customer BGP-4 peering session is configured to accept only a maximum number of prefixes. This is achieved through the use of the **neighbor maximum-prefix** command on each PE-CE BGP peering session. USCom also uses *route dampening* (with the same set of parameters) for all its customers who attach to the VPN service via external BGP. This is stringently applied to all customers because route flaps (constant routing information changes) can cause instability in the control plane of the USCom network. The policy applied for dampening is as follows: Any route that flaps receives a *penalty* of 1000 for each flap. A *reuse limit* of 750 is configured so that a route, once suppressed, can be readvertised when the limit reaches 750. After a period of 15 minutes (the *half-life time*), the total value of the accumulated *penalty* is reduced in value by 50 percent. If the accumulated penalty ever reaches a *suppress limit* of 3000, MP-BGP suppresses advertisement of the route regardless of whether it is active.

Both of these parameters are configured using the template shown in Example 3-5.

Example 3-5 *Restricting the Number of Prefixes on PE-CE BGP Links Template*

```
router bgp 32765
 address-family ipv4 vrf vrf-name
  neighbor 23.50.0.6 remote-as customer-ASN
  neighbor 23.50.0.6 activate
  neighbor 23.50.0.6 maximum-prefix 100
  no auto-summary
  no synchronization
  bgp dampening route-map vpn-dampen
  exit-address-family
!
route-map vpn-dampen permit 10
 set dampening 15 750 3000 60
```

NOTE USCom uses the same set of dampening parameters for all eBGP PE-CE peering sessions. It also uses a route map for ease of provisioning. The parameters contained in the route map are inherited by all customer accesses that use external BGP.

The maximum prefix setting is determined at service provisioning time. It differs from customer to customer.

NOTE USCom currently does not tune any of the BGP timers to decrease convergence times.

PE-CE IGP Routing Design Considerations

In recent months USCom has seen an increase in the number of customers requesting either OSPF or EIGRP support on their PE-CE links. These customers typically have large, and often complex, IGP topologies.

A number of benefits may be gained by running IGP on the PE-CE links:

- The service provider MPLS VPN network may be used for WAN connectivity while remaining within the customer's IGP domain. This provides a "drop and insert" approach to migrating the existing network onto the new infrastructure.
- A relatively seamless routing domain from the attached customers' perspective may be obtained. This avoids the extra costs associated with staff retraining to support an additional routing protocol such as BGP-4.
- IGP fast convergence enhancements can be deployed, especially in the case of multihomed sites, which may be useful in the case of a PE router or PE-CE link failure.
- External routes can be prevented within the IGP topology.
- IGP routing metrics can be maintained across sites, and the USCom network can remain transparent to the end user from a routing perspective.
- In the presence of customer back-door links (direct connectivity between customer sites, such as via leased lines), superior loop-avoidance and path-selection techniques can be used, such as sham links (OSPF) and site of origin (EIGRP).

A provider could offer a specific routing protocol as the only choice to avoid the costs associated with provisioning, maintaining, and troubleshooting different routing protocols. However, such an offering might force the VPN customers to compromise their design requirements and would ultimately hurt the provider through restriction of its customer base. If multiple routing protocol choices are to be offered on the PE-CE links, it is important to carefully consider the convergence characteristics (which are important to the VPN customer) and the service's scalability (which is important to both VPN customer *and* service provider).

USCom chose to offer RIPv2, EIGRP, and OSPF, all of which are provided on a restricted basis (in terms of the number of sites permitted to attach to a given PE router for each protocol). These restrictions are currently set at 25 for each protocol, although this figure is not a hard rule. It depends on the specific customer attachment needs (such as the number of routes and so forth) and is monitored to obtain more deployment experience. The IGPs are configured on a per-customer basis. The complexity of the configuration is driven by the complexity of the attached customer topology.

Specifics of the OSPF Service Deployment

USCom currently has two large customers who run OSPF on their PE-CE links. A number of features are included in the service provider design at the PE routers to support these customers.

A different OSPF process ID is used for each VPN. By default the same process ID is used for the VPN on all PE routers that have attached sites for that VPN. This is important. Otherwise, the OSPF routes transported across the MPLS VPN network are inserted as external routes (Type 5 LSAs) at a receiving OSPF site. This is typically undesirable because externals are by default flooded throughout the OSPF domain. Using the same process ID causes the PE router to generate interarea (Type 3 LSAs) routes instead, which are not flooded everywhere and therefore are bounded.

USCom uses the following command for *all* OSPF deployments. It protects the PE router from a large flood of Link-State Advertisements (LSAs) from any attached CE router.

```
[no] max-lsa maximum [threshold] [warning-only] [ignore-time value]
  [ignore-count value] [reset-time value]
```

Restricting the number of LSAs at the PE router is important because it protects the OSPF routing process from an unexpectedly large number of LSAs from a given VPN client. That might result from either a malicious attack or an incorrect configuration (such as redistributing the global BGP-4 table into the customer OSPF process).

Using this functionality, the PE routers can track the number of non-self-generated LSAs of any type for each VPN client that runs OSPF on the PE-CE links. When the maximum number of received LSAs is exceeded, the PE router does not accept any further LSAs from the offending OSPF process. If after 1 minute the level is still breached, the PE router shuts down all adjacencies within that OSPF process and clears the OSPF database.

USCom leaves the threshold, ignore time, ignore count, and reset time at their default values of 75 percent, 5 minutes, 5, and 2 times ignore time, respectively. Because only two OSPF clients exist at this time, the maximum LSA count is set to 10,000. USCom will continue to monitor this as new OSPF deployments arrive so as to optimize the default value.

Each router within an OSPF network needs to hold a unique identifier within the OSPF domain. This identifier is used so that each router can recognize self-originated LSAs and so that other routers can know during routing calculation which router originated a particular LSA. The LSA common header has a field known as the *advertising router* . It is set to the originating router's router ID.

The router ID used for the VRF OSPF process within Cisco IOS is selected from the highest loopback interface address within the VRF or, if no loopback interface exists, the highest interface address. This may be problematic if the interface address selected for the router ID fails, because a change of router ID is forced, and the OSPF process on the router must restart, causing a rebuild of the OSPF database and routing table. This clearly may cause instability in the OSPF domain. Therefore, USCom allocates a separate loopback address for each VRF that has OSPF PE-CE connectivity. This address is used as the router ID as well as for any sham links that may be required.

Specifics of the EIGRP Service Deployment

USCom found that a number of large Enterprise customers requested EIGRP connectivity with their PE routers. This protocol is widely deployed within Enterprise networks. Therefore, USCom felt that offering support for this protocol was a "service portfolio" differentiator. USCom deploys a number of features at the PE routers to support this protocol.

Automatic summarization is disabled as a matter of course for all EIGRP customers. The default behavior is for this functionality to be enabled. However, because the MPLS VPN backbone is considered transparent, USCom uses the **no auto-summary** command to disable it.

To support external routes within a customer EIGRP domain, a default metric of 1000 100 255 100 1500 is used, but this may be changed on a per-customer basis.

USCom supports the EIGRP site-of-origin (SoO) cost community. This community attribute is applied automatically at the point of insertion (POI) (the originating PE router) when an EIGRP route is redistributed into MP-BGP. Supporting this functionality allows USCom to support back-door links within a customer EIGRP topology by affecting the BGP best path calculation at a receiving PE router. This is achieved by carrying the original EIGRP route type and metric within the MP-BGP update and allowing BGP to consider the POI before other comparison steps.

USCom also supports the SoO attribute. This is configured by default for every site that belongs to a given EIGRP customer. This feature allows a router that is connected to a back-door link to reject a route if it contains its local SoO value. Example 3-6 shows this default configuration.

Example 3-6 *EIGRP SoO Attribute Configuration Template*

```
interface Serial 1/0
 ip vrf forwarding vrf-name
 ip vrf sitemap customer-name-SoO
!
route-map customer-name-SoO permit 10
 set extcommunity soo per-customer-site-id
 exit
```

USCom protects the PE routers from saturation of routing information by using the maximum-prefix feature. The following shows the syntax of this command:

```
maximum-prefix maximum [threshold] [warning-only]
  [[restart interval] [restart-count count] [reset-time interval] [dampened]]
```

At this point in time the default values for *threshold,* **restart**, **restart-count**, and **reset-time** are used. These values are 75 percent, 5 minutes, 3, and 15 minutes, respectively.

NOTE It is worth noting that running an IGP between the PE router and the CE router requires some significant extra configuration for USCom.

IP Address Allocation for PE-CE Links

USCom decided within its design that it would allocate the PE-CE link IP addresses from one of its registered blocks. This allows more flexibility in determining a filtering template that can be applied to all PE routers so that unwanted traffic can be dropped at the edge. It also avoids any conflicts with customers' IP address space, because many will have selected IP addressing from the [PRIVATE] private ranges.

The block of addresses chosen for this purpose is taken from the 23.50.0.0/16 address block. Because the customer access routers are unmanaged, each PE-CE link is assigned a 255.255.255.252 network mask that allows two hosts. For example, 23.50.0.4/30 provides IP addresses 23.50.0.5 and 23.50.0.6 with which to address the PE-CE link of a given VPN customer. These addresses are redistributed into MP-BGP so that they are available within the VPN for troubleshooting purposes.

NOTE USCom also decided to allow customer address space for the PE-CE links. However, this would be on an exception basis, and the IP addresses must be from a registered block.

Controlling Route Distribution with Filtering

Each PE router within the USCom network has finite resources that are distributed between all services that are offered at the edge. Because many VPN clients will access the network via the same PE routers, USCom would like to be able to restrict the number of routes that any one customer can carry within its routing table. This is achieved by applying the maximum routes command to all VRFs, as shown in Example 3-7.

Example 3-7 *Maximum Routes Configuration Template*

```
hostname USCom.cityname.PErouter-number
!
ip vrf vpn-name
 rd 32765:1-4294967295
 route-target export 32765:101-65535
 route-target import 32765:101-65535
 maximum routes maximum-#-of-routes {warning-threshold-% | warning-only}
```

USCom considered what values should be set within this command. It noticed that if the value of the limit imposed were set too low, valid routes would be rejected, causing a denial of service for some customer locations. Also, USCom noted that the **maximum routes** value must be able to cater to all types of routes injected into the VRF, including static routes, connected routes, and routes learned via a dynamic protocol. USCom decided to start with a **maximum routes** limit that was set for each VRF to be 50 percent more than the actual number of routes in steady state, with a warning at 20 percent more than the actual number of routes in steady state.

NOTE When a link-state IGP, such as OSPF, is run on the PE-CE links, restricting route input to the VRF does not stop the link-state database from being populated. Therefore, additional protection mechanisms are required. These are discussed in the "Specifics of the OSPF Service Deployment" section earlier in this chapter.

USCom decided not to use any filtering for customer route distribution during its initial deployment of the Layer 3 MPLS VPN service. Because of this, all RRs carry the same set of routes, and each PE router relies on the Automatic Route Filtering (ARF) feature to ignore any routing updates that contain routes that are not locally imported into any attached VRFs.

Security Design for the Layer 3 MPLS VPN Service

Security of the network infrastructure is one of the most important considerations when designing any robust network. [ISP-security] provides an excellent overview of security best practices for ISP networks. Most of the material presented is also relevant to the USCom Layer 3 MPLS VPN service, because it presents basic router security tips, and USCom already follows these for its Internet service.

Although the Layer 3 MPLS VPN service separates customer routing from backbone routing, existing tools such as traceroute provide a method of revealing the core topology of the USCom network from within a customer VPN. Because of this, USCom chose to disable this behavior through the use of the **mpls ip propagate-ttl forwarded** command throughout the network. This command is discussed in detail in Chapter 13 of [VPN-Arch-Volume-1]. It basically has two effects: It propagates the IP TTL into the label header, and it propagates the label TTL into the IP header (where the packet exits the MPLS backbone). By disabling the TTL propagation (via the **no mpls ip propagate-ttl forwarded** command), USCom can hide its internal infrastructure from the output of any customer-initiated traceroutes that are sourced from within a Layer 3 VPN. This command, however, does not protect any of the Internet PE routers because packets are IP-forwarded rather than label-switched toward the Internet. Therefore, USCom has a policy of not advertising the IP address blocks used for its internal infrastructure toward the Internet. It also applies packet filters toward these addresses at the external-facing interfaces of the Internet PE routers.

Although the core addressing is hidden from traceroute through the use of the **no mpls ip propagate-ttl forwarded** command, the same cannot be guaranteed for the subnet used for PE-CE circuit addressing. DoS attacks can always be performed if the VPN client knows one or more of the IP addresses of the PE router. This is easy to determine through the use of the **traceroute** command. Visibility of PE router circuit information could allow the VPN client to intrude or perform DoS attacks on the PE router.

If the CE router is managed, which is not the case for USCom in its initial deployment, the PE router circuit address can be hidden by a filter that prevents it from being redistributed into the

customer network. In addition, various inbound filters can be applied at the PE router to restrict CE router access; this is what USCom has adopted for its unmanaged service. Example 3-8 shows the filter template chosen for deployment.

Example 3-8 *PE-CE Link Filter Template*

```
ip access-list extended PE-CE-Filter
 permit icmp host CE-host-address host PE-router-PE-CE-link-address
 permit bgp host CE-interface-address host PE-interface-address
 deny ip any 23.49.0.0 0.0.255.255
 permit ip any any
```

The first line of the access list (the second line of Example 3-8) permits ICMP packets such as pings to be sent from the VPN customer to the directly connected PE router interface. Allowing such packets is useful for the customers, because they can perform diagnostics and management activity. The second line permits the BGP routing protocol to exchange routes by allowing communication between the CE router and PE router interface addresses. If a VPN customer is using a different protocol, this needs to be explicitly allowed within the filter.

The third line of the access list blocks any IP packets that are addressed to a destination within the USCom backbone network. The last line permits all other IP traffic to pass-through the PE router.

Quality of Service Design

The quality of service (QoS) marketed by a service provider and actually experienced by its customer base, is a key element of customer satisfaction. This is particularly true because most customers have already migrated, or soon will migrate, their mission-critical applications, as well as their voice and video applications, to IP services. In turn, this means that QoS is a key element of service provider competitiveness and success in the marketplace for both Internet and Layer 3 MPLS VPN services.

The levels of performance offered as part of Internet services in some parts of the world (such as the U.S.) have increased tremendously in recent years. This section discusses the Internet SLA that USCom offers in such a context. It also reviews the Layer 3 MPLS VPN SLA that USCom offers. Its objective is to allow customers to successfully carry all their mission-critical applications, as well as converge their data, voice, and video traffic. Finally, this section presents the design, both in the core and on the edge, deployed by USCom to meet these SLAs.

SLA for Internet Service

USCom offers an SLA for its Internet service. This SLA is made up of availability commitments, as well as performance commitments, as summarized in Table 3-6.

Table 3-6 *USCom Internet SLA Commitments*

SLA Parameter	SLA Commitment
Service availability (single-homed, no backup)	99.4%
Mean Time To Repair (MTTR)	4 hours
POP-to-POP Round-Trip Time (RTT)	70 ms
POP-to-POP Packet-Delivery Ratio (PDR)	99.5%

The availability commitments are provided to each Internet site. They are characterized by service availability of 99.4 percent (for a single-homed site attached via a leased line and without dial/ISDN backup) and an MTTR of 4 hours. Higher-availability commitments are offered with optional enhanced access options such as dial/ISDN backup and dual homing. Service availability is defined as the total number of minutes in a given month during which the Internet site can transmit and receive IP packets to and from the USCom backbone, divided by the total number of minutes in the month. USCom calculates service availability and MTTR based on trouble ticket information reported in the USCom trouble ticketing system for each site and customer.

Because USCom does not manage the Internet CE routers, the performance commitments of its Internet SLA are not end-to-end (not site-to-site). Instead, they apply POP-to-POP. The performance commitments are made up of an RTT of 70 ms and a PDR of 99.5 percent, which apply between any two POPs.

Using active measurements and averaging, USCom computes the POP-to-POP RTT and PDR. Dedicated devices located in every POP are used to generate two sets of sample traffic every 5 minutes to every other POP. The first sample traffic is a series of ten ICMP ping packets, which the sample traffic source uses to measure RTT. The second sample traffic is a series of ten UDP packets. The sample traffic destination uses them to measure the PDR (the ratio of received sample packets divided by the total number of transmitted sample packets). The worst RTT value measured over every hour is retained as the "worst hourly value." These "worst hourly values" are then averaged over the day, and the daily averages are averaged over the month. This yields the monthly average RTT value to be compared against the 70-ms SLA commitment specified in Table 3-6.

SLA for the Layer 3 MPLS VPN Service

Several considerations influenced the SLA definition of the Layer 3 MPLS VPN service. The first was the need to offer QoS levels that allow VPN customers to converge their data, voice, and video applications onto a common infrastructure. The second was the fact that it is

relatively easy and cheap for USCom to "throw bandwidth" at the QoS problem in the core (from POP to POP). As discussed in the next section, the most attractive approach for USCom to offer the appropriate QoS to all traffic of all application types, including the most demanding applications such as voice, was to offer indiscriminately the highest required QoS to all the traffic. In turn, this means that the SLA needed to specify only a single set of POP-to-POP performance commitments that were applicable to all traffic.

USCom handles all traffic the same way in the core and does not allocate more resources to some types of traffic over others. Therefore, there is no need for the company to charge differently depending on the mix of traffic types from the customer or to limit the rate at which some traffic types from a given customer site might enter the network. This results in a very simple service for both USCom and the customer in which the customer can transmit as much traffic as he wants, of any traffic type, without USCom's having to know or even care about it. In turn, this means that the service charge for a Layer 3 MPLS VPN site is a flat fee that depends only on the site's port access speed.

The next consideration was the fact that, unlike in the core, "throwing bandwidth" at the QoS problem on the access links (CE-to-PE and PE-to-CE links) is not easy or cheap for USCom. This is primarily because these links are dedicated to a single customer and need to be provisioned over access technology where capacity is usually still scarce or at a premium. This means that congestion is to be expected on some of the CE-to-PE and PE-to-CE links. Therefore, prioritization mechanisms (such as differentiated services) are required on these links to protect the high QoS of the most demanding and important applications.

The final main SLA consideration was the fact that USCom does not manage the CE routers that access the Layer 3 MPLS VPN service. This means that prioritizing traffic onto the CE-to-PE links, and the resulting QoS experienced by various applications on this segment, is entirely under the control of the customer (because such mechanisms have to be performed by the device that transmits onto the link) and conversely is entirely out of USCom's control. In turn, this means that USCom cannot offer any SLA performance commitment for the CE-to-PE segment. This does not mean that QoS cannot be achieved on that segment. Instead, it recognizes that such QoS is under the customer's operational domain and thus is the customer's responsibility.

Similarly, USCom does not restrict in any way the proportion of each traffic type that a CE router can send, nor does it restrict which remote VPN site this traffic is destined for. Therefore, USCom has no way of knowing, or controlling, how much traffic will converge onto a remote PE router for transmission to a given CE router on the corresponding PE-to-CE link. In addition, sizing of the corresponding PE-to-CE link is under the customer's control. Thus, although USCom manages the upstream device on the PE-to-CE link, it cannot offer any SLA performance commitment on the PE-to-CE link either.

To help illustrate this point, imagine that a given customer has a VPN containing five sites—S1, S2, S3, S4, and S5—each of which is connected to its respective PE router via a T1 link. Imagine that S1, S2, S3, and S4 all transmit 500 kbps worth of voice traffic destined for S5. The egress PE router that attaches to S5 receives 2 Mbps of voice traffic destined for S5. It is clear

that no matter what scheduling/prioritization mechanism may be used on the link to S5 by this PE router, and leaving aside any other traffic, trying to squeeze 2 Mbps of voice traffic onto a T1 link will result in poor QoS, at least on some voice calls (if not all).

However, the upstream end of the PE-to-CE link (the PE router itself, where prioritization and differentiation have to be implemented to offer appropriate QoS to the various applications on that segment) is not in the customer operational domain. So, unlike the CE-to-PE link case, the customer is not in a position to implement the required mechanisms independent of USCom to ensure that the right QoS is provided to all applications on the PE-to-CE link.

To ensure that the customer can achieve QoS on the PE-to-CE link in cases where that link may become congested, USCom decided to offer a service option (the "PE-to-CE QoS" option) whereby USCom would activate on the PE-to-CE link a fixed set of DiffServ per-hop behaviors (PHBs) that the customers can use as they see fit. Using this service offering, by properly sizing the PE-to-CE link, by controlling the amount of traffic for a given type of application (such as voice) sent from and to a given CE router, and by triggering the appropriate PHB made available by USCom for that type of traffic, the customer can ensure that the required QoS is provided to important applications. To trigger the appropriate PHB for a given packet, the customer simply needs to mark the packet's DS field in the IP header according to the Differentiated Services Codepoint (DSCP) values specified by USCom for each supported PHB. This marking must be performed at the ingress VPN site (on the ingress CE router) so that the DS field is already marked when the egress PE router applies the PHB during transmission of the packet onto the PE-to-CE link (after forwarding takes place and the MPLS header is popped). This is described in detail in the "QoS Design on the Network Edge" section.

The "PE-to-CE QoS" service option is marketed with a flat fee corresponding to the value-add provided to the end user and reflecting the extra processing load on the PE router. This option is kept very simple, without any customizable parameters.

For each application to experience the required QoS end-to-end, the corresponding requirements must be met on every segment of the traffic path—that is, within the ingress VPN site, from the ingress CE router to the ingress PE router, from USCom POP to POP, from egress PE router to egress CE router, and finally within the egress VPN site. The following summarizes how the USCom SLA performance commitments play out across each segment of the end-to-end path:

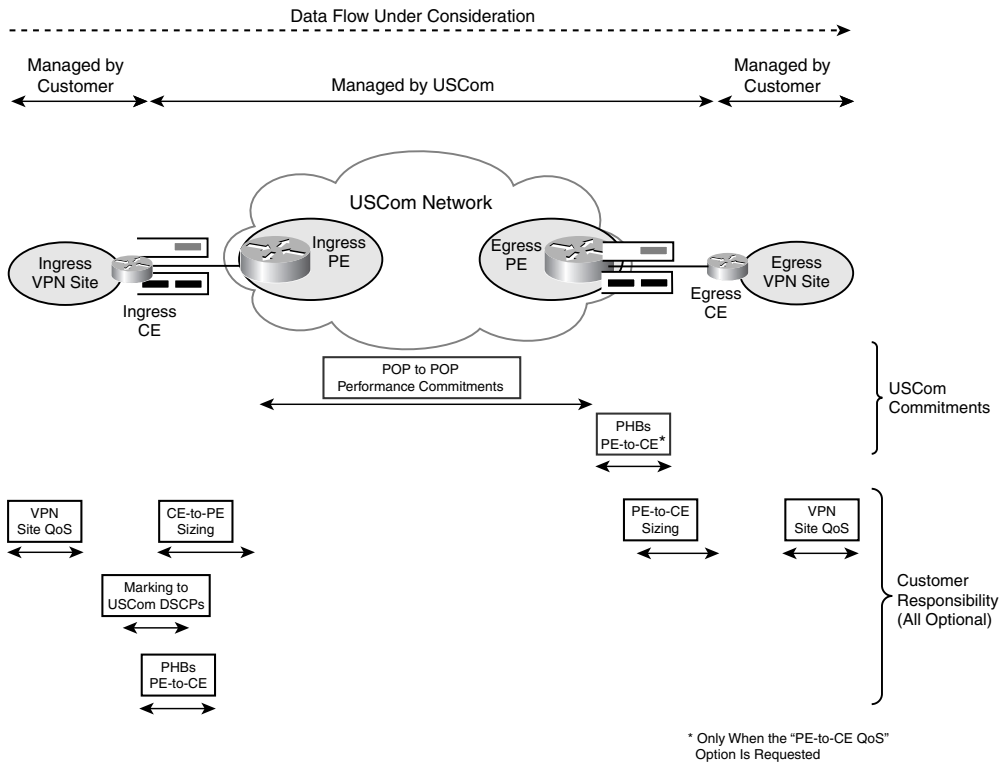
- **POP-to-POP**—USCom offers arbitrarily to all traffic (independent of its application type and/or actual requirements) SLA performance commitments across the backbone (from POP to POP). This is compatible with the most stringent requirements of any application (including voice).
- **CE-to-PE link**—USCom provides no SLA performance commitments on this segment. It is the customer's responsibility to ensure that the QoS required by the various applications is offered appropriately. The customer may achieve this by ensuring that the

CE-to-PE link is sufficiently overengineered for the total aggregate load or by deploying DiffServ mechanisms on the ingress CE router, including classifying traffic into separate DiffServ classes and applying the corresponding PHBs.

- **PE-to-CE link**—USCom provides no SLA performance commitments on this segment. However, as a service option, USCom can take the responsibility of applying DiffServ PHBs. It is then the customer's responsibility to use these PHBs, in combination with traffic control on ingress CE routers and capacity planning for the PE-to-CE link, to ensure that the QoS required by the different applications is provided on that segment.
- **Layer 3 VPN sites**—Because this is entirely out of the USCom realm of operation, USCom leaves it to the customer to ensure that the right QoS is offered inside the VPN. The customer may achieve this by overengineering the VPN site network (that is, via switched Gigabit Ethernet technology) and/or deploying DiffServ within the VPN Site.

These SLA points are illustrated in Figure 3-13.

Figure 3-13 *USCom VPN SLA Performance Commitments and Customer Responsibility*



As with the Internet SLA, all the performance commitments apply POP-to-POP. The RTT and PDR commitments provided in the Internet SLA are appropriate for any multimedia

application, so those are also used in the Layer 3 MPLS VPN SLA. However, because the performance commitments must meet the QoS requirements of all applications, including real-time/VoIP, a jitter commitment is added in the VPN SLA to the RTT and PDR commitments.

NOTE Because in the core no distinction is made between Internet traffic and VPN traffic, Internet traffic experiences the same performance level as VPN traffic. Nonetheless, USCom elected to include only the jitter commitment in the VPN SLA. The first reason for this is that USCom's service of choice for customers running real-time traffic is the Layer 3 MPLS VPN service. The second reason is that it is possible that, in the future, it will become more economical for USCom to deploy some DiffServ differentiation in the core and prioritize real-time traffic over other traffic so that very low POP-to-POP jitter may no longer be provided to all traffic by default.

As with the RTT and PDR, USCom uses active measurement and averaging to compute the jitter. The same series of sample traffic of ten UDP packets sent every 5 minutes that is used to measure the packet delivery ratio is also used to measure the jitter. Note that the sample source and sample destination do not need to synchronize their internal clocks because jitter can be computed by the destination only using its local timestamp on packet arrival and analyzing the variation over the known transmitted interpacket interval. The worst value measured over every hour is retained as the "worst hourly value." These "worst hourly values" are then averaged over the day, and the daily averages are averaged over the month.

Table 3-7 lists the Layer 3 MPLS VPN SLA commitments.

Table 3-7 *USCom Layer 3 MPLS VPN SLA Commitments*

SLA Parameter	SLA commitment
Service availability (single-homed, no backup)	99.4%
Mean Time To Repair (MTTR)	4 hours
POP-to-POP Round-Trip Time (RTT)	70 ms
POP-to-POP Packet-Delivery Ratio (PDR)	99.5%
POP-to-POP jitter	20 ms
Optional "PE-to-CE QoS"	Optional support of three PHBs on the PE-to-CE link

When unable to meet the commitments listed in Table 3-6 over the one-month measurement period, USCom offers refunds to its VPN customers in the form of service credits. The SLA specifies how the service credits are computed, depending on the observed deviation from the commitment for each SLA parameter.

QoS Design in the Core Network

This section presents the QoS design USCom deployed in the core network to support the Internet SLA and the Layer 3 MPLS VPN SLA performance commitments described in the previous sections. As discussed in the section "USCom's Network Environment," thanks to its DWDM optical infrastructure, and thanks to the use of Gigabit Ethernet switching within its POPs, USCom can enforce an overengineering policy. Therefore, it can maintain a low aggregate utilization everywhere in the core without incurring any excessive additional capital expenditure. USCom elected to take full advantage of this by

- Relying exclusively on aggregate capacity planning and overengineering to control QoS in the core and not deploying any DiffServ mechanisms or MPLS Traffic Engineering. This results in simpler engineering, configuration, and monitoring of the core.
- Pushing this overengineering policy approach further so that, in most cases, the aggregate utilization is kept low even during a single link, node, or SRLG failure. In turn, this ensures that QoS is maintained during most failures. (Protection against SRLG failure is discussed later, in the "Network Recovery Design" section.)
- Factoring in a safety margin when determining USCom's maximum utilization for capacity planning purposes to compensate for the shortcomings of capacity planning. This is discussed in the "Core QoS Engineering" section of Chapter 2, "Technology Primer: Quality of Service, Traffic Engineering, and Network Recovery."

Thus, USCom is adhering to the 1/1/0 model (or 3/1/0 model when the PE-to-CE QoS option is used) presented in the "QoS Models" section of Chapter 2.

The maximum distance between any two POPs in the USCom network is 4000 km. Assuming 25 percent of extra distance to cope with a longer actual physical route and additional distance when transiting via intermediate POPs, the one-way maximum distance is 5000 km and the round-trip maximum distance is 10,000 km. Assuming a 5-ms per 1000 km of light propagation delay through fiber, the maximum round-trip propagation delay in the USCom network is 50 ms.

NOTE USCom's optical network is quite dense so that the IP topology is generally congruent with the underlying optical topology. This is why only 25 percent of extra distance is factored in when computing the maximum one-way distance.

The SLA RTT commitment of 70 ms leaves 20 ms of round-trip queuing delay. Assuming a maximum of 12 hops in one direction, such a round-trip queuing delay is safely met if the delay at each hop is kept below 0.8 ms. In fact, the round-trip queuing delay is likely to be significantly better than 20 ms because delay commitment is statistical in nature and therefore does not accumulate linearly. However, USCom uses the simpler linear rule because exact accumulation formulas are not strictly known, and estimate functions are quite complex.

Similarly, the SLA jitter commitment of 20 ms can be safely met if the jitter is kept below 0.8 ms at every hop, which is all the more true if the queuing delay itself is bounded at 0.8 ms, as identified to meet the RTT commitment.

USCom determined through mathematical analysis and simulation of aggregate queuing through a single hop and by applying an empirical safety margin that the per-hop queuing delay requirement can be safely met with a maximum aggregate utilization of 70 percent for any of the link speeds used in its core. In other words, USCom characterized the shape of the QoS versus utilization curve (discussed in the section "The Fundamental QoS Versus Utilization Curve" in Chapter 2) for its particular environment and various core link speeds.

This analysis also indicated that the level of loss caused by excessive queue occupancy under such conditions would be well below what is necessary to achieve the SLA's packet delivery ratio (in fact, it would actually be negligible). However, the packet delivery ratio also accounts for other causes of loss, such as those due to failures and routing reconvergence.

Based on this, USCom specified its capacity planning policy whereby additional core capacity is provisioned whenever

- The measured link utilization exceeds 40 percent in the absence of any failure in the network.

  or

- The link utilization exceeds 70 percent in the case of a single failure of a link, node, or SRLG.

Clearly, this policy ensures that POP-to-POP performance commitments are met because the link utilization is significantly below the maximum aggregate utilization in the absence of failure and is below or equal to the maximum aggregate utilization in the case of a single failure.

To enforce this policy, USCom monitors link utilization at 10-minute intervals on every link. When the utilization reaches 40 percent, an alarm is triggered. If this level is reached in the absence of failure and is not caused by any exceptional event, additional capacity is provisioned.

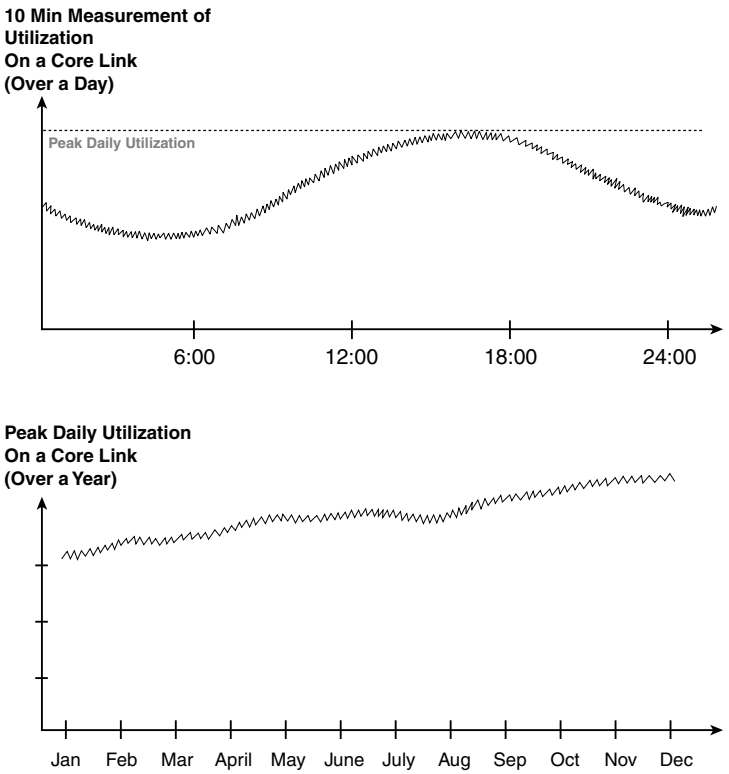
Also, USCom uses a network engineering and simulation tool with "what-if" analysis capabilities. On a regular basis, the measured maximum utilization figures for all links are fed into the tool. The tool then determines what the maximum utilization would be on all links should any link, node, or SRLG fail. If this exceeds 70 percent and cannot be reduced by adjusting IS-IS metrics (without redirecting excessive traffic onto another link), additional capacity is provisioned.

NOTE USCom has identified a small number of nodes and SRLGs in the current network whose single failure would result in a load on other links possibly exceeding 70 percent. This means the performance could possibly be somewhat degraded should any of those actually fail. But because such failure scenarios are very rare and affect performance only during the duration of the failure, it is very unlikely that they would prevent USCom from meeting its SLA commitments over its one-month period. Thus, USCom decided that these few exceptions to the capacity planning rules were tolerable.

It is therefore clear that as long as USCom can enforce its high overengineering policy (based on the capacity planning rule of keeping utilization on all links below 40 percent in the absence of failure and below 70 percent in the presence of failure), the SLA performance commitments can be met without deploying any additional QoS tools in the core network, such as MPLS DiffServ or MPLS Traffic Engineering.

Because the DWDM optical core is currently far from reaching capacity limitations (that is, all lambdas used on a given fiber), the link provisioning lead time is only a few weeks. Because traffic growth on the USCom backbone is relatively steady and free from huge spikes (as shown in Figure 3-14), USCom felt it will indeed be able to enforce its overengineering policy, at least in the next one to two years. Thus, USCom has not yet deployed MPLS DiffServ or MPLS Traffic Engineering. However, if in the longer term enforcing the high overengineering policy becomes difficult, USCom will then consider such technologies.

Figure 3-14 *USCom Utilization and Traffic Growth*



In summary, although USCom offers tight POP-to-POP SLA commitments for Internet and Layer 3 MPLS VPN traffic, its core QoS design is very simple. It relies entirely on capacity

planning, with enforcement of a high overengineering policy applied on an aggregate basis to all traffic. It does not involve any additional QoS mechanism in the core network.

QoS Design on the Network Edge

This section presents the QoS design deployed by USCom on the customer-facing interfaces of the PE routers. Because USCom does not implement any differentiated service in the core network and does not care about the mix of traffic classes received from a CE router, no QoS mechanism is configured on the ingress of the PE routers for both Internet customers and Layer 3 MPLS VPN customers.

On the egress side of the PE router, by default no QoS mechanisms are activated. However, if the Layer 3 MPLS VPN customer requests the PE-to-CE QoS option, a fixed QoS service policy is applied on the egress side of the PE router that activates three DiffServ PHBs.

Because USCom does not perform any QoS mechanism on the ingress side of the PE router or in the core, the Precedence field (or even the full Differentiated Services field) of an IP packet is carried transparently through the USCom network. Its value at the time of transmission by the egress PE router onto the PE-to-CE link (that is, after popping of the complete MPLS header) is unchanged from when the packet was transmitted by the customer ingress CE router. Therefore, USCom can use the Precedence field in the IP header as the classification criteria to apply the PHBs on the PE-to-CE link. To control which packets receive what PHB, the customer just has to mark the Precedence field on the ingress CE router (or upstream of it) in accordance with the Precedence-to-PHB mappings defined by USCom and specified in Table 3-8.

NOTE The ability to transparently transport the Differentiated Services field of the IP header over a Layer 3 MPLS VPN service provider network without modifying the value set by the customer is often called the QoS transparency feature.

Table 3-8 *Precedence-to-PHB Mapping for the PE-to-CE QoS Option*

Precedence Values	PHB*	Targeted Traffic
0, 1, 2, 3	BE	Best-effort traffic
4, 6, 7	AF41	High-priority traffic
5	EF	Real-time traffic

*See the section “The IETF DiffServ Model and Mechanisms” in Chapter 2.

For example, the customer could configure its CE router to

- Set the Precedence field to a value of 5 when sending a VoIP packet to the CE-to-PE link.
- Set the Precedence field to a value of 4 when sending an Enterprise Resource Planning (ERP) packet to the CE-to-PE link.
- Set the Precedence field to a value of 0 when sending other packets to the CE-to-PE link.

The result of this configuration is that when USCom transmits packets to the PE-to-CE link, it applies the EF PHB to the voice packets, the AF41 PHB to the ERP packets, and the BE PHB to the rest of the traffic. This effectively allows the customer to ensure that its applications are prioritized as it sees fit on the PE-to-CE link in case of congestion on that link.

Note that the customer would also probably configure its CE router to apply some custom PHBs on the CE-to-PE link to manage potential congestion on the CE-to-PE link. This set of custom PHBs does not have to be the same as the ones applied by USCom for the PE-to-CE QoS option, but it must be consistent with it, and its DS-field-to-PHB mapping must be consistent with the one from USCom. For example, the customer could decide to perform finer differentiation and activate a set of four PHBs with the Precedence-to-PHB mappings shown in Table 3-9.

Table 3-9 *Sample Precedence-to-PHB Mapping for Custom CE PHBs*

Precedence Values	PHB	Targeted Traffic
0, 1, 2	BE	Best-effort traffic
3	AF31	High-priority noninteractive traffic
4, 6, 7	AF41	High-priority interactive traffic
5	EF	Real-time traffic

End-to-end QoS operation when the PE-to-CE QoS option is not used, and when it is used by a customer, are illustrated in Figures 3-15 and 3-16, respectively.

NOTE Because Precedence values of 6 and 7 are set aside for network control (including IP routing) and network administration (including some network management traffic), some service providers apply preferential treatment inside their network to traffic marked with these Precedence values to guarantee the stability of their core. For example, as explained in the "Quality of Service Design" section in Chapter 4, Telecom Kingland schedules its internal control, management, and routing traffic inside a dedicated queue completely separate from the queue used for customer traffic. To make sure customer traffic does not interfere with their own network control or network administration traffic, such service providers either drop traffic sent by customers with a Precedence field set to 6 or 7 or remark it to a different value so that it does not map to the same MPLS EXP value as their own Precedence 6 and 7 traffic. However, because USCom does not enforce any traffic differentiation in its core, it decided to accept, as

is, traffic from customers with Precedence set to 6 or 7 and to schedule those as high-priority traffic in its PE-to-CE QoS option (as shown in Table 3-8). Should USCom deploy preferential treatment of its network control or network administration traffic inside its core in the future, it would have to ensure that customer traffic cannot interfere with this preferential treatment. (For example, USCom would have to remark the Precedence or impose on such customer traffic an EXP value different from the one it uses for its network administration and control traffic.)

Figure 3-15 *End-to-End QoS Operations Without the PE-to-CE QoS Option*

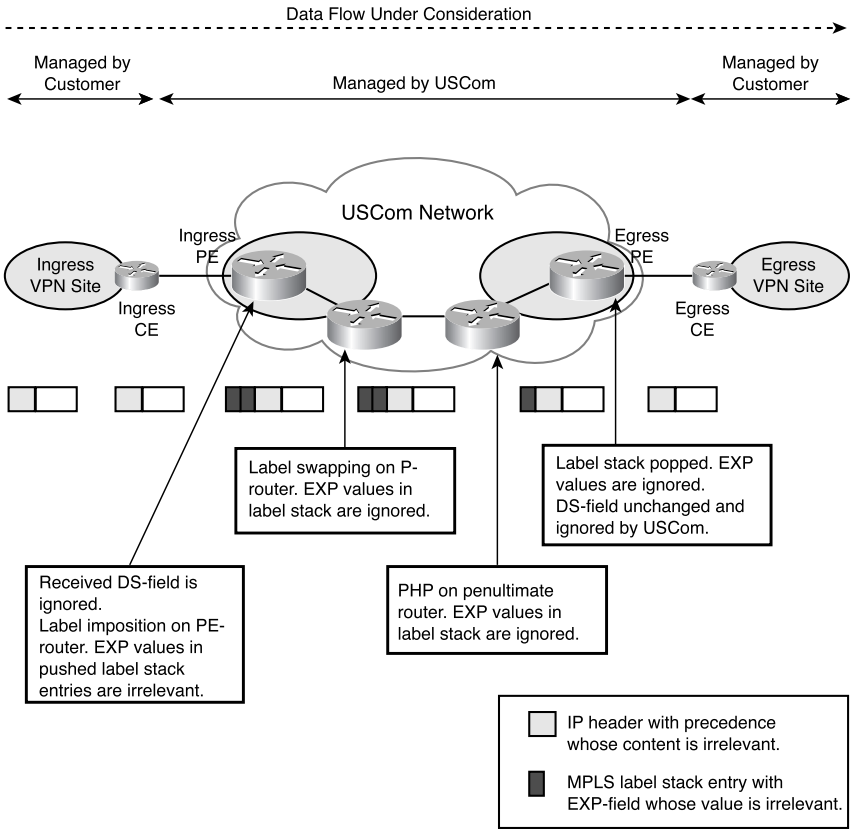
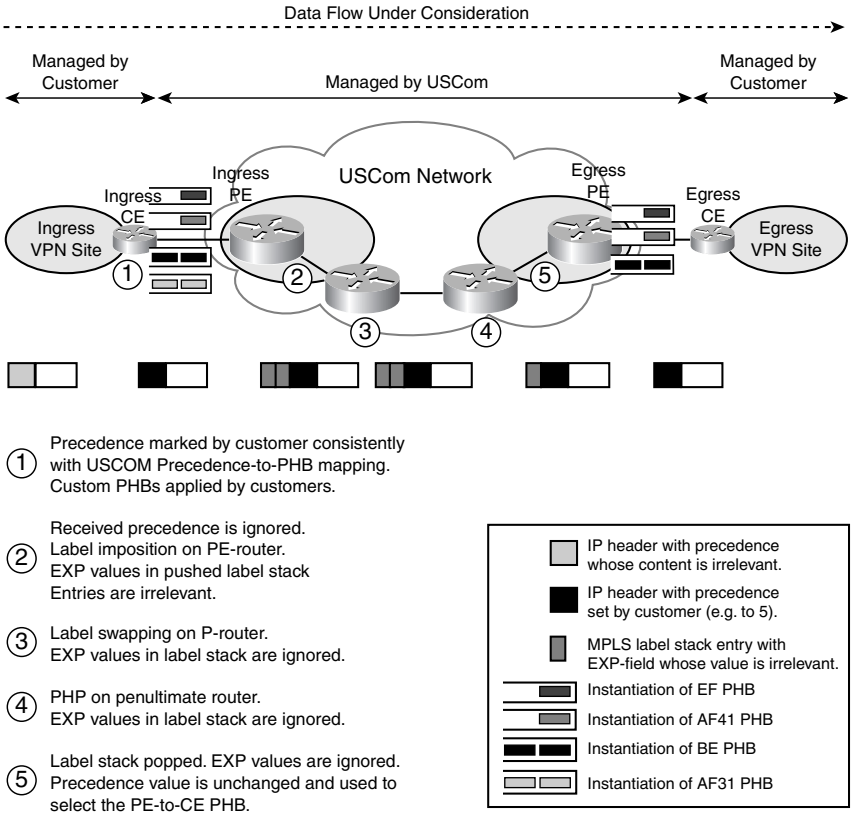


Figure 3-16 *End-to-End QoS Operations with the PE-to-CE QoS Option*



USCom elected to perform classification for the PE-to-CE QoS based on Precedence rather than the full DS field because it offers the end customer the flexibility to perform traffic marking either on the Precedence field or on the full DS field. For example, if the customer elected to mark VoIP packets with the DS field set to the EF DSCP (101110), these packets would be classified by the egress PE router appropriately because the first 3 bits of the packet's DS field, which constitute the Precedence field, are set to 101, which is Precedence 5.

Because USCom wanted to support a simple fixed set of PHBs for the PE-to-CE QoS option without any customizable parameters, it selected a versatile set of PHBs, as shown in Table 3-10, and a versatile PHB instantiation intended to be suitable for typical customer needs.

Table 3-10 *PHB Instantiation for the PE-to-CE QoS Option*

PHB	Instantiation
EF	Priority queue with 40% of the link bandwidth allocated. In the absence of congestion, bandwidth is not limited. In the presence of congestion, bandwidth is limited to 40% (excess is dropped) to protect the mission-critical applications expected to be handled by the AF41 PHB.
AF41	Class queue with most of the remaining bandwidth allocated (50% of the link bandwidth). This ensures strong prioritization of AF41 over BE. In case of contention across all classes, this queue is granted 50% of the link bandwidth. However, this queue is not limited to 50%. It can use more if the other queues are not currently using their allocated bandwidth. Random Early Detection (RED), as discussed in the section "The IETF DiffServ Model and Mechanisms" of Chapter 2, optimizes performance for TCP traffic, which is expected to be common in this class.
BE	Class queue with remaining bandwidth allocated (10% of the link bandwidth). In case of contention across all classes, this queue is granted 10% of the link bandwidth. However, this queue is not limited to 10%. It can use more if the other queues are not currently using their allocated bandwidth. Random Early Detection (RED) optimizes performance for TCP traffic, which is expected to be common in this class.

Example 3-9 illustrates how USCom configures PHB instantiation using Cisco IOS Modular QoS CLI (MQC) and applies it as the egress service policy of a PE router for a PE-to-CE link (see [QoS-CONF] and [QoS-REF] for details on how to configure QoS on Cisco devices using MQC). Note that the service policy is applied on the ATM1/0/0.100 and ATM1/0/0.101 interfaces because the PE-to-CE QoS option has been requested for the corresponding attached site. Expressing bandwidth as a percentage of the link bandwidth (rather than in absolute values) in the policy map is extremely convenient. It allows the use of a single policy map on all physical and logical interfaces regardless of their actual link speed.

Example 3-9 *Egress Service Policy for the PE-to-CE QoS Option*

```
ip vrf v101:USPO
 description VRF for US Post Office
 rd 32765:239
 route-target export 32765:101
 route-target import 32765:101
!
ip vrf v102:SoccerOnline
 description VRF for SoccerOnline International
 rd 32765:240
 route-target export 32765:102
 route-target export 32765:102
!
ip vrf v103:BigBank
 description VRF for BigBank of Massachusetts
```

continues

Example 3-9 *Egress Service Policy for the PE-to-CE QoS Option (Continued)*

```
 rd 32765:241
 route-target export 32765:103
 route-target import 32765:103
!
interface ATM1/0/0.100 point-to-point
 description ** BigBank_Site2 with PE-to-CE QoS option
 ip vrf forwarding v103:BigBank
 ip address 23.50.0.17 255.255.255.252
 pvc 10/50
  vbr-nrt 1200 1000 2
  encapsulation aal5snap
  service-policy out policy-PE-CE-QoS
!
interface ATM1/0/0.101 point-to-point
 description ** SoccerOnline_Site1 International with PE-to-CE QoS option
 ip vrf forwarding v102:SoccerOnline
 ip address 23.50.0.9 255.255.255.252
 pvc 10/60
  vbr-nrt 1500 1500 3
  encapsulation aal5snap
  service-policy out policy-PE-CE-QoS
!
interface ATM1/0/0.102 point-to-point
 description ** US Post Office_Site10 without PE-to-CE QoS option
 ip vrf forwarding v101:USPO
 ip address 23.50.0.13 255.255.255.252
 pvc 10/50
  vbr-nrt 1200 1000 2
  encapsulation aal5snap
!
class-map class-PrecHigh
  match precedence 4 6 7
class-map class-PrecVoice
  match precedence 5
!
policy-map policy-PE-CE-QoS
  class class-PrecVoice
   priority percent 40
  class class-PrecHigh
   bandwidth percent 50
   random-detect
  class class-default
   bandwidth percent 10
   random-detect
```

In summary, the USCom QoS edge design is very simple: By default, no QoS mechanism is activated on the PE routers. When a customer selects the PE-to-CE QoS option, a fixed service policy is applied in the egress direction onto the PE-to-CE link in order to instantiate a traditional set of three PHBs targeted at real-time, mission-critical, and best-effort applications.

Traffic Engineering Within the USCom Network

As established earlier in the "QoS Design in the Core Network" section, one of the fundamental network design rules adopted by USCom is the overprovisioning of available network resources, hence ensuring bounded link utilization in the core. This implies the following:

- Sufficient bandwidth must be provisioned in the core optical network.
- The metrics used by IGP (IS-IS in the case of USCom) must be computed so as to efficiently balance the traffic load in the core. In other words, traffic engineering is achieved through manipulating the IGP metrics. To ensure that link utilization remains below 40 percent in the absence of failures at all times, USCom decided to develop an internal tool that computes a set of IS-IS metrics that are used to traffic-engineer the network. The tool is run on a regular basis (approximately every 6 months) to accommodate some traffic growth. It is triggered by the monitoring of link utilization in the network by the USCom management system.

NOTE It is worth mentioning that changing the IGP metrics is not a completely cost-free operation. It requires some nonnegligible work for the network operations staff. Indeed, each IS-IS link metric must be changed individually to give the network time to converge before changing another IS-IS metric. Moreover, although transient states may lead to temporary congestion, as already stated in the section devoted to the network QoS design, this is unlikely to impact the overall SLA because it is averaged over a period of one month. Furthermore, such changes are performed during maintenance windows. Another constraint added to the IGP metric computation tool is to keep the link utilization below 70 percent in the case of a single link, node, or SRLG failure. In the case of the USCom network, this was an achievable objective in most cases.

Network Recovery Design

Network recovery is undoubtedly a key component of the overall network design because it impacts the network availability of the various service offerings and consequently the SLAs presented by USCom. In particular, USCom had an objective to offer high availability for both the Layer 3 MPLS VPN and Internet services. USCom's existing customers clearly required network availability for its VPN traffic equivalent to that given by the regular Layer 2-based network (Frame Relay, ATM, and so on). As far as the Internet traffic was concerned, although the requirements of this service are usually less stringent, USCom decided to arbitrarily provide high network availability to both the Layer 3 MPLS VPN and Internet traffic. As specified earlier in the SLA section, a network availability of 99.4 percent is guaranteed for both types of traffic.

Before determining its network recovery design, USCom had to consider several objectives and network design constraints, such as the required network availability, the failure scope coverage (link/SRLG/node failure), the requirement for covering single versus multiple failures, the

traffic rerouting time, QoS during failure, and single versus multiple class of recovery (CoR). Other criteria, such as the operational constraints and cost aspects, were also taken into account.

Network Availability Objectives

When considering the network failure scope, USCom had a requirement that the network be able to survive any single failure, including the failure of an SRLG (which is considered a single failure). In terms of rerouting time, the goal was to provide a 50-ms convergence time upon an inter-POP link or SRLG failure.

NOTE A link failure can be provoked by a fiber cut or optical equipment failure.

As mentioned in the "USCom's Network Environment" section, although a very limited set of data was available for the newly deployed optical network, USCom expected that link failures would be by far the most common failure scenario (90 percent of the failures were expected to be link failures). Consequently, the objective was to provide a rerouting time similar to SONET (60 ms) in case of link and SRLG failures only.

Because the USCom network was designed to engineer customer traffic flows based on the computed set of IS-IS metrics, link utilization does not exceed 40 percent during steady state and 70 percent during a single link/SRLG failure. Because of this, QoS can be guaranteed during failure (along the backup path) without the requirement of any type of DiffServ deployment in the core network. Hence, the only objective that USCom had in the design was to provide a backup path and to implement fast recovery (50 ms) upon link and SRLG failure. Based on this design, the rerouted traffic flows should not suffer from QoS degradation.

In terms of class of recovery, USCom decided to provide equivalent network availability to all traffic without any discrimination between types. In other words, both the Internet and Layer 3 MPLS VPN traffic should benefit from the same rerouting time objectives.

Operational Constraints on Network Recovery Design

One of USCom's key objectives was to carefully minimize the network management complexity for all service offerings. The adoption of a new technology, such as MPLS Fast Reroute, could not be justified if the cost of such an implementation unreasonably increased the network management complexity. Although such criteria might be somehow subjective, trying to keep the network as simple as possible was a clear objective, and it is reflected in the resulting network design.

Cost Constraints for the Network Recovery Design

Obviously USCom could have selected from a large set of network recovery mechanisms to be able to reach its particular network availability objectives. Every mechanism has some benefits and drawbacks in terms of efficiency, complexity, scalability, scope of recovery, and so on (as discussed in the Chapter 2 section "Core Network Availability"). For USCom, the cost of the chosen network recovery strategy to meet the set of objectives for network availability was of the utmost importance and had to be kept as low as possible. In particular, the purchase of additional equipment at any layer (optical or IP/MPLS) was to be avoided if at all possible.

Network Recovery Design for Link Failures

SONET link failures are handled at the SONET layer. In the case of the optical links, USCom decided to deploy unprotected light paths and MPLS-based Traffic Engineering Fast Reroute (FRR) to provide 50-ms rerouting time upon link and SRLG failure for every unprotected light path. An important objective was to keep the operation as simple as possible. Moreover, the only constraint to be taken into account as far as the backup tunnel path was concerned was the SRLG diversity. Indeed, as pointed out in the "Traffic Engineering Within the USCom Network" section, thanks to the in-house IGP metric computation tool, the network was designed such that any recovery path offers an acceptable QoS during failure.

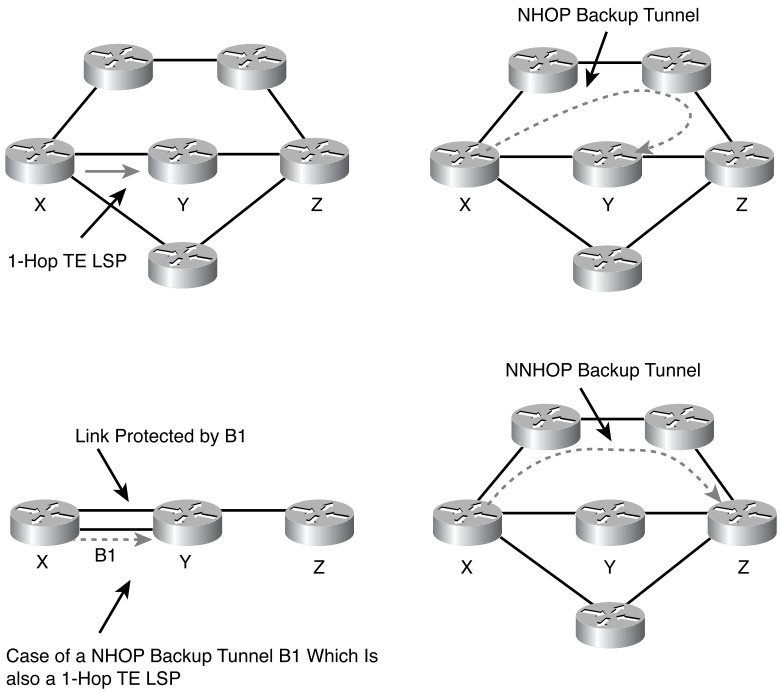
USCom elected to pursue the following MPLS Traffic Engineering Fast Reroute design for each light path that is to be protected:

- Configuration of a one-hop unconstrained primary TE Label-Switched Path (LSP)
- Dynamic configuration of an SRLG diverse next-hop (NHOP) backup tunnel

Before reviewing each of these aspects, it is useful to revisit the definitions of the terms one-hop, NHOP, and next-next hop (NNHOP).

As shown in Figure 3-17, a one-hop TE LSP is defined as a TE LSP that starts on router X and terminates on router Y, where Y is a direct neighbor of X. The signaling aspects of such a TE LSP are identical to any other TE LSP. The forwarding is different because it does not require any additional MPLS labels. Indeed, when a packet is sent to a one-hop TE LSP, no additional label is pushed (because of the penultimate hop popping operation).

Figure 3-17 *One-Hop, NHOP, and NNHOP TE LSP*



An NHOP backup tunnel simply refers to the fact that a backup tunnel originating on router X terminates on a direct neighbor of X (router Y in Figure 3-17). As shown in Figure 3-17, such a backup tunnel can be a one-hop tunnel if it protects a link via another parallel link or a multihop backup tunnel.

An NNHOP backup tunnel is a backup tunnel that originates on router X and terminates on router Z, where Z is one of X's neighbor's neighbors.

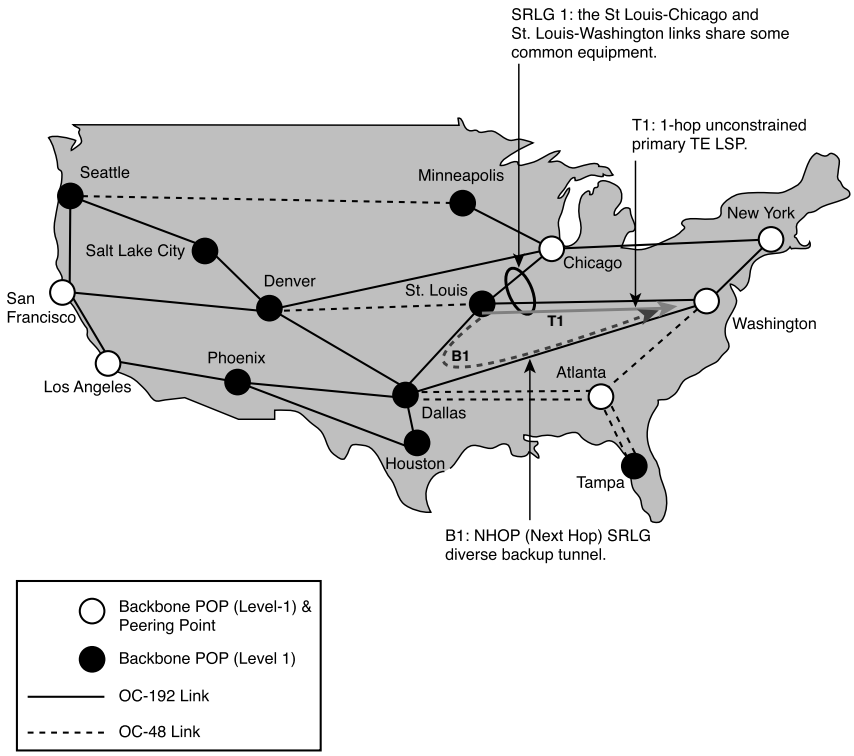
If you review each of these elements in more detail (along with the corresponding parameter tuning), you can see that configuration of a one-hop unconstrained primary TE LSP is possible because MPLS Traffic Engineering is just used with the aim of providing Fast Reroute protection. A single one-hop primary TE LSP is required so as to carry all the traffic routed through the link in question. (This is ensured because the TE LSP does not have any constraint, so its path just follows the IS-IS shortest path.) The use of such a primary one-hop TE LSP allows for the automatic protection of all the IP prefixes routed by the IGP along the same link that the one-hop tunnel follows.

Dynamic configuration of an SRLG diverse NHOP backup tunnel is made possible by flooding SRLG-related information within the IGP, as specified in [ISIS-GMPLS] and [OSPF-GMPLS]. In turn, this allows every router acting as a Point of Local Repair (PLR) to dynamically and

automatically compute an NHOP backup tunnel path, SRLG diverse from the protected link (a path that does not have any SRLG in common with the protected link). USCom decided to make use of such technology to reduce the management complexity.

Figure 3-18 shows an example of SRLG (the links St. Louis–Chicago and St. Louis–Washington share SRLG 1). It also illustrates an example of a one-hop unconstrained primary TE LSP and NHOP SRLG diverse backup tunnel for the St. Louis–Washington OC-192 link.

Figure 3-18 *USCom MPLS Traffic Engineering Fast Reroute Design*



In this example, the router in St. Louis has to compute an SRLG diverse path for the backup tunnel B1 that will be used to protect the link St. Louis–Washington.

NOTE The USCom team in charge of managing the optical layer provided all the information related to the design shown in Figure 3-18 (in particular, the SRLG membership).

The first step in deploying this design is to configure the SRLG membership on each router. This information is then flooded throughout the network by means of the relevant IS-IS extensions. Example 3-10 provides the necessary Cisco IOS configuration used by USCom for this process.

Example 3-10 *Configuration of SRLG Membership*

```
hostname USCom.StLouis.P1
!
interface POS0/0
 description ** St Louis - Washington OC-192 link
 mpls traffic-engineering srlg 1
!
interface POS1/0
 description ** St Louis - Chicago OC-192 link
 mpls traffic-engineering srlg 1
```

In addition, each router has been configured to automatically configure and set up a one-hop unconstrained primary TE LSP and an NHOP SRLG diverse backup tunnel for each protected link (the unprotected light path).

Example 3-11 illustrates how to automatically configure and set up unconstrained one-hop TE LSPs to each neighbor. These TE LSPs terminate at the IP address that is connected to each next-hop neighbor. They are fast-reroutable (protected by means of Fast Reroute). They do not have any other constraints such as bandwidth, affinities, and so on. This is because the aim of these TE LSPs is to use MPLS TE Fast Reroute as a fast local protection mechanism as opposed to using MPLS TE to effectively perform some traffic engineering functions.

Example 3-11 *Automatic Configuration of the One-Hop Primary Unconstrained TE LSP*

```
hostname USCom.StLouis.P1
!
mpls traffic-engineering auto-tunnel primary onehop
```

NOTE On a Cisco IOS router, additional commands allow the operator to tune other parameters.

NOTE As discussed in the Chapter 1 section "Forwarding of Layer 3 MPLS VPN Packets," the property of penultimate hop popping (PHP) consists of removing the TE LSP label at the penultimate router. Consequently, as pointed out, in the case of a one-hop tunnel, when PHP is in use, no label is added because the headend router is also the penultimate hop. Consequently, when one-hop tunnels are used for protection, as in the USCom network, no additional label is required when forwarding to a one-hop tunnel.

Example 3-12 shows the configuration that triggers the setup of one SRLG diverse backup tunnel per protected interface.

Example 3-12 *Automatic Configuration of NHOP SRLG Diverse Backup Tunnel*

```
hostname USCom.StLouis.P1
!
mpls traffic-engineering auto-tunnel backup nhop-only
mpls traffic-engineering auto-tunnel backup srlg exclude
```

Referring back to Figure 3-18, given the previous configuration, all traffic routed to the St. Louis–Washington link according to the IS-IS routing table is carried to the primary tunnel, T1.

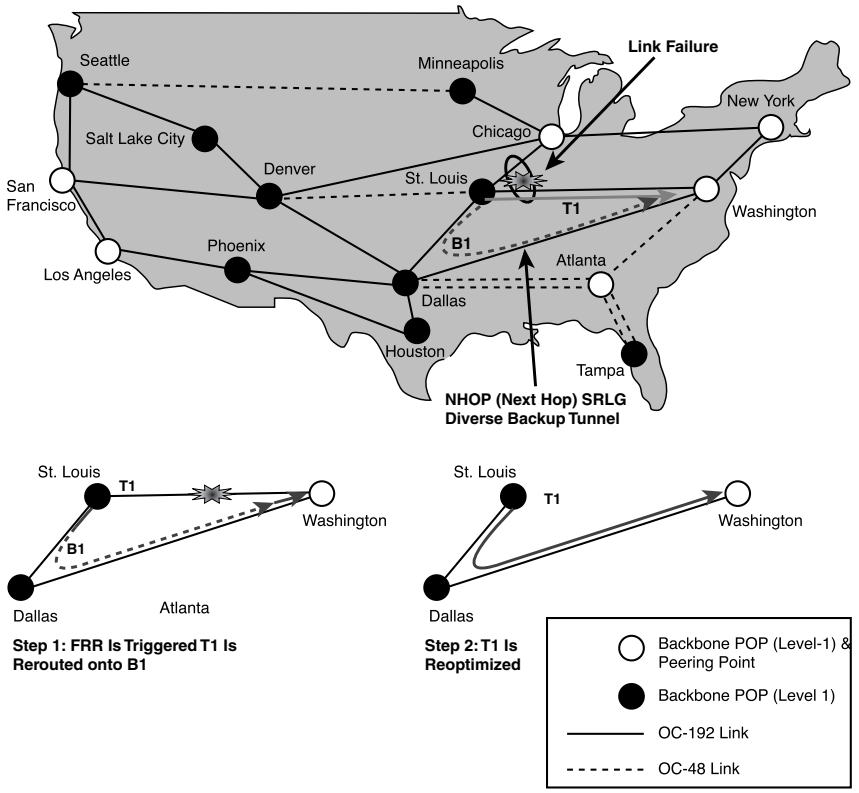
As shown in Figure 3-19, the router in St. Louis has an NHOP backup tunnel B1 configured to protect any fast reroutable TE LSP traversing the protected link St. Louis–Washington. Hence, upon a failure of the St. Louis–Washington link, T1 is rerouted to B1 within a few tens of milliseconds. Consequently, in the case of a failure of this link, all the traffic routed to the link is rerouted along the path followed by the backup tunnel B1. In a second step occurring right after the rerouting, the primary tunnel T1 is reoptimized by the PLR in St. Louis along a more optimal path. Because T1 is unconstrained, that path corresponds to the IS-IS shortest path along the new topology, as shown in Figure 3-19.

In the case of a link failure, the design selected by USCom guarantees a traffic restoration time within a few tens of milliseconds. This meets USCom's rerouting time requirements.

NOTE Thanks to the IGP network engineering rules, the IS-IS metrics can be computed by the USCom in-house IGP metric computation offline tool. The path followed by the backup tunnel and dynamically computed by each router in the network is guaranteed to offer an acceptable QoS to all traffic.

NOTE Because the TE LSPs are unidirectional, one one-hop unconstrained primary TE LSP and one NHOP SRLG diverse backup tunnel are required in each direction to fully protect a link with MPLS TE Fast Reroute.

Figure 3-19 *MPLS Traffic Engineering Mode of Operation in the Case of the St. Louis–Washington Link Failure*



Prefix Prioritization Within the USCom Network

When FRR is triggered, the fast reroutable traffic engineering LSP T1 is immediately rerouted to the selected backup tunnel. At a lower level of detail, this means that all the IP prefixes routed by means of T1 (shown in Figures 3-17 and 3-18) must have their forwarding entries updated to reflect the path change. Upon failure detection, MPLS TE Fast Reroute is triggered by the PLR. This operation consists of updating the forwarding entry for each affected IP prefix in a serialized fashion. Consequently, some prefixes are rerouted faster than others. (Note that the total rerouting time for all prefixes still occurs within a very short period.) USCom adopted an interesting design solution that consists of giving a higher priority to important prefixes so that they get rerouted before less-important prefixes.

NOTE Note that this is just an optimization, but it may be useful in large networks such as USCom, which has more than 3,000 IS-IS prefixes.

Given the Layer 3 MPLS VPN service offered by USCom, and the desire to maintain the same level of service for its Internet customers, USCom chose to use prefix prioritization during the FRR process. To ensure that IP and VPNv4 traffic is restored first in the case of a link failure, the IP addresses that represent a BGP next hop (a loopback from either an Internet or Layer 3 MPLS VPN PE router) were chosen for prioritization. This optimizes the reroute of these services, because these addresses are used by recursive resolution to reach all IP and VPNv4 prefixes advertised by the USCom PE routers. IP addresses of internal links, such as those between P routers, were considered less important, or at least did not require such a stringent convergence time.

NOTE It is worth observing that the Internet traffic is IP-routed in steady state (because of LDP label filtering). During failure the traffic is label-switched along the backup tunnel and then the primary multihop TE LSP.

This prioritization is achieved using the configuration shown in Example 3-13.

Example 3-13 *Configuration of Prefix Prioritization*

```
hostname USCom.StLouis.P1
!
mpls traffic-engineering fast-reroute acl prefix-priority
!
ip access-list standard prefix-priority
 permit 23.49.16.0 0.0.1.255
 permit 23.49.20.0 0.0.1.255
 permit 23.49.10.0 0.0.1.255
```

NOTE In Example 3-13, the subnets 23.49.16/24, 23.49.20/24, and 23.49.10/24 represent the main and reserved Layer 3 MPLS VPN PE router loopbacks and the Internet PE router loopback addresses, respectively.

NOTE The configuration of such prefix prioritization triggers an FRR database sorting function that ensures that the important prefixes are rerouted first.

Temporary Loop Avoidance

The MPLS TE Fast Reroute design elected by USCom allows the company to meet its 50-ms rerouting time objective in case of link failure, but it requires a bit of extra work to be entirely

satisfactory. By their very nature, IGP link-state protocols may lead to temporary loops during network convergence. (Until all the routers have synchronized their Link-State Database [LSDB] and have converged, a loop-free state cannot be guaranteed.) By default, the knowledge of a TE LSP is kept local to the router that is the headend for that TE LSP. The FRR design chosen by USCom does not escape this rule (the TE LSP [T1] is not visible to any other router in the USCom network). The consequence of this upon a link/SRLG failure is that the router in St. Louis locally reroutes all the traffic traversing the St. Louis–Washington link. After a period of time (determined by the IS-IS timer tuning, discussed later in this chapter), both the routers in St. Louis and Washington originate a new IS-IS LSP that reflects the new network topology and, in particular, the loss of adjacency between those two routers. The IS-IS LSP is then flooded throughout the network, and each router triggers a new routing table calculation.

IP routing is distributed, so the sequence of events is not deterministic. In particular, although the Dijkstra algorithm guarantees the computation of a loop-free path during steady state, this may not be the case during network convergence, when the routers' LSDB may not be synchronized.

To help illustrate this point, consider the following sequence of events:

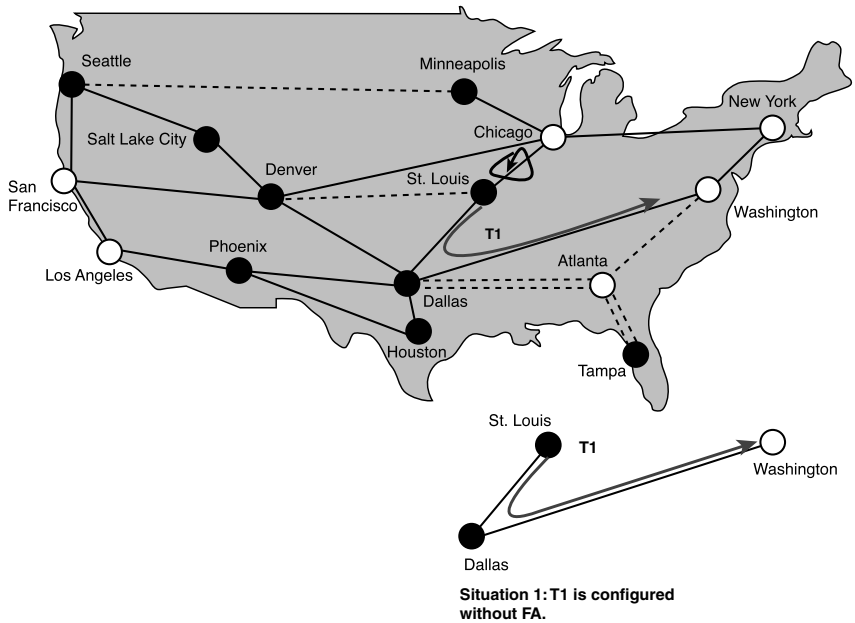
- **Time t0**—The St. Louis–Washington link fails. (More precisely, the interface connecting the link to Washington fails on the router in St. Louis.)
- **Time t1**—The router in St. Louis detects the failure and triggers a local reroute by means of MPLS Fast Reroute. IS-IS originates a new IS-IS LSP and recalculates its routing table and forwarding database (note that MPLS Fast Reroute and IS-IS operate independently).
- **Time t2**—The router in Chicago receives the newly originated IS-IS LSP and recalculates its routing table and forwarding database.

During the time interval (t2–t1), the LSDBs of the routers in St. Louis and Chicago are not synchronized with each other. Assume that the IS-IS link metrics have been computed by the in-house tool such that

- The shortest path from Chicago to Washington is Chicago–St. Louis–Washington (in the absence of failure).
- In case of failure of the St. Louis–Washington link, the shortest path from St. Louis to Washington is via Chicago and New York.

During t2–t1, a temporary loop appears between the St. Louis and Chicago routers for the traffic sent from Chicago to Washington. This happens because during t2–t1, the St. Louis router sends the traffic for Washington back to the Chicago router, as shown in Figure 3-20.

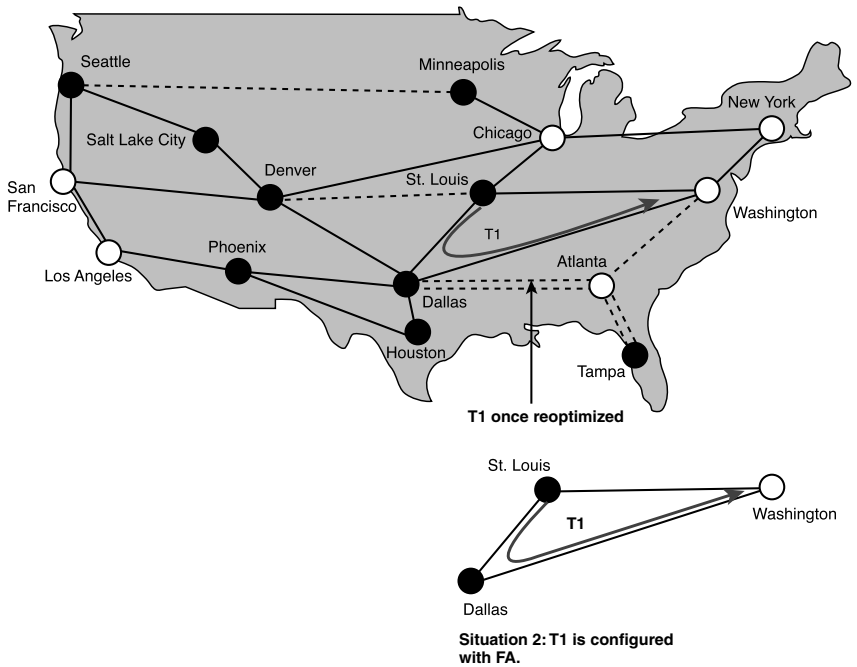
Figure 3-20 *Temporary Loop Effect During IS-IS Network Convergence*



Forwarding Adjacency for Loop Avoidance

The solution to this problem is to configure the primary tunnel T1 as a forwarding adjacency (FA). Configuring a primary TE LSP as an FA has the effect of flooding the TE LSP as an IP link within the IGP. This means that as long as the TE LSP is operational, the node in St. Louis advertises T1 as a physical link in its IS-IS LSP. Consequently, when the physical link St. Louis–Washington fails, T1 is rerouted and then reoptimized along the path St. Louis–Dallas–Washington. T1 is still advertised in the IS-IS LSP originated by the St. Louis node as a physical link. Hence, upon link failure, no new IS-IS LSP is advertised, and the other routers in the USCom network do not detect any network topology change. (Of course, this requires configuring the FA with the same cost as the primary link it traverses.) This avoids the undesirable temporary loop effect just described. This process is illustrated in Figure 3-21.

Figure 3-21 *Avoiding the Temporary Loop with Forwarding Adjacency*



The advantage of using a forwarding adjacency is that any temporary loop during network convergence can be avoided. Moreover, in the case of a temporary failure of a few seconds, FA prevents the generation of two network convergence sequences throughout the network (which impacts hundreds of routers in the case of USCom).

As with all design choices, there are trade-offs. In the case of a forwarding adjacency, some failures such as fiber cuts may last for several days or even weeks, although USCom has not yet gathered a significant optical failure history. In such a case, the path followed by the traffic routed across the St. Louis–Washington link may follow a nonoptimal path for a long period of time because IS-IS is unaware of any topology changes. For instance, the traffic following the path Denver–St. Louis–Washington would actually follow the path Denver–St. Louis–Dallas–Washington, because the router in Denver does not actually see the failure of the link St. Louis–Washington. However, the path Denver–Dallas–Washington might have been more optimal. This is because T1 is still advertised as a physical link; hence, the other routers do not see any network topology change. In the USCom case, this was not considered an issue, because the network is overprovisioned. Therefore, the backup path, although potentially not optimal for some period of time, still provides the required QoS guarantees. Moreover, USCom has a monitoring system capturing the SNMP traps. Therefore, a procedure can be put into place to detect link failures and potentially deconfigure forwarding adjacencies on some primary tunnels if the failure lasts too long, such as in the case of a fiber cut. Of course, the

deconfiguration of forwarding adjacency triggers an IGP convergence and is equivalent to the previous case where temporary loops occur.

Reuse of a Restored Link

An important consideration in the USCom FRR design was the reuse of a restored link after failure. Once a link is restored, a couple of strategies can be put into place:

- Reuse the restored resources as soon as possible.
- Wait for a period of time before reusing the link to maximize the network stability.

A multitude of link failure profiles are possible. For example, a link can fail and then be reliably restored (an up-down effect caused by a temporary network element failure, such as a laser desynchronization). On the other hand, a link can become unstable, experiencing a set of successive failures (in other words, the link is flapping). In such cases, waiting for a period of time before reusing a link helps determine whether it is safe to reuse the link. A flapping effect is highly undesirable; it can generate network instabilities, triggering storms of LSP flooding, SPF computation on each router, and so on. To solve such issues, multiple techniques can be used, such as back off and dampening. These can be implemented at various layers (such as the interface level, IGP, MPLS Traffic Engineering, and so on). In a nutshell, the key idea is to dampen the use of a link that suffers from instabilities to preserve network stability.

Various back-off/dampening algorithms have been designed (based on accumulated penalties such as BGP dampening, exponential back off, and so on). In the case of MPLS Traffic Engineering, on a Cisco router, the triggering of the TE LSP reoptimization always drives the reuse of a link. When a router tries to reoptimize a TE LSP path by means of a CSPF algorithm, it first determines whether a more optimal path other than the path currently in use can be found. If it can, the TE LSP is gracefully rerouted along the more optimal path. A Cisco router has various configurable reoptimization triggers that can be individually activated and deactivated:

- **Timer-based trigger**—Every (Tr) seconds, a headend router attempts to reoptimize its set of TE LSPs.
- **User-triggered**—Reoptimization forced by the user.
- **Link-up**—Each time the IGP signals a link, every router tries to see whether its set of TE LSPs can benefit from that new link.

USCom therefore had to make a decision regarding the following trade-offs:

> Reuse a restored link as soon as possible to quickly alleviate some congestion, but with the potential risk of generating network instability
> or
> Wait for a period of time before reusing a restored link whose state does not change

Because the USCom backbone is overprovisioned, the immediate reuse of a restored link was not considered a priority when compared to preserving network stability. Hence, USCom decided to be conservative in the reuse of a restored link. It would rely on IS-IS to declare the

link operational according to the IS-IS back-off mechanism, described in the Chapter 2 section "Use of Dynamic Timers for LSA Origination and SPF Triggering." No TE LSP reoptimization is triggered on a link-up event. On the traffic engineering side, USCom decided to use the timer-based approach, with a timer value of 15 minutes. Hence, every 15 minutes, a router tries to see whether the link is restored so as to reoptimize the one-hop unconstrained primary TE LSP along the link. As soon as a link is restored, in the worst case, the TE LSP is rerouted to it in 15 minutes.

NOTE Note that even if the IS-IS adjacency is reestablished across the restored link, the traffic routed between the two routers according to the computed routing table is steered to the TE LSP. This means that the traffic traverses the restored link only when the TE LSP is reoptimized along that path.

Multiple Failures Within the USCom Network

When assessing its multiple-failure requirements, USCom found that the only cause of concern was multiple failures provoked by the failure of an SRLG. Such a situation is handled by the design because every backup tunnel path is dynamically computed to be "SRLG diverse" from the protected link. (The links visited along the backup tunnel path do not share any SRLG with the link protected with FRR.) Any other case of multiple failures (such as an SRLG failure followed by the failure of another link that does not belong to the failed SRLG) was not considered a requirement because of the low probability of multiple simultaneous failures of independent elements. It is worth pointing out that the USCom network could survive double failures without experiencing a disconnected network but would not have any guarantee in terms of rerouting times and QoS during those multiple failures.

Link Failure Detection Within the USCom Network

The main challenge when protecting links in a switched environment (such as the intra-POP Gigabit Ethernet links) is quickly detecting a link failure. In the case where two routers are interconnected via a direct Gigabit Ethernet link, in only a few milliseconds the neighbor can detect a link failure caused by a fiber cut or a router interface failure. On the other hand, if the routers are interconnected by means of an intermediate Layer 2 switch, as in the case of the USCom Level 3 switched POP, this presents the challenge of link failure detection, because it requires the use of a fast hello (keepalive) protocol. Indeed, consider the following two cases:

- Two routers connected by a direct PoS or Gigabit Ethernet link. The failure of the link or one of the router interfaces is quickly detected by means of the alarms provided by Layer 1 or 2.

- Two routers connected by means of a Layer 2 switch. In this case, the failure of the link or router interface is seen only by the switch and the router connected to the failed element. Hence, the routing neighbor of the router attached to the failed element cannot detect such failures other than with a hello protocol between the two routers.

Because router interface failures are pretty rare and intra-POP links also very rarely fail, USCom decided not to protect these intra-POP links and to just rely on the IS-IS convergence.

Node Failures Within the USCom Network

When assessing the requirements for protection against node failure within the network, USCom chose to differentiate between the case of planned router maintenance and unplanned router failure.

Planned Router Maintenance

A software or hardware upgrade may require a router to be taken out of operation for a period of time (typically 10 minutes on average in USCom's case, as indicated in Table 3-2). In this case, for the core routers, USCom adopted the approach of setting up the IS-IS overload bit of the router in question via an administrative procedure. This has the effect of triggering a network-wide IS-IS convergence (rerouting) of the traffic around the router in question. As soon as the network has fully converged, the upgrade can finally take place without any traffic disruption. Such an approach is particularly suited to the USCom environment. The network overengineering rules are such that the network does not experience any congestion, even in such circumstances as a single network element failure. After the router has been reloaded, the original link metrics are restored.

In the case of the planned router maintenance of edge routers (Layer 3 MPLS VPN and Internet PE routers), things are quite different. USCom considered three scenarios:

- **Internet customer sites that are dual-attached**—Before upgrading the Internet PE routers, USCom relies on a script to automatically increase the MED value for all the BGP routes announced to the set of affected CE routers. This allows each CE router to smoothly reroute its traffic to the second PE router it is connected to; this avoids any traffic disruption. The actual PE router maintenance takes place 5 minutes after the BGP routing changes.
- **VPN customer sites that are dual-attached**—This could be two colocated CE routers connected to two different Layer 3 MPLS VPN PE routers or a single CE router connected to two different Layer 3 MPLS VPN PE routers. A similar procedure is applied to the case of dual-attached CE routers if BGP is used as the routing protocol between CE router and PE router. No particular measure is taken for other routing protocols or the CE routers using static routing.

- **Internet and VPN customer attached to a single PE router**—In this case, USCom handles the router maintenance, which inevitably provokes some traffic disruption, during a maintenance period.

In all these cases, USCom managed to get a maintenance period window of 4 hours on the first Sunday of every month from 3 a.m. to 7 a.m.

Unexpected Router Failures

USCom noted that several types of router failures have highly variable effects on data forwarding. These effects can vary from the traffic being black-holed to absolutely no consequences on the traffic, depending on the router platform, failure types, and so on.

Two examples are provided in the next section to illustrate how the USCom IS-IS design met the requirements of unexpected router failures:

- The case of a power supply failure at a core router
- The case of a router failure that does not trigger any link failure, or the failure cannot be detected by the neighbors

Convergence of IS-IS

When designing the tuning of IS-IS from a convergence perspective, USCom had the objective of providing a convergence of 5 seconds in the case of a router failure or intra-POP link failure (when Layer 2 switches are used to interconnect routers). Link failures were considered outside the scope of IGP tuning because MPLS Traffic Engineering Fast Reroute covers them. This convergence time includes detection of the failure, propagation of the topology change, and local convergence (computing a new routing table).

Of course, a number of IS-IS parameters come into play when tuning IS-IS for faster convergence. As mentioned in Chapter 2's "Core Network Availability" section, the IGP convergence time is basically made up of three main components:

- The failure detection time
- The flooding of the new IS-IS LSP reporting a topology change
- The routing table computation on each router (SPF algorithm, Routing Information Base (RIB) update, and so on)

IS-IS Failure Detection Time

When a router failure occurs, also implying multiple link failures (such as a power supply failure), the SONET or light path link failure is detected within tens of milliseconds. On the other hand, as mentioned, the case of a link or a router failure within a switched POP (Level 3) requires a hello protocol. USCom decided to set the IS-IS hello frequency to 1 second (one IS-

IS hello message [IIH]Drew, I changed back to parenthesis since this does not refer to a reference but to the name of the hello messages. Thanks. is sent to every adjacent neighbor every second). Note that in the USCom network topology the maximum number of adjacent neighbors stays within a very reasonable limit (less than 30). Hence, sending an IS-IS Hello message every second to each neighbor is not of concern. The Hold timer is set to 3 seconds. If no IS-IS message is received during this period, the routing adjacency is declared down.

Flooding of New IS-IS LSPs

The flooding of new IS-IS LSPs is basically a function of the LSP origination time (discussed later in this section), the propagation delay, and the processing time at each router hop. In the USCom network, because the optical network is pretty dense, the worst-case propagation delay from coast to coast is 50 ms. Based on several internal tests, USCom determined that the worst-case processing delay of an IS-IS LSP even on a pretty heavily loaded router would rarely exceed 10 ms. This calculation supposes that the flooding of the newly received IS-IS LSP always occurs before the triggering of the new SPF.

NOTE Note that the ability to systematically flood an LSP before triggering an SPF may not exist on some router platforms but is quite important to limit the convergence time of any link-state routing protocol. Indeed, upon receiving a new LSP (an LSP reflecting a topology change), a router should always first flood the LSP instead of triggering an SPF and then flooding the LSP.

Furthermore, every router has to be configured to ensure that the queuing delay experienced by the IS-IS control messages is bounded and negligible so it won't severely impact the total convergence time. It is of the utmost importance to provide a high priority to the IS-IS control messages. This applies to hello messages to avoid losing a routing adjacency in case of congested links (not in the case of USCom, however). It also ensures a quick LSP update because hello messages may reflect a topology change (if the LSP is not a refresh), which is required to quickly converge to reroute the traffic to alternate paths. Because IS-IS control messages do not rely on IP, internal mechanisms need to ensure that IS-IS messages get the relevant precedence over other user traffic.

NOTE The serialization delay on a link from OC-3 to OC-192 of an IS-IS LSP is not a significant factor in the overall IS-IS convergence.

Based on the previous flooding time analysis, USCom determined that the total flooding time should never exceed 200 ms. (This is the time to originate the new IS-IS LSP plus the total

propagation delay between the originating routers and the routers where the traffic is rerouted along an alternate path by IS-IS.)

Routing Table Computation on Each Node

The final component of the IS-IS convergence to consider is the routing table computation time, which is itself made up of two components:

- The SPF computation
- The routing table computation and update to router line cards (in the case of a distributed router architecture)

Some testing on the USCom network showed that the SPF computation time was 100 ms, and the complete routing table update was 500 ms.

IS-IS Configuration Within the USCom Network

Example 3-14 provides the configuration of the St. Louis P router shown in Figure 3-1 to achieve the IS-IS convergence objective of 5 seconds upon a node failure. Similar configurations are adopted for all the routers in the network.

Example 3-14 *Fast IS-IS Configuration*

```
hostname USCom.StLouis.P1
!
interface pos0/0
 isis hello-interval 1
 isis hello-multiplier 3
!
router isis
 lsp-gen 5 50 20
 spf-interval 5 50 20
 prc-interval 5 50 20
```

NOTE The hello interval and hello multiplier must be configured on each interface that is associated with the IS-IS process.

Considering the syntax lsp-gen A B C, USCom decided to set B to 50 ms so that every router would get a chance to detect all the possible local failures (caused by SRLG failure) before originating a new IS-IS LSP. Indeed, upon SRLG failure, multiple local links may fail, and these failures might not be detected simultaneously. Thus, the 50 ms of waiting time before originating the new IS-IS LSP provides an accurate network topology state. If a second failure occurs, the router originates a second LSP after 20 ms (C = 20).

This also applies to the triggering of the SPF. In the syntax spf-interval A B C, B is set to 50 ms. This gives a chance, in case of an SRLG failure, to receive all the IS-IS LSPs reflecting the new topology and consequently the SRLG failure before triggering a new (second) SPF. See the section "Core Network Availability" in Chapter 2 for an explanation of the various IS-IS tuning parameters.

To help illustrate the outcome of the IS-IS parameter settings, consider two extremes:

- **The case of a power supply failure at a core router**—In this case the links attached to the router will also likely fail, which will provide a fast failure indication to the neighbors of the failing routers. Each neighbor originates a new IS-IS LSP that is flooded throughout the network, and each router converges. In such a case, the failure is detected before the expiration of the hold-time timer. The propagation delays and SPF/RIB computation time are such that the objective of 5 seconds total convergence time is easily met.
- **The case of a router failure that does not trigger any link failure, or the failure cannot be detected by the neighbors**—For the sake of illustration, two situations should be considered:
  - A router fails, with impact on traffic forwarding, but the attached links do not fail.
  - A router fails, with impact on traffic forwarding. Its attached links also fail, but its neighbors cannot detect these failures. Typically this is the case with a switched POP.

In these two situations, the failure detection occurs by means of the IS-IS adjacency maintenance procedure—hence, within 3 seconds (until the hold-time timer expires). This still provides 2 seconds for the neighbors of the failing router to originate their new IS-IS LSP, for the new LSP(s) to be flooded throughout the network, and finally for all the nodes to converge. Hence, this guarantees that the 5-second rerouting time objective is also met with the previously mentioned IS-IS parameter tuning. Note that only a subset of the routers is required to converge for the impacted traffic (traffic routed through the failing router) to be restored.

It is worth mentioning that other router failures do not affect data forwarding, such as a control plane failure on a distributed platform. In such failures, if the control plane cannot be restored within 3 seconds (the value of the hold-time timer), the IS-IS neighbor declares a loss of adjacency, and IS-IS converges (the traffic is rerouted around the failing router). However, the user traffic is unaffected because the alternate paths offer an equivalent QoS in the case of USCom.

It is worth noting that an edge router failure always has an impact on the traffic originated by locally attached CE routers as well as the traffic to those sites. USCom decided not to initially implement any high-availability (HA) functionality on the Internet or Layer 3 MPLS VPN PE routers, but this will be assessed at a later stage. Hence, this applies to any type of router failure. Because the customer sites are out of the realm of the USCom operation (they are unmanaged), the customers, depending on the routing protocol in use and their parameter settings, control the convergence time.

Design Lessons to Be Taken from USCom

A number of observations can be made from USCom's design decisions:

- Straightforward engineering rules such as structured VRF naming conventions, route distinguisher/route target allocation schemes, and well-defined configuration templates allow for a simpler Layer 3 MPLS VPN service deployment.
- Operation of the Internet service can be kept exactly as before deployment of the Layer 3 MPLS VPN service by separating forwarding of Internet and VPN traffic in the core. VPN traffic is carried over MPLS LSPs, while Internet traffic remains forwarded as IP traffic.
- PE router protection techniques, such as limiting the number of routes within a VRF or restricting the number of prefixes received from a given client, should be a mandatory part of the Layer 3 MPLS VPN service deployment.
- Simple tuning of certain router parameters, such as the input hold-queue and Selective Packet Discard (SPD), can considerably enhance convergence of the BGP control plane.
- Route reflectors should be deployed to help scale the number of BGP TCP sessions required at the PE routers.
- Enabling path MTU discovery at the PE routers and route reflectors allows the TCP protocol used by BGP to run more efficiently, thus providing better convergence times.
- Where core bandwidth is plentiful/cheap/quick to provision, the core QoS design can rely on pure overengineering to maintain QoS during single failures and to achieve a good SLA that satisfies mission-critical and multimedia applications. This is a low operational expenses (opex) design because of simpler engineering, configuration, monitoring, troubleshooting, and fine-tuning. This is usually an attractive avenue for "facilities-owned" operators with an optical infrastructure.
- Even when no QoS mechanism is supported in the core, and unmanaged CE routers are deployed, it is a good idea to offer an optional QoS mechanism on egress PE routers. Doing so provides added value for customers because it manages congestion on the last weak link in the chain (the first weak link, the CE-PE link, can be managed by the customer anyway) and does not add significant complexity to the design.
- A network can follow a simple design to be able to offer a 50-ms convergence time upon link or SRLG failure by means of MPLS Traffic Engineering Fast Reroute, at a minimal cost in terms of opex and capital expenditure (capex). Such backup tunnels can be automatically configured and set up with minimal configuration.
- Node failures may be covered by minimal IGP tuning to obtain a few seconds of rerouting time upon a router failure that affects data forwarding. USCom might consider more-aggressive IS-IS parameter settings if it has to increase its network availability in the future.