Implementing Advanced Features on MPLS-Based VPNs

QoS for Critical Applications

QoS Design Overview

Next generation (NG)-WAN/MAN networks form the backbone of business-ready networks. These networks transport a multitude of applications, including real-time voice, high-quality video, and delay-sensitive data. NG-WAN/MAN networks must therefore provide predictable, measurable, and sometimes guaranteed services by managing bandwidth, delay, jitter, and loss parameters on a network via QoS technologies.

QoS technologies refer to the set of tools and features available within Cisco hardware and software to manage network resources; these include classification and marking tools, policing and shaping tools, congestion management and congestion avoidance tools, as well as link-efficiency mechanisms. QoS is considered the key enabling technology for network convergence. The objective of QoS technologies is to make voice, video, and data convergence appear transparent to end users. QoS technologies allow different types of traffic to contend inequitably for network resources. Voice, video, and critical data applications may be granted priority or preferential services from network devices so that the quality of these strategic applications does not degrade to the point of being unusable. Therefore, QoS is a critical, intrinsic element for successful network convergence. However, QoS tools are not only useful in protecting desirable traffic, but also in providing deferential services to undesirable traffic such as the exponential propagation of worms.

A successful QoS deployment is comprised of multiple phases, including the following:

- Strategically defining the business objectives to be achieved via QoS.
- Analyzing the service level requirements of the various traffic classes for which to be provisioned.
- Designing and testing QoS policies before production network rollout.
- Rolling out the tested QoS designs to the production network.
- Monitoring service levels to ensure that the QoS objectives are being met.

These phases may need to be repeated as business conditions change and evolve. The following sections focus on the first three phases of a QoS deployment and specifically adapt best-practice QoS design to the NG-WAN/MAN.

Strategically Defining the Business Objectives

QoS technologies are the enablers for business/organizational objectives. Therefore, the way to begin a QoS deployment is not to activate QoS features simply because they exist, but to start by clearly defining the objectives of the organization. For example, among the first questions that arise during a QoS deployment are the following: How many traffic classes should be provisioned for? What should they be?

To help answer these fundamental questions, organizational objectives need to be defined, such as the following:

- Is the objective to enable VoIP only or is video also required?
- If video is required, is video-conferencing required or streaming video? Or both?
- Are there applications that are considered mission-critical and if so, what are they?
- Does the organization wish to squelch certain types of traffic and if so, what are they?

To help address these crucial questions and to simplify QoS, Cisco has adopted a new initiative called the "QoS Baseline." The QoS Baseline is a strategic document designed to unify QoS within Cisco from enterprise to service provider and from engineering to marketing. The QoS Baseline was written by the most qualified Cisco QoS experts, who have developed or contributed to the related IETF RFC standards (as well as other technology standards) and are thus eminently qualified to interpret these standards. The QoS Baseline also provides uniform, standards-based recommendations to help ensure that QoS designs and deployments are unified and consistent. The QoS Baseline defines up to 11 classes of traffic that may be viewed as critical to a given enterprise. A summary these classes and their respective standards-based marking recommendations are presented in Table 4-1.

Table 4-1 Cisco QoS Baseline/Technical Marketing (Interim) Classification and Marking Recommendations

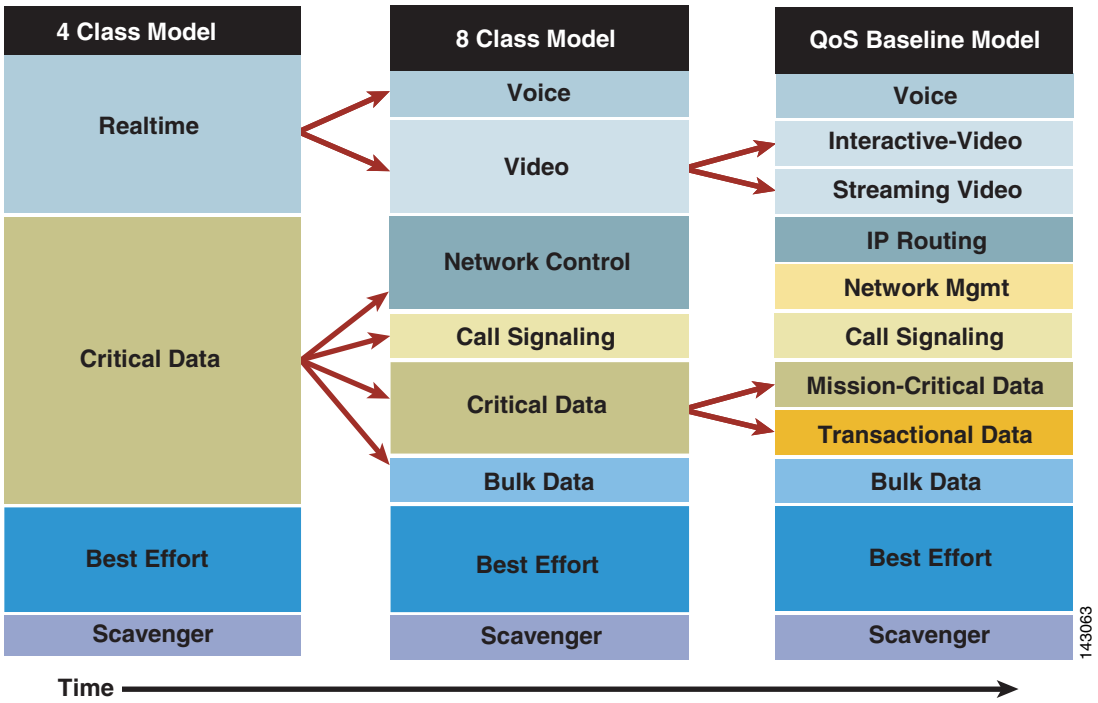
Application	Classification		Referencing Standard	Recommended Configuration
	PHB	DSCP		
IP Routing	CS6	48	RFC 2474-4.2.2	Rate-based Queuing + RED
Voice	FF	46	RFC 3246	RSVP Admission Control + Priority
Interactive-Video	AF 41	34	RFC 2957	RSVP + Rate-Based Queuing + DSCP
Streaming Video	CS4	32	RFC 2474-4.2.2	RSVP + Rate-Based Queuing + RED
Mission-Critical	AF 31	26	RFC 2597	Rate-Based Queuing + DSCP-WRED
Call Signaling	CS3	24	RFC 2474-4.2.2	Rate-Based Queuing + RED
Transactional Data	AF 21	18	RFC 2597	Rate-Based Queuing + DSCP-WRED
Network Mgmt	CS2	16	RFC 2474-4.2.2	Rate-based Queuing + RED
Bulk Data	AF 11	10	RFC 2597	Rate-Based Queuing + DSCP-WRED
Scavenger	CS1	8	Internet 2	No BW Guarantee + RED
Best Effort	0	0	RFC 2474-4.1	BW Guarantee Rate-Based Queuing

Note The QoS Baseline recommends marking Call-Signaling to CS3. However, originally Cisco IP telephony products marked Call-Signaling to AF31. A marking migration is under way within Cisco to change all IP telephony products to mark Call-Signaling to CS3 by default. For companies deploying IP telephony products that still might be using AF31 for Call-Signaling, a recommended interim marking strategy is to use both AF31 and CS3 for Call-Signaling marking and to mark Locally-Defined Mission-Critical Data applications to a temporary placeholder (non-standard) DSCP, such as 25. Upon completion of the migration, the QoS Baseline marking recommendations of CS3 for Call-Signaling and AF31 for Locally-Defined Mission-Critical Data applications should be used. These marking recommendations are more in line with RFC 2474 and RFC 2597.

Enterprises do not need to deploy all 11 classes of the QoS Baseline model. This model is intended to be a forward-looking guide that considers as many classes of traffic with unique QoS requirements as possible. Familiarity with this model can assist in the smooth expansion of QoS policies to support additional applications as future requirements arise.

However, at the time of QoS deployment, the enterprise needs to clearly define their organizational objectives, which correspondingly determine how many traffic classes are required. This consideration should be tempered with the determination of how many application classes the networking administration team feels comfortable with deploying and supporting. Platform-specific constraints or service provider constraints may also affect the number of classes of service. At this point, you should also consider a migration strategy to allow the number of classes to be smoothly expanded as future needs arise, as shown in Figure 4-1.

Figure 4-1 Example Strategy for Expanding the Number of Classes of Service over Time



Platform limitations do not necessarily have to be considered as gating factors to the number of classes that can be supported over the NG-WAN/MAN. On platforms that support more classes, these may be configured; on platforms that support fewer classes, some classes must be collapsed to accommodate hardware limitations. The main consideration is that policies need to be kept consistent and complementary to achieve expected per-hop behaviors.

A strategic standards-based guide such as the QoS Baseline coupled with a working knowledge of QoS tools and syntax is a prerequisite for any successful QoS deployment. However, you must also understand the service level requirements of the various applications requiring preferential or deferential treatment within the network.

Analyzing the Service Level Requirements

QoS Requirements of VoIP

This section includes the following topics:

- Voice (bearer traffic)
- Call -Signaling traffic

VoIP deployments require provisioning explicit priority servicing for VoIP (bearer stream) traffic and a guaranteed bandwidth service for Call-Signaling traffic. These related classes are examined separately.

VoIP (Bearer) Traffic

The following is a summary of the key QoS requirements and recommendations for Voice (bearer traffic):

- Voice traffic should be marked to DSCP EF per the QoS Baseline and RFC 3246.
- Loss should be no more than 1 percent.
- One-way latency (mouth-to-ear) should be no more than 150 ms.
- Average one-way jitter should be targeted under 30 ms.
- 21–320 kbps of guaranteed priority bandwidth is required per call (depending on the sampling rate, VoIP codec, and Layer 2 media overhead).

Voice quality is directly affected by all three QoS quality factors of loss, latency, and jitter.

Loss causes voice clipping and skips. The packetization interval determines the size of samples contained within a single packet. Assuming a 20 ms (default) packetization interval, the loss of two or more consecutive packets results in a noticeable degradation of voice quality. VoIP networks are typically designed for very close to zero percent VoIP packet loss, with the only actual packet loss being because of Layer 2 bit errors or network failures.

Excessive latency can cause voice quality degradation. The goal commonly used in designing networks to support VoIP is the target specified by ITU standard G.114, which states that 150 ms of one-way, end-to-end (mouth-to-ear) delay ensures user satisfaction for telephony applications. A design should apportion this budget to the various components of network delay (propagation delay through the backbone, scheduling delay because of congestion, and the access link serialization delay) and service delay (because of VoIP gateway codec and de-jitter buffer).

If the end-to-end voice delay becomes too long, the conversation begins to sound like two parties talking over a satellite link or even a CB radio. Although the ITU G.114 states that a 150 ms one-way (mouth-to-ear) delay budget is acceptable for high voice quality, lab testing has shown that there is a negligible difference in voice quality mean opinion scores (MOS) using networks built with 200 ms

delay budgets. Cisco thus recommends designing to the ITU standard of 150 ms, but if constraints exist where this delay target cannot be met, then the delay boundary can be extended to 200 ms without significant impact on voice quality.

Jitter buffers (also known as play-out buffers) are used to change asynchronous packet arrivals into a synchronous stream by turning variable network delays into constant delays at the destination end systems. The role of the jitter buffer is to balance the delay and the probability of interrupted play-out because of late packets. Late or out-of-order packets are discarded.

If the jitter buffer is either set arbitrarily large or arbitrarily small, then it imposes unnecessary constraints on the characteristics of the network. A jitter buffer set too large adds to the end-to-end delay, meaning that less delay budget is available for the network such that the network needs to support a delay target tighter than practically necessary. If a jitter buffer is too small to accommodate the network jitter, then buffer underflows or overflows can occur.

An underflow occurs when the buffer is empty when the codec needs to play out a sample. An overflow occurs when the jitter buffer is already full and another packet arrives that cannot therefore be queued in the jitter buffer. Both jitter buffer underflows and overflows cause packets to be discarded.

Adaptive jitter buffers aim to overcome these issues by dynamically tuning the jitter buffer size to the lowest acceptable value. Where such adaptive jitter buffers are used, you can in theory engineer out explicit considerations of jitter by accounting for worst-case per hop delays. Advanced formulas can be used to arrive at network-specific design recommendations for jitter based on maximum and minimum per-hop delays. Alternatively, a 30 ms value can be used as a jitter target because extensive lab testing has shown that when jitter consistently exceeds 30 ms, voice quality degrades significantly.

Because of its strict service level requirements, VoIP is well suited to the expedited forwarding per-hop behavior, as defined in RFC 3246 (formerly RFC 2598). It should therefore be marked to DSCP EF (46) and assigned Strict Priority servicing at each node, regardless of whether such servicing is done in hardware (as in the Cisco 7600 or 12000 routers via hardware priority queuing) or in software (as in Cisco 7200 routers via LLQ).

Call-Signaling Traffic

The following are key QoS requirements and recommendations for Call-Signaling traffic:

- Call-Signaling traffic should be marked as DSCP CS3 per the QoS Baseline (during migration, it may also be marked the legacy value of DSCP AF31).
- 150 bps (plus Layer 2 overhead) per phone of guaranteed bandwidth is required for Voice control traffic; more may be required, depending on the call signaling protocol(s) in use.

Call-Signaling traffic was originally marked by Cisco IP telephony equipment to DSCP AF31. However, the Assured Forwarding classes, as defined in RFC 2597, were intended for flows that could be subject to markdown and subsequently the aggressive dropping of marked-down values. Marking down and aggressively dropping Call-Signaling can result in noticeable delay-to-dial-tone (DDT) and lengthy call setup times, both of which generally translate to poor user experiences.

The QoS Baseline changed the marking recommendation for Call-Signaling traffic to DSCP CS3 because class selector code points, as defined in RFC 2474, were not subject to markdown/aggressive-dropping per-hop behaviors. Most Cisco IP telephony products have already begun transitioning to DSCP CS3 for Call-Signaling marking. If the enterprise is still in a migration between older and new IP telephony products and software, during the interim period both code points (CS3 and AF31) should be reserved for Call-Signaling marking until the transition is complete.

QoS Requirements of Video

This section describes the two main types of video traffic and includes the following topics:

- Interactive-Video
- Streaming Video

Interactive-Video

When provisioning for Interactive-Video (IP videoconferencing) traffic, the following guidelines are recommended:

- Interactive-Video traffic should be marked to DSCP AF41; excess Interactive-Video traffic can be marked down by a policer to AF42 or AF43.
- Loss should be no more than 1 percent.
- One-way latency should be no more than 150 ms.
- Jitter should be no more than 30 ms.
- Interactive-Video queues should be overprovisioned by 20 percent to accommodate bursts.

Because IP Videoconferencing (IP/VC) includes an audio codec for voice and relies on extending a real-time user experience to the video conference, it has the same loss, delay, and delay variation requirements as voice, but the traffic patterns of videoconferencing are radically different from voice.

Because (unlike VoIP) IP/VC packet sizes and rates vary given the motion-based nature of the video codec, the header overhead percentage varies as well, so an absolute value of bandwidth utilization and overhead cannot be accurately calculated for all streams. Testing, however, has shown a conservative rule of thumb for IP/VC bandwidth provisioning is to overprovision the guaranteed/priority bandwidth by 20 percent over the video call rate, which accounts for Layer 2 and Layer 3 overhead and a maximum transmission rate. For example, a user running a 384kbps video call (64kbps audio, 320 kbps video) uses a maximum bandwidth of 384kbps plus 20 percent, for approximate peak usage of 460kbps.

Streaming Video

When addressing the QoS needs of Streaming Video traffic, the following guidelines are recommended:

- Streaming Video (whether unicast or multicast) should be marked to DSCP CS4 as designated by the QoS Baseline.
- Loss should be no more than 5 percent.
- Latency should be no more than 4–5 seconds (depending on video application buffering capabilities).
- There are no significant jitter requirements.
- Guaranteed bandwidth (CBWFQ) requirements depend on the encoding format and rate of the video stream.

Streaming video applications have more lenient QoS requirements because they are delay-insensitive (the video can take several seconds to cue-up) and are largely jitter-insensitive (because of application buffering). However, streaming video may contain valuable content, such as e-learning applications or multicast company meetings, and therefore may require service guarantees.

The QoS Baseline recommendation for Streaming Video marking is DSCP CS4.

Non-organizational video content (or video that is strictly entertainment-oriented in nature such as movies, music videos, humorous commercials, and so on) might be considered for a ("less-than-Best-Effort") Scavenger service. This means that these streams play if bandwidth exists, but they are the first to be dropped during periods of congestion.

QoS Requirements of Data

There are hundreds of thousands of data networking applications. Some are TCP, others are UDP; some are delay sensitive, others are not; some are bursty in nature, others are steady; some are lightweight, others require high bandwidth, and so on. Not only do applications vary one from another, but even the same application can vary significantly between versions.

Given this, determining how to best provision QoS for data is a daunting proposition. The Cisco QoS Baseline identifies four main classes of data traffic, according to their general networking characteristics and requirements:

- Best Effort
- Bulk Data
- Transactional/Interactive Data
- Locally-Defined Mission-Critical Data

Best Effort Data

The Best Effort class is the default class for all data traffic. An application is removed from the default class only if it has been selected for preferential or deferential treatment.

When addressing the QoS needs of Best Effort data traffic, Cisco recommends the following guidelines:

- Best Effort traffic should be marked to DSCP 0.
- Adequate bandwidth should be assigned to the Best Effort class as a whole, because the majority of applications default to this class; reserve at least 25 percent for Best Effort traffic.

Typical enterprises have several hundred, if not thousands, of data applications running over their networks (the majority of which default to the Best Effort class). Therefore, you need to provision adequate bandwidth for the default class as a whole to handle the sheer volume of applications that will be included in it. Otherwise, applications defaulting to this class are easily drowned out, which typically results in an increased number of calls to the networking help desk from frustrated users. Cisco therefore recommends that you reserve at least 25 percent of link bandwidth for the default Best Effort class.

Bulk Data

The Bulk Data class is intended for applications that are relatively non-interactive and drop-insensitive and that typically span their operations over a long period of time as background occurrences. Such applications include the following:

- FTP
- E-mail
- Backup operations
- Database synchronizing or replicating operations
- Content distribution
- Any other type of background operation

When addressing the QoS needs of Bulk Data traffic, Cisco recommends the following guidelines:

- Bulk Data traffic should be marked to DSCP AF11; excess Bulk Data traffic can be marked down by a policer to AF12; violating bulk data traffic may be marked down further to AF13 (or dropped).
- Bulk Data traffic should have a moderate bandwidth guarantee, but should be constrained from dominating a link.

The advantage of provisioning moderate bandwidth guarantees to Bulk Data applications rather than applying policers to them is that Bulk Data applications can dynamically take advantage of unused bandwidth and thus speed up their operations during non-peak periods. This in turn reduces the likelihood of their bleeding into busy periods and absorbing inordinate amounts of bandwidth for their time-insensitive operations.

Transactional/Interactive Data

The Transactional/Interactive Data class, also referred to simply as Transactional Data, is a combination of two similar types of applications: Transactional Data client-server applications and Interactive Messaging applications.

The response time requirement separates Transactional Data client-server applications from generic client-server applications. For example, with Transactional Data client-server applications such as SAP, PeopleSoft, and Data Link Switching (DLSw+), the transaction is a foreground operation; the user waits for the operation to complete before proceeding.

E-mail is not considered a Transactional Data client-server application, because most e-mail operations occur in the background and users do not usually notice even several hundred millisecond delays in mailspool operations.

When addressing the QoS needs of Transactional Data traffic, Cisco recommends the following guidelines:

- Transactional Data traffic should be marked to DSCP AF21; excess Transactional Data traffic can be marked down by a policer to AF22; violating Transactional Data traffic can be marked down further to AF23 (or dropped).
- Transactional Data traffic should have an adequate bandwidth guarantee for the interactive, foreground operations they support.

Locally-Defined, Mission-Critical Data

The Locally-Defined Mission-Critical Data class is probably the most misunderstood class specified in the QoS Baseline. Under the QoS Baseline model, all traffic classes (with the exclusion of Scavenger and Best Effort) are considered critical to the enterprise. The term “locally-defined” is used to underscore the purpose of this class, which is to provide each enterprise with a premium class of service for a select subset of their Transactional Data applications that have the highest business priority for them.

For example, an enterprise may have properly provisioned Oracle, SAP, BEA, and DLSw+ within their Transactional Data class. However, the majority of their revenue may come from SAP, and therefore they may want to give this Transactional Data application an even higher level of preference by assigning it to a dedicated class such as the Locally-Defined Mission-Critical Data class.

Because the admission criteria for this class is non-technical (being determined by business relevance and organizational objectives), the decision of which applications should be assigned to this special class can easily become an organizationally- and politically-charged debate. Cisco recommends that you assign as few applications to this class from the Transactional Data class as possible. You should also obtain executive endorsement for application assignments to the Locally-Defined Mission-Critical Data class, because the potential for QoS deployment derailment exists without such an endorsement.

For the sake of simplicity, this class is referred to simply as Mission-Critical Data. When addressing the QoS needs of Mission-Critical Data traffic, Cisco recommends the following guidelines:

- Mission-Critical Data traffic should be marked to DSCP AF31; excess Mission-Critical Data traffic can then be marked down by a policer to AF22 or AF23. However, DSCP AF31 is currently being used by Cisco IP telephony equipment as Call-Signaling, so until all Cisco IPT products mark Call-Signaling to DSCP CS3, a temporary placeholder code point (DSCP 25) can be used to identify Mission-Critical Data traffic.
- Mission-Critical Data traffic should have an adequate bandwidth guarantee for the interactive, foreground operations they support.

QoS Requirements of the Control Plane

This section includes the following topics:

- IP Routing
- Network Management

Unless the network is up and running, QoS is irrelevant. Therefore, it is critical to provision QoS for control plane traffic, which includes IP Routing and Network Management traffic.

IP Routing

By default, Cisco IOS software (in accordance with RFC 791 and RFC 2474) marks Interior Gateway Protocol (IGP) traffic such as Routing Information Protocol (RIP/RIPv2), Open Shortest Path First (OSPF), and Enhanced Interior Gateway Routing Protocol (EIGRP) to DSCP CS6. However, Cisco IOS software also has an internal mechanism for granting internal priority to important control datagrams as they are processed within the router. This mechanism is called PAK_PRIORITY.

As datagrams are processed though the router and down to the interfaces, they are internally encapsulated with a small packet header, referred to as the PAKTYPE structure. Within the fields of this internal header there is a PAK_PRIORITY flag that indicates the relative importance of control packets to the internal processing systems of the router. PAK_PRIORITY designation is a critical internal Cisco IOS software operation and, as such, is not administratively configurable in any way.

Note that Exterior Gateway Protocol (EGP) traffic such as Border Gateway Protocol (BGP) traffic is marked by default to DSCP CS6, but does not receive such PAK_PRIORITY preferential treatment and may need to be explicitly protected to maintain peering sessions.

When addressing the QoS needs of IP Routing traffic, Cisco recommends the following guidelines:

- IP Routing traffic should be marked to DSCP CS6; this is default behavior on Cisco IOS platforms.
- IGPs are usually adequately protected with the Cisco IOS internal PAK_PRIORITY mechanism; Cisco recommends that EGPs such as BGP have an explicit class for IP Routing with a minimal bandwidth guarantee.
- Cisco IOS automatically marks IP Routing traffic to DSCP CS6.

Additional information on PAK_PRIORITY can be found at the following URL: http://www.cisco.com/warp/public/105/rtgupdates.html.

Network Management

When addressing the QoS needs of Network Management traffic, Cisco recommends the following guidelines:

- Network Management traffic should be marked to DSCP CS2.
- Network Management applications should be explicitly protected with a minimal bandwidth guarantee.

Network Management traffic is important to perform trend and capacity analysis and troubleshooting. Therefore, you can provision a separate minimal bandwidth queue for Network Management traffic, which could include SNMP, NTP, Syslog, NFS, and other management applications.

Scavenger Class QoS

The Scavenger class, based on an Internet-II draft, is intended to provide deferential services, or "less-than-Best-Effort" services, to certain applications. Applications assigned to this class have little or no contribution to the organizational objectives of the enterprise and are typically entertainment-oriented. These include peer-to-peer (P2P) media-sharing applications (such as KaZaa, Morpheus, Grokster, Napster, iMesh, and so on), gaming applications (Doom, Quake, Unreal Tournament, and so on), and any entertainment video applications.

Assigning a minimal bandwidth queue to Scavenger traffic forces it to be squelched to virtually nothing during periods of congestion, but allows it to be available if bandwidth is not being used for business purposes, such as might occur during off-peak hours. This allows for a flexible, non-stringent policy control of non-business applications.

When provisioning for Scavenger traffic, Cisco recommends the following guidelines:

- Scavenger traffic should be marked to DSCP CS1.
- Scavenger traffic should be assigned the lowest configurable queuing service; for instance, in Cisco IOS this would mean assigning a CBWFQ of 1 percent to Scavenger.

The Scavenger class is a critical component to the data plane policing DoS/worm mitigation strategy presented in the Enterprise QoS SRND 3.1 at www.cisco.com/go/srnd.

Designing the QoS Policies

After a QoS strategy has been defined and the application requirements are understood, end-to-end QoS policies can be designed for each device and interface as determined by its role in the network infrastructure. Because the Cisco QoS toolset provides many QoS design and deployment options, a few succinct design principles can help simplify strategic QoS deployments.

Additionally, these best-practice design principles need to be coupled with topology-specific considerations and constraints. Therefore, the second part of this design section contains a discussion of QoS design considerations specific to the NG-WAN/MAN.

QoS Design Best Practices

For example, one such design principle is to always enable QoS policies in hardware rather than software whenever a choice exists. Lower-end to mid-range Cisco IOS routers (such as the Cisco 1700 through Cisco 7500) perform QoS in software, which places incremental loads on the CPU, depending on the complexity and functionality of the policy. On the other hand, Cisco Catalyst switches and high-end routers (such as the Cisco 7600 and Cisco 12000) perform QoS in dedicated hardware ASICS and as such do not tax their main CPUs to administer QoS policies. This allows complex policies to be applied at line rates at even GE, 10 GE, or higher speeds.

Other simplifying best-practice QoS design principles include the following:

- Classification and marking principles
- Policing and markdown principles
- Queuing and dropping principles

Classification and Marking Design Principles

When classifying and marking traffic, an unofficial differentiated services design principle is to classify and mark applications as close to their sources as technically and administratively feasible. This principle promotes end-to-end differentiated services and per-hop behaviors (PHBs).

Furthermore, it is recommended to use DSCP markings whenever possible, because these are end-to-end, more granular, and more extensible than Layer 2 markings. Layer 2 markings are lost when media changes (such as a LAN-to-WAN/VPN edge). There is also less marking granularity at Layer 2; for example, 802.1Q/p CoS supports only three bits (values 0–7), as does MPLS EXP. Hence only up to eight classes of traffic can be supported at Layer 2 and inter-class relative priority (such as RFC 2597 Assured Forwarding Drop Preference markdown) is not supported. On the other hand, Layer 3 DSCP markings allow for up to 64 classes of traffic, which is more than enough for most enterprise requirements for the foreseeable future.

As the line between enterprises and service providers continues to blur and the need for interoperability and complementary QoS markings is critical, you should follow standards-based DSCP PHB markings to ensure interoperability and future expansion. Because the QoS Baseline marking recommendations are standards-based, enterprises can easily adopt these markings to interface with service provider classes of service. Network mergers (whether the result of acquisitions, mergers, or strategic alliances) are also easier to manage when you use standards-based DSCP markings.

Policing and Markdown Design Principles

There is little reason to forward unwanted traffic only to police and drop it at a subsequent node, especially when the unwanted traffic is the result of DoS or worm attacks. The overwhelming volume of traffic that such attacks can create can cause network outages by driving network device processors to their maximum levels. Therefore, you should police traffic flows as close to their sources as possible.

Whenever supported, markdown should be done according to standards-based rules, such as RFC 2597 (“Assured Forwarding PHB Group”). For example, excess traffic marked to AFx1 should be marked down to AFx2 (or AFx3 whenever dual-rate policing such as defined in RFC 2698 is supported). Following such markdowns, congestion management policies, such as DSCP-based WRED, should be configured to drop AFx3 more aggressively than AFx2, which in turn should be dropped more aggressively than AFx1.

Queuing and Dropping Design Principles

Critical applications such as VoIP require service guarantees regardless of network conditions. The only way to provide service guarantees is to enable queuing at any node that has the potential for congestion, regardless of how rarely this may occur. There is simply no other way to guarantee service levels than by enabling queuing wherever a speed mismatch exists.

When provisioning queuing, some best practice rules of thumb also apply. For example, as discussed previously, the Best Effort class is the default class for all data traffic. Only if an application has been selected for preferential/deferential treatment is it removed from the default class. Because many enterprises have several hundred, if not thousands, of data applications running over their networks, you must provision adequate bandwidth for this class as a whole to handle the sheer volume of applications that default to it. Therefore, it is recommended that you reserve at least 25 percent of link bandwidth for the default Best Effort class.

Not only does the Best Effort class of traffic require special bandwidth provisioning consideration, so does the highest class of traffic, sometimes referred to as the Real-time or Strict Priority class (which corresponds to RFC 3246, “An Expedited Forwarding Per-Hop Behavior”). The amount of bandwidth assigned to the Real-time queuing class is variable. However, if you assign too much traffic for Strict Priority queuing, then the overall effect is a dampening of QoS functionality for non-real-time

applications. Remember that the goal of convergence is to enable voice, video, and data to transparently co-exist on a single network. When real-time applications such as voice or interactive video dominate a link (especially a WAN/VPN link), then data applications fluctuate significantly in their response times, destroying the transparency of the converged network.

Extensive testing and customer deployments have shown that a general best queuing practice is to limit the amount of Strict Priority queuing to 33 percent of link capacity. This Strict Priority queuing rule is a conservative and safe design ratio for merging real-time applications with data applications.

Cisco IOS software allows the abstraction (and thus configuration) of multiple Strict Priority LLQs. In such a multiple LLQ context, this design principle applies to the sum of all LLQs to be within one-third of link capacity.

Note This Strict Priority queuing rule (limit to 33 percent) is simply a best practice design recommendation and is not a mandate. There may be cases where specific business objectives cannot be met while holding to this recommendation. In such cases, enterprises must provision according to their detailed requirements and constraints. However, it is important to recognize the tradeoffs involved with over-provisioning Strict Priority traffic and its negative performance impact on non-real-time-application response times.

Whenever a Scavenger queuing class is enabled, it should be assigned a minimal amount of bandwidth. On some platforms, queuing distinctions between Bulk Data and Scavenger traffic flows cannot be made because queuing assignments are determined by CoS values and these applications share the same CoS value of 1. In such cases you can assign the Scavenger/Bulk Data queuing class a bandwidth percentage of 5. If you can uniquely assign Scavenger and Bulk Data to different queues, then you should assign the Scavenger queue a bandwidth percentage of 1.

NG-WAN/MAN QoS Design Considerations

A few NG-WAN/MAN-specific considerations also come into play when drafting QoS designs for this network. Most of these fall under two main headings:

- MPLS DiffServ tunneling modes, which are discussed in this section.
- Platform-specific capabilities/constraints, discussed in Appendix A, “Platform-Specific Capabilities and Constraints.”

To maintain the class-based per hop behavior, we recommend that you implement the 8 class model where ever possible (see Figure 4-1 for reference) within the MPLS MAN. In scenarios where this is not feasible, at least 5 classes should be deployed—Realtime, Critical, Video, Bulk, and Best Effort. This allows the video traffic to be in a separate queue and keeps bulk data separated from critical data.

MPLS DiffServ Tunneling Modes

A marking disparity exists between MPLS VPNs and DiffServ IP networks:

Layer 2 MPLS labels support only 3 bits for marking (referred to as MPLS EXP bits) offering 8 levels of marking options.

Layer 3 IP packets support 6 bits for marking (the Differentiated Services Code Point [DSCP]).

Because of the reduced level of granularity in marking at Layer 2 via MPLS EXP bits, there is a potential for a “loss in translation” as (potentially) 64 levels of marking cannot be faithfully reduced to 8, nor can 64 levels be faithfully re-created from 8.

To address this disparity, RFC 3270, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services" presents the following three modes to manage the translation and/or preservation (tunneling) of DiffServ over MPLS VPN networks:

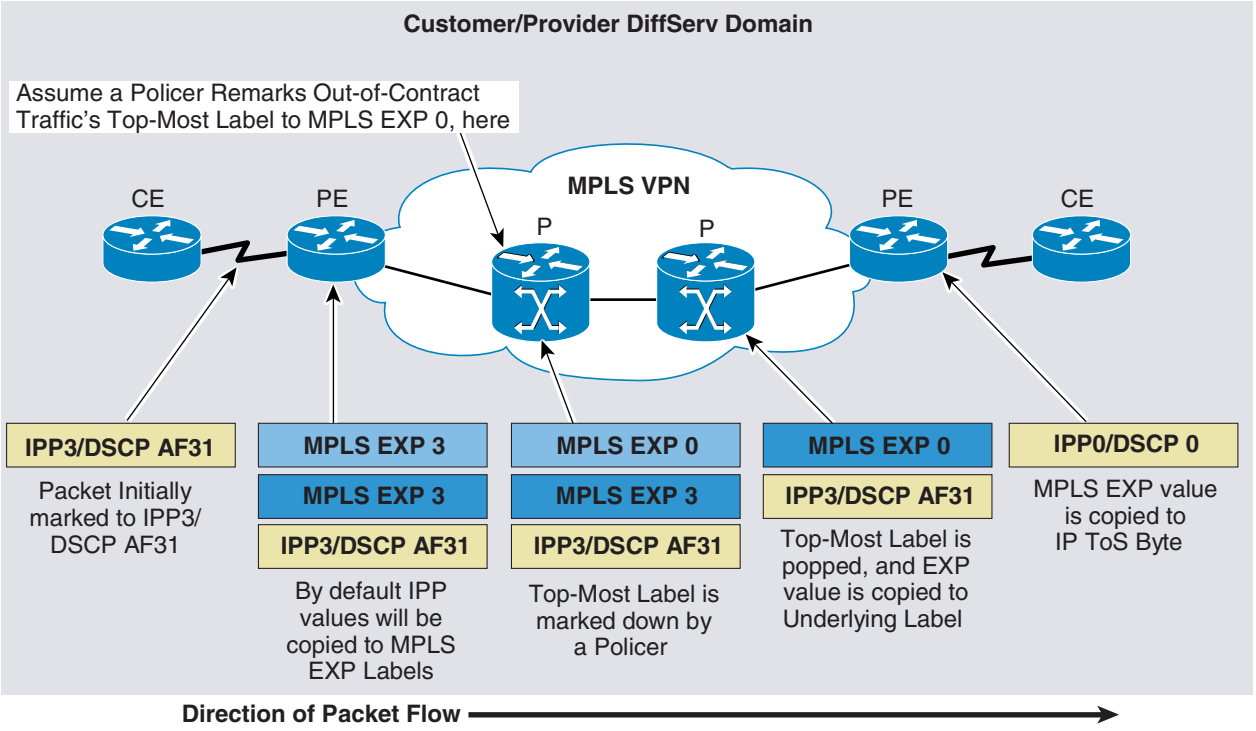
- Uniform mode
- Short-pipe mode
- Pipe mode

Uniform Mode

Uniform mode is used when the customer and service provider share the same DiffServ domain, meaning that a single administrative marking policy is applied over the entire network. In uniform mode, packets are treated uniformly in the IP and MPLS networks; that is, the IP precedence value and the MPLS EXP bits always correspond to the same PHB. Whenever a router changes or recolors the PHB of a packet that change must be propagated to all encapsulation markings. The propagation is performed by a router only when a PHB is added or exposed because of label imposition or disposition on any router in the packet path. The color must be reflected everywhere at all levels. For example, if a packet QoS marking is changed in the MPLS network, the IP QoS marking reflects that change.

Uniform mode is the preferred MPLS DiffServ Tunneling mode for the NG-WAN/MAN and is shown in Figure 4-2.

Figure 4-2 Uniform Mode MPLS DiffServ Tunneling

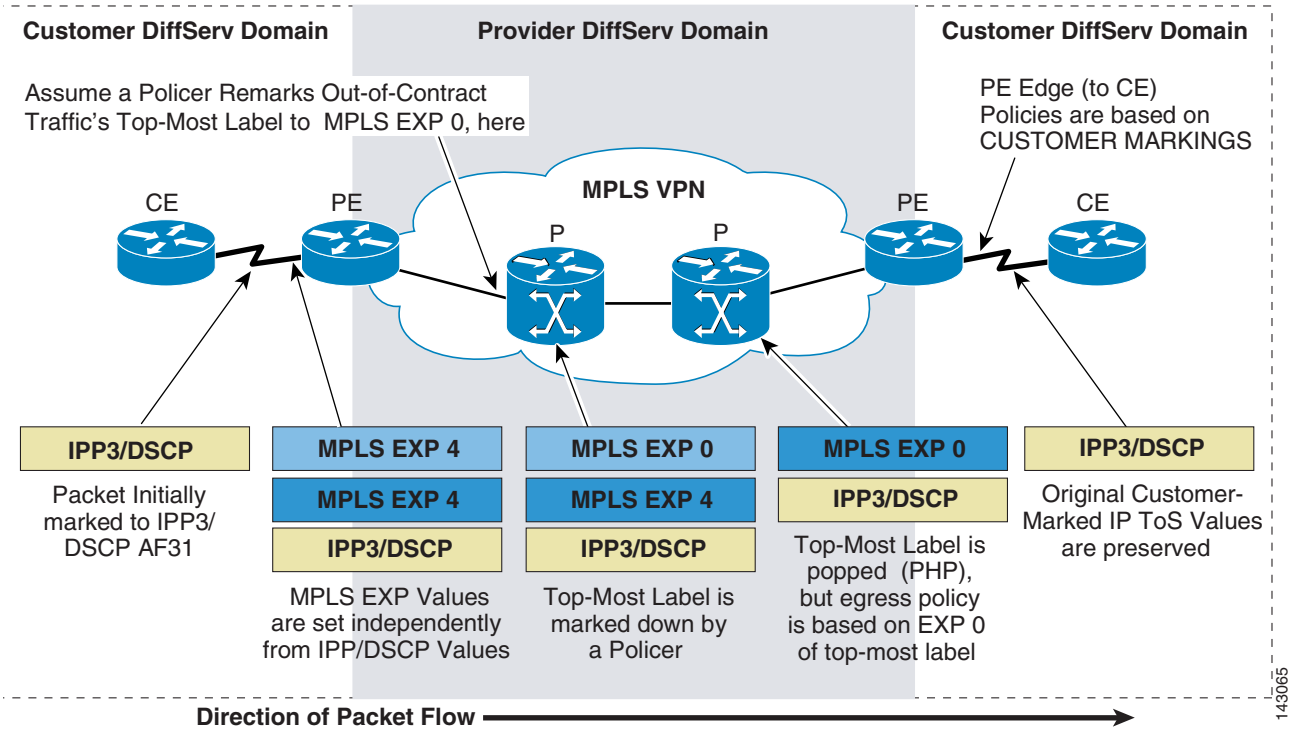


Short-Pipe Mode

Short-pipe mode is used when the customer and service provider are in different DiffServ domains. It allows the service provider to enforce its own DiffServ policy while preserving customer DiffServ information, which provides a DiffServ transparency through the service provider network.

QoS policies implemented in the core do not propagate to the Layer 3 IP packet ToS byte. The classification based on MPLS EXP value ends at the customer-facing egress PE interface; classification at the customer-facing egress PE interface is based on the original IP packet header and not the MPLS header. The presence of an egress IP policy (based on the customer PHB marking and not on the provider PHB marking) automatically implies the short-pipe mode. Short-pipe mode is illustrated in Figure 4-3.

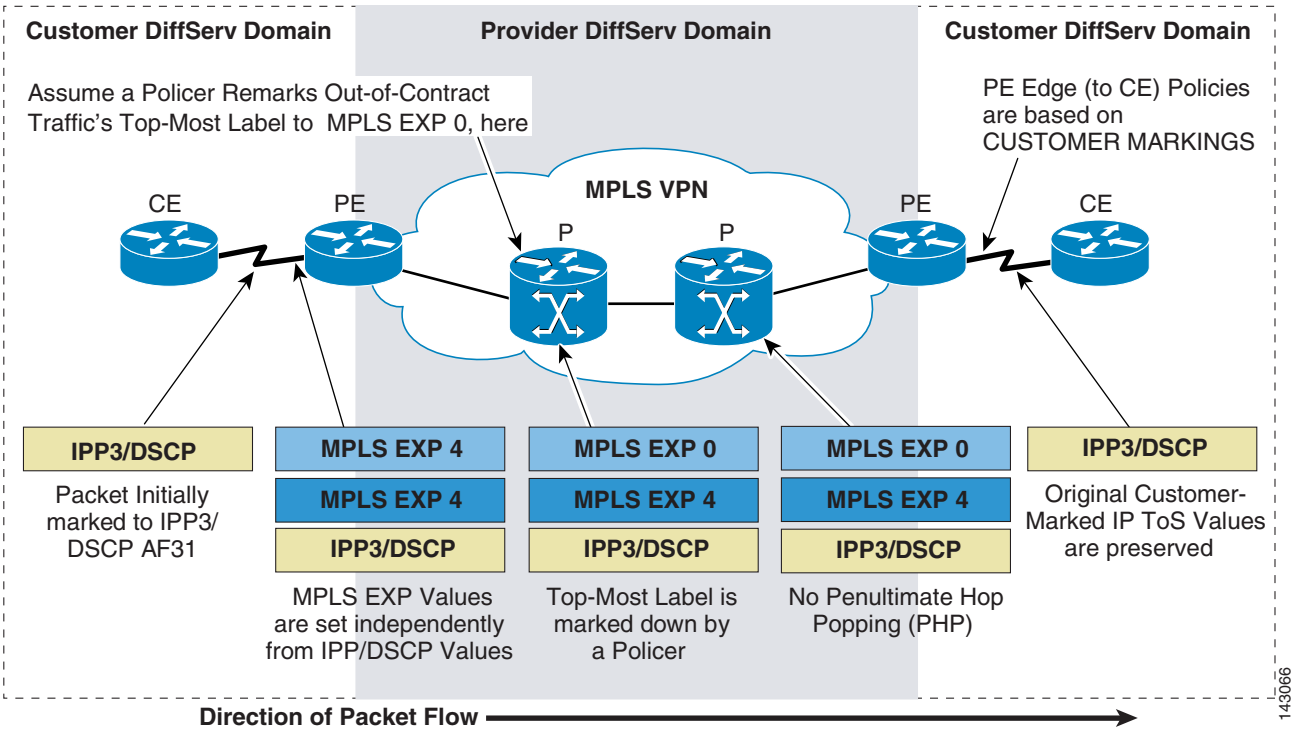
Figure 4-3 Short-Pipe Mode MPLS DiffServ Tunneling



Pipe Mode

Pipe mode is very similar to short-pipe mode, because the customer and service provider are in different DiffServ domains. The difference between the two is that with pipe mode the service provider derives the outbound classification for congestion management and congestion avoidance policies based on the service provider DiffServ policy (rather than according to the enterprise customer markings). This affects how the packet is scheduled/dropped on the egress PE toward the customer CE. Egress scheduling/dropping markings are maintained through the use of **qos-groups** and **discard-class** commands on the egress PE policy maps. This implementation avoids the additional operational overhead of per-customer configurations on each egress interface on the egress PE, as shown in Figure 4-4.

Figure 4-4 Pipe Mode MPLS DiffServ Tunneling



Note Platform-specific design considerations for the Cisco 7200, Cisco 7304,Cisco 7600, and Cisco 12xxx are disccused in Appendix A, "Platform-Specific Capabilities and Constraints."

Security

IP VPNs have a similar security level to that of separate Frame Relay or ATM circuits. Although the separation between the groups is logical, it is very difficult for a hacker to leak traffic from one VPN to another.

In the case of VPNs, isolation is based on the fact that each VPN has a separate logical control plane. This means that devices in one VPN do not know about the IP prefixes in other VPNs and therefore cannot reach these. This protects one VPN from another, but also protects the global routing space from being accessed by users/devices in any of the customer VPNs. By deploying Layer 3 VPNs, the core is therefore made invisible to the customers serviced in the different VPNs, which raises the level of security and availability of the network core.

Encryption

MPLS VPNs provide traffic separation through the logical isolation of the different control planes. MPLS VPNs do not provide encryption of the data in the VPNs. However, the deployment of MPLS VPNs does not preclude encryption of the data in each VPN. Because each VPN provides any-to-any IP connectivity between CE devices, it is possible to overlay many types of encryption architectures on top

of a VPN. Thus, solutions such as DMVPN or site-to-site IPsec can be used to encrypt traffic between CE devices within a VPN. These encryption solutions are a topic in themselves and are discussed in detail in the companion IPsec design guide.

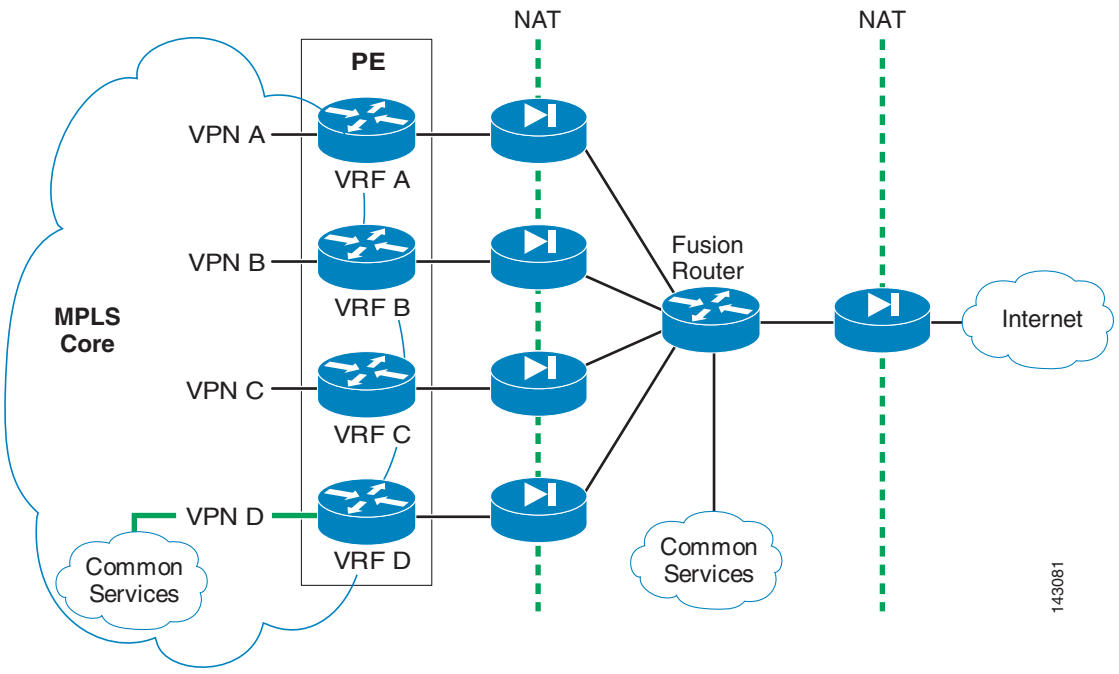
VPN Perimeter—Common Services and the Internet

The default state of a VPN is to be totally isolated from other VPNs. In this respect, VPNs can be seen as physically separate networks. However, because VPNs actually belong to a common physical network, it is desirable for these VPNs to share certain services such as Internet access, DHCP services, DNS services, or server farms. These services are usually located outside of the different VPNs or in a VPN of their own, so these VPNs must have a gateway to connect to the "outside world." The outside world is basically any network outside the VPN such as the Internet or other VPNs. Because this is the perimeter of the VPN, it is also desirable that this perimeter be protected by security devices such as firewalls and IDS. Typically, the perimeter is deployed at a common physical location for most VPNs. Thus, this location is known as the central services site.

The creation of VPNs can be seen as the creation of security zones, each of which has a unique and controlled entry/exit point at the VPN perimeter. Routing within the VPNs should be configured so that traffic is steered to the common services site as required.

Figure 4-5 illustrates a typical perimeter deployment for multiple VPNs accessing common services. Because the services accessed through the VPN perimeter are protected by firewalls, they are referred to as "protected services."

Figure 4-5 Central Site Providing VPN Perimeter Security

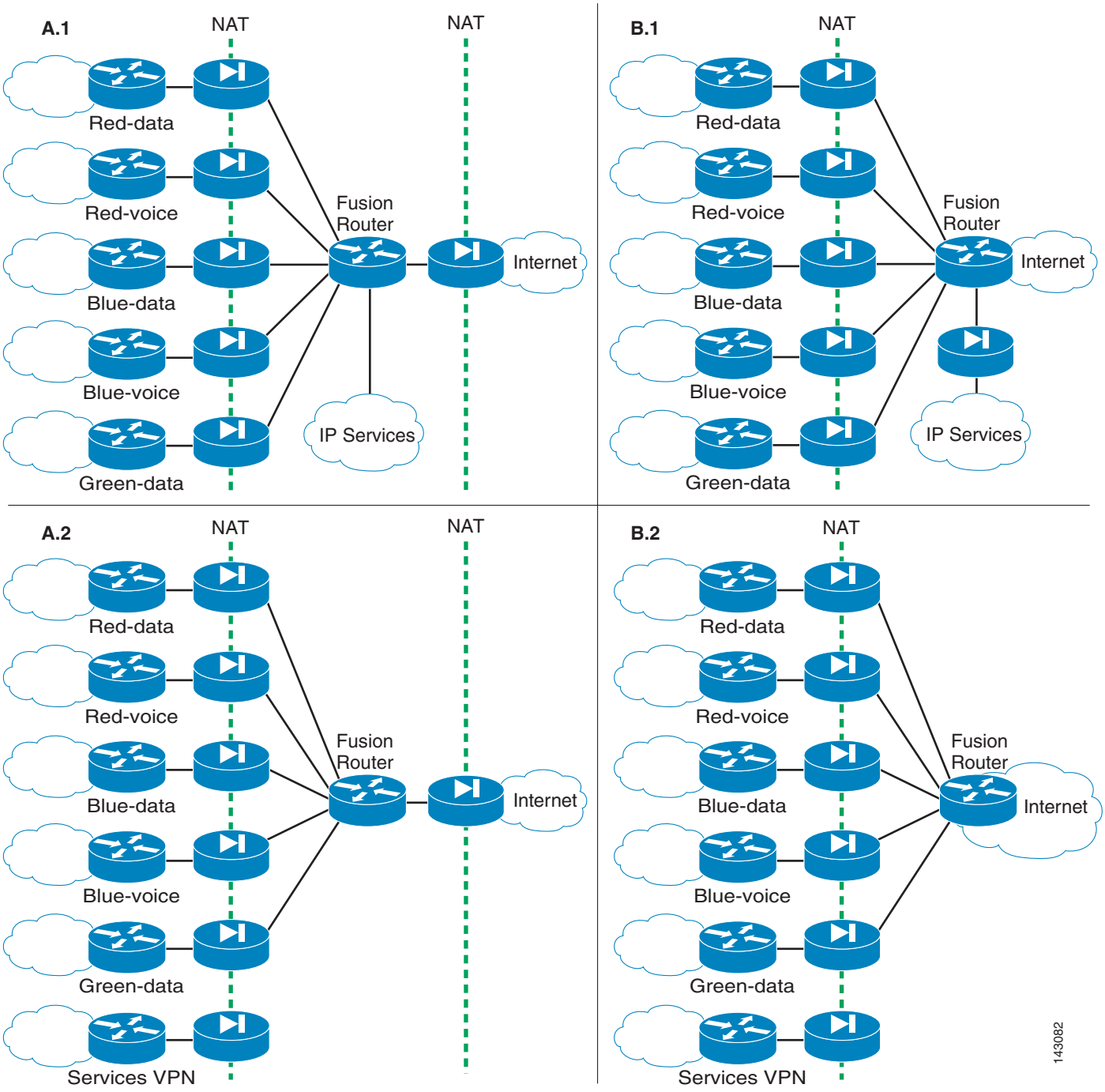


As seen in Figure 4-5, each VPN is head-ended by a dedicated firewall, which allows the creation of security policies specific to each VPN and independent from each other. To access the shared services, all firewalls are connected to a "fusion" router. The fusion router can provide the VPNs with connectivity to the common services, the Internet, or even inter-VPN connectivity. The presence of this fusion router raises two main concerns: the potential for traffic leaking between VPNs and the risk of routes from one VPN being announced to another VPN. The presence of dedicated per VPN firewalls prevents the leaking of traffic between VPNs through the fusion router by allowing only established connections to return through the VPN perimeter. It is important to configure the routing on the fusion device so that routes from one VPN are not advertised to another through the fusion router. The details of the routing configuration at the central site are discussed in Common Services.

Figure 4-5 shows an additional firewall separating the fusion area from the Internet. This firewall is optional. Whether to use it or not depends on the need to keep common services or transit traffic in the fusion area protected from the Internet.

Figure 4-6 illustrates the different scenarios for common services positioning and the Internet firewall.

Figure 4-6 Common Services Positioning



When the common services are not present or placed in their own VPN (and therefore front-ended by a dedicated firewall context), the additional Internet firewall can be removed as shown in diagram B.2. If there is a concern about transit traffic being on the Internet, then the firewall can be kept (see diagram A.2). The common services can be separated from the rest of the network by having their own firewall, yet not be included in a VPN, as shown in diagram B.1.

For scenarios B.1 and B.2, it is important to note that the fusion router is actually part of the Internet and thus the NAT pool employed at the firewalls must use valid Internet addresses. The deployment of the optional Internet firewall should follow standard Internet edge design guidance as documented in the Data Center Internet Edge SRND:

- http://www.cisco.com/application/pdf/en/us/guest/netsol/ns304/c649/ccmigration_09186a008014ee4e.pdf

Throughout this design guide, scenario A.1 is used to illustrate the relevant design and deployment considerations.

Unprotected Services

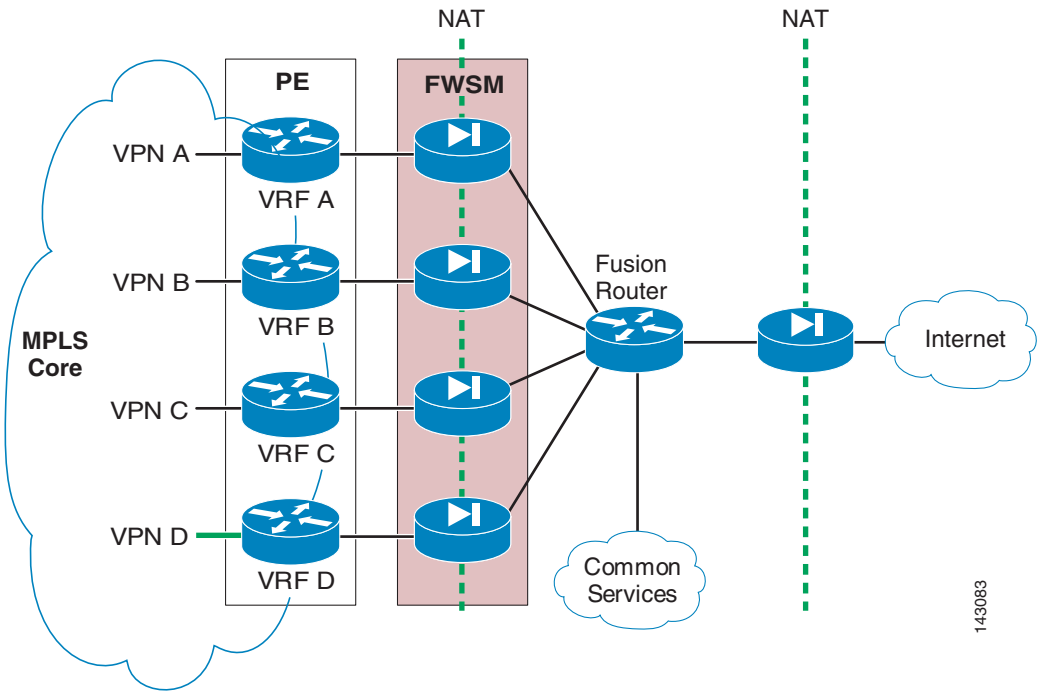
Unlike circuit-based technologies such as ATM or Frame Relay, the IP nature of MPLS VPNs allows enough flexibility for traffic to be leaked between VPNs in a controlled manner by importing and exporting routes between VPNs to provide IP connectivity between the VPNs. Thus, the exchange of traffic between the VPNs may happen within the IP core and does not have to pass through the VPN perimeter firewalls at the central site. This type of inter-VPN connectivity can be used to provide services that do not need to be protected by the central site firewall or that represent an unnecessary burden to the VPN perimeter firewalls. Because of the any-to-any nature of an IP cloud, there is very little chance of controlling inter-VPN traffic after the routes have been exchanged. These are referred to as “unprotected services.” This type of connectivity must be deployed very carefully because it can potentially create unwanted backdoors between VPNs and break the concept of the VPN as a “security zone” protected by a robust VPN perimeter front end. You must also consider the fact that importing and exporting routes between VPNs precludes the use of overlapping address spaces between the VPNs.

Note Although these services are not protected by the VPN perimeter firewalls, the IP segment to which they belong can potentially be head-ended by a firewall and therefore “protected.” However this creates routing and policy management challenges.

Firewalling for Common Services

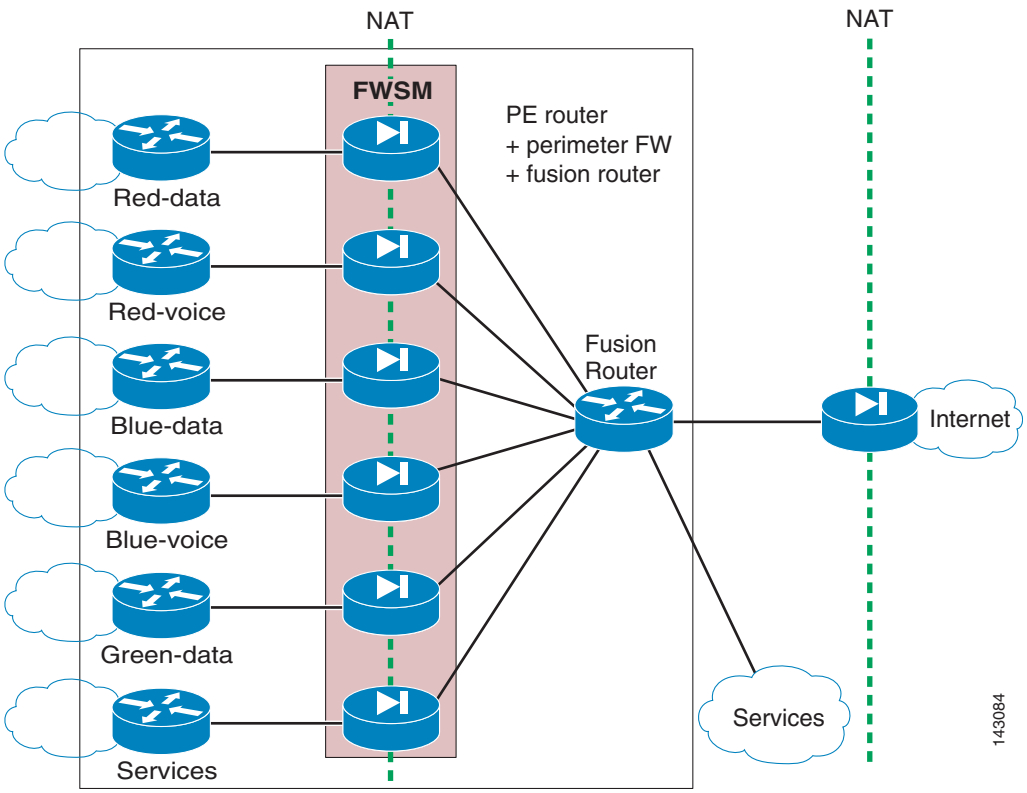
As VPNs proliferate, head-ending each VPN onto its own firewall can become both expensive and hard to manage. Cisco firewalls can be virtualized and thus offer a separate context for each VPN on the same physical appliance. The resulting topology is depicted in Figure 4-7. What has changed here is that a single physical firewall now provides a dedicated logical firewall to each VPN.

Figure 4-7 Virtual Firewall Contexts



The concept of virtual firewalls or firewall contexts has been implemented in the integrated Firewall Service Module (FWSM) for the Cisco Catalyst 6500. The integration of the firewall functionality onto the PE platform allows the topology shown in Figure 4-7 to be consolidated onto a single physical device, as shown in Figure 4-8.

Figure 4-8 Single Box Implementation of the VPN Perimeter Gateway



The logical topology remains unchanged: the firewall functionality is carried out by an FWSM within the PE and the fusion router is implemented by the creation of a VRF inside the same PE. Note that the "fusion VRF" does not connect to the MPLS cloud directly and acts as a separate router, with certain limitations that are explored in a subsequent section.

A single box perimeter implementation is feasible when there is a single common services/Internet site. However, when there is more than one services site and both resiliency and load distribution are desired among those sites, it is necessary to move the fusion VRF outside the PE router and to use a separate physical fusion router. The topologies and necessary routing configuration for single and multiple service site support are discussed in Common Services.

Network Address Translation—NAT

When operating in routed mode, a firewall establishes a connection between the inside and the outside for each flow that traverses the firewall. These connections are in the form of NAT entries, regardless of whether address translation is configured on the firewall.

The default behavior of firewalls is to allow the establishment of flows that are initiated from the inside network. Provided that the access lists allow it, upstream traffic flows through the firewall without a problem. However, a valid NAT entry in the connection table is required for the firewall to allow return traffic through. This NAT entry is dynamically created when the flow is initiated from the inside; connections initiated from the outside do not dynamically create an entry in the firewall.

This unidirectional mechanism prevents connections from being initiated from the outside of the network. For a connection to be successfully initiated from the outside of the network, a NAT entry for the internal destination address must exist in the firewall table before the connection can be established. Thus, if connections initiated from the outside network are required, static NAT entries must be created

to make the specific prefixes available to the outside of the firewall. To allow outside initiated connections, the creation of a static NAT entry is necessary even if the firewall is configured to not translate addresses (nat 0).

Benefits of NAT

The many benefits of being able to translate addresses include:

- Internal networks are hidden from the outside world. With NAT, it is not necessary for the Internet to be aware of the internal addressing scheme of the enterprise to be accessed. This provides an added layer of security.
- Internal networks can use private address spaces as defined in RFC 1918. This is particularly useful when deploying VPNs because this can accelerate the depletion of the IP address space available to the enterprise. This requires restricting extra-VPN communication through the VPN perimeter where addresses can be determined through the use of NAT; that is, inter-VPN route leaking does not work if there are any address overlaps between the private spaces employed.

Dynamic NAT

Address translation can be done dynamically. When an inside station attempts to connect to the outside of the firewall, a dynamic mapping of the source address of the inside station to a globally significant address (outside) is made. The globally significant address to be used is defined by a configured address pool. Thus, each connection is identified by a unique NAT entry. There is the potential for the number of connections to exceed the number of addresses available in the translation pool, in which case any new connection is not successful.

An alternative to regular NAT is Port Address Translation (PAT). With PAT, it is possible to use a single IP address for the global pool. Multiple connections can be associated to the same IP address and are uniquely identified by a unique Layer 4 port number. Hence a single global address can accommodate thousands of connections.

Static NAT

When internal resources must be made available outside the firewall, it is necessary to provide a predictable presence for the internal resource on the outside of the firewall.

By default, all addresses internal to the firewall are not visible to the outside. When addresses are not being translated, they might be visible to the outside but they are still not reachable because reachability from the outside requires an entry in the firewall connection table to be present ahead of time.

Static NAT assigns a globally significant address to the internal resource and also adds an entry to the firewall connection table. This address is fixed so that it can be reached from the outside in a consistent manner. The entry in the connection table makes it possible for the outside to connect to the inside resource provided that the necessary policy is in place.

Common Services

Single Common Services—Internet Edge Site

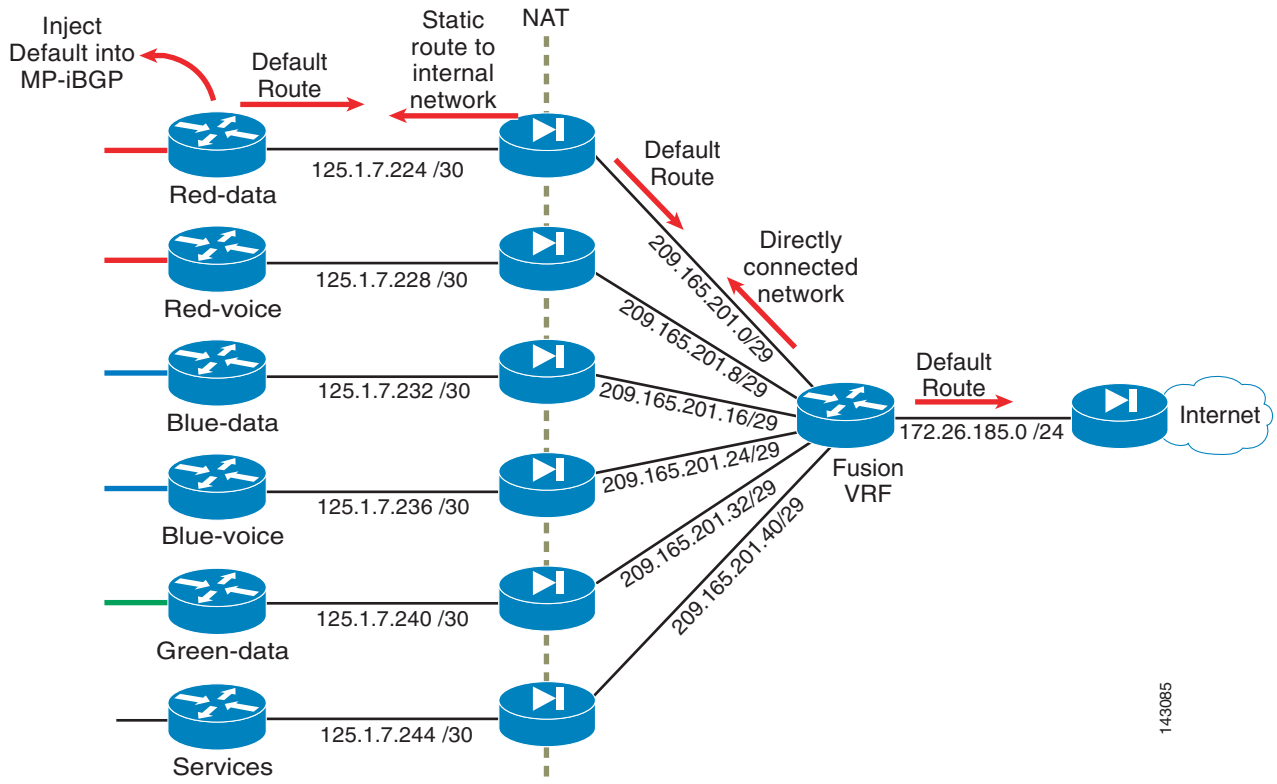
The routing between the fusion router, the different contexts, and VPNs must be configured with care.

Because of its place in the topology, the fusion router has the potential to mix the routes from the different VPNs when exchanging routes dynamically with the different VPNs. However, because the firewall in routed mode supports only static routing when configured for multiple contexts, the mixing

of VPN routes is not a concern. Connectivity between VPNs is achieved by the sole configuration of the fusion router; however the firewalls are configured to allow “established” connections only, which means only connections initiated from the inside of the firewall. Hence all VPNs can reach the fusion router and the fusion router can return traffic to all the VPNs. However, the VPNs are not able to communicate with each other through the fusion router unless very specific policies are set on the different firewall contexts to allow inter-VPN communication through the VPN perimeter gateway.

The static routing configuration for the perimeter gateway is shown in Figure 4-9.

Figure 4-9 Routing Considerations at the VPN Perimeter



The following steps configure static routing for the perimeter gateway. Detail is provided for only one VPN; other VPNs require similar configuration.

Step 1 Create a default route for the internal VRF (red-data):

```
7600-DC1-SS1(config)# ip route vrf red-data 0.0.0.0 0.0.0.0 125.1.7.226
```

Step 2 Create a static route for the inside of the firewall to reach the internal network (red-data VPN):

```
np-fwsm/red-data(config)# route inside 125.1.0.0 255.255.0.0 125.1.7.225 1
```

Step 3 Create a static default route for the outside of the firewall to send traffic to the fusion router/VRF:

```
np-fwsm/red-data(config)# route outside 0.0.0.0 0.0.0.0 209.165.201.2 1
```

Note The fusion router is able to reach the outside prefixes because they are directly connected, so no configuration is required.

Step 4 Create a static default route for the fusion router/VRF to communicate with the ISP. This is the standard configuration of an Internet access router and is not covered in this document.

```
7200-IGATE-DC1(config)# ip route vrf fusion 0.0.0.0 0.0.0.0 172.26.185.1
```

Step 5 Inject the default route created in Step 1 into MP-iBGP:

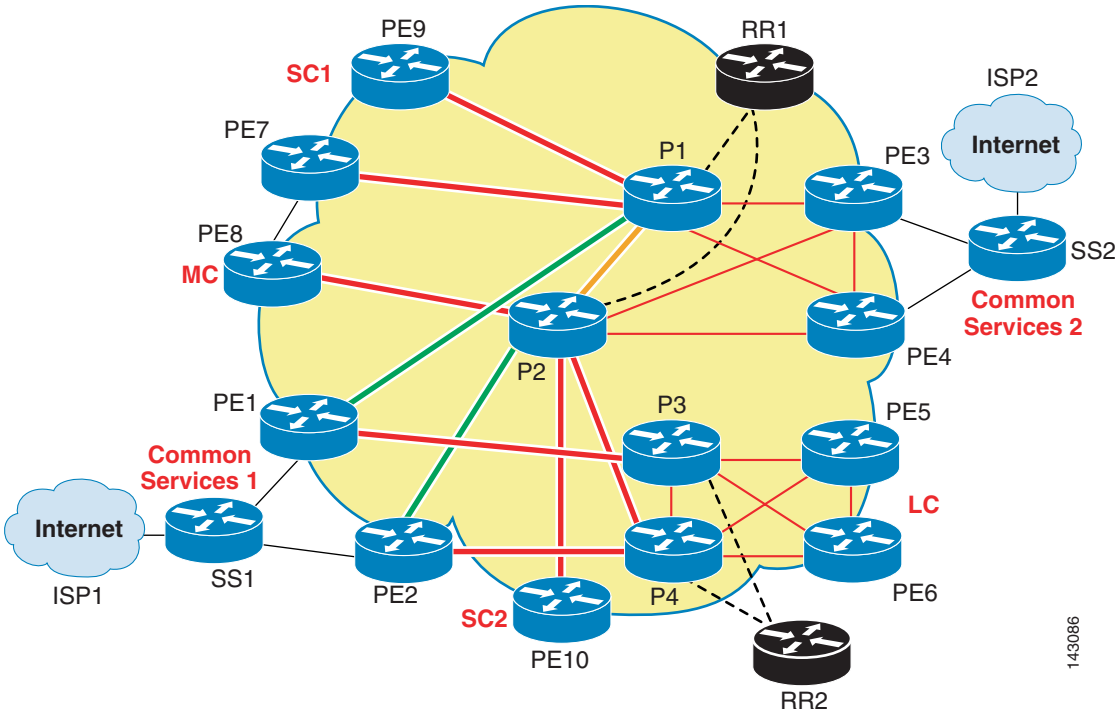
```
7600-DC1-SS1(config)#router bgp 1
7600-DC1-SS1(config-router)#address-family ipv4 vrf red-data
7600-DC1-SS1(config-router-af)#redistribute static
7600-DC1-SS1(config-router-af)#default-information originate

address-family ipv4 vrf red-data
 redistribute connected
 redistribute static
 default-information originate
 no auto-summary
 no synchronization
 exit-address-family
```

Multiple Common Services—Internet Edge Sites

Multiple sites are usually deployed for access to the Internet, so this section focuses on Internet access. However, the same principles apply for other shared services if these are accessible over a common network. When using multiple access points to the Internet (or the common services area), resiliency and load balancing are among the main goals.

In the proposed solution, two common services sites inject a default route into the MPLS MAN. As the default routes are received at the different PEs, the preferred route is chosen by the PE based on its proximity to the common services sites. This proximity is determined based on the core IGP metric (all other BGP attributes should be equal between the two advertised default routes). In the particular case of Internet access, some sites use the first Internet edge site, while others use the second. This achieves site-based load balancing and minimizes the use of the internal MAN links by choosing the closest Internet gateway to send traffic to the Internet in the most efficient manner (see Figure 4-10).

Figure 4-10 Both Internet Edge Sites and IGP Proximity

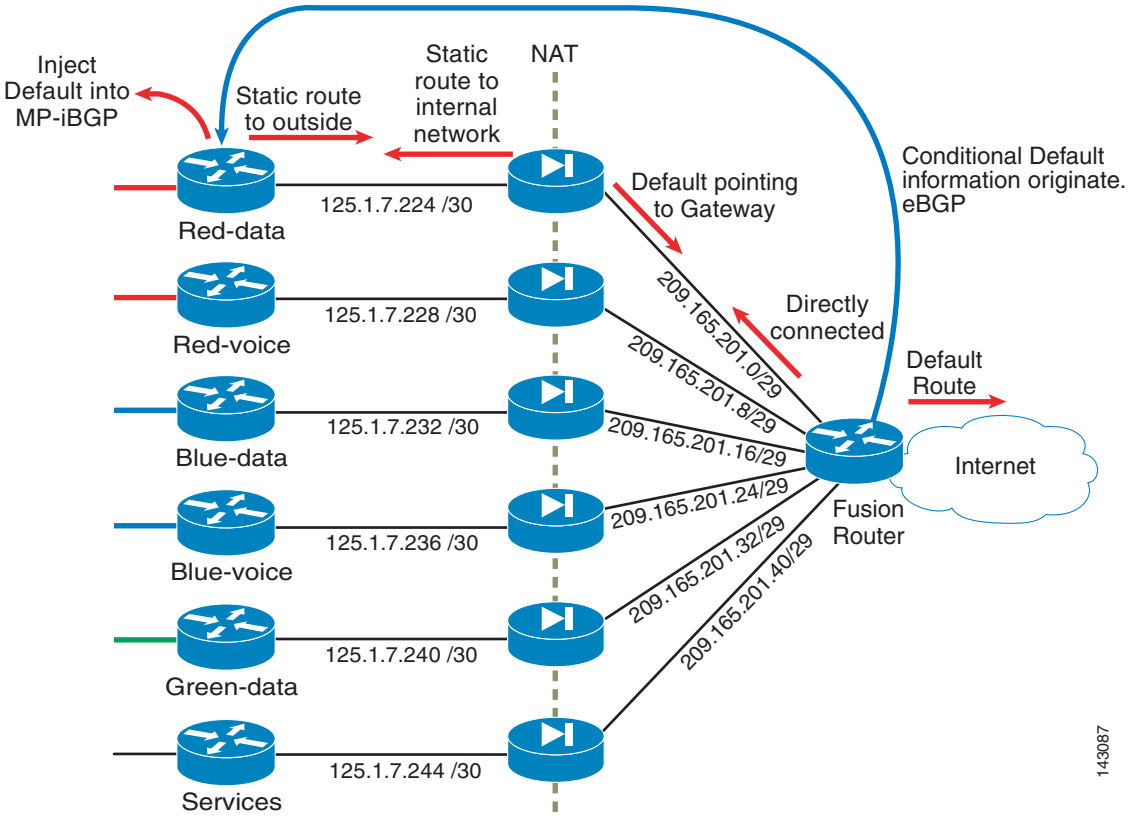


Internet Edge Site Considerations

In the case where one Internet edge site fails, all Internet traffic should be re-routed to the live Internet site. For failures within the MAN, this failover is provided by the reconvergence of the core IGP and the overlaid MP-iBGP. However because static default routes are being injected into the MAN, an ISP failure remains undetected and traffic is black-holed unless there is a dynamic mechanism to report this failure and to trigger a routing re-convergence to use the second Internet edge site. To do this, a dynamic routing protocol can be used to conditionally inject the default routes into the MPLS MAN. Hence a default route is originated and injected into the MPLS MAN from the Internet edge router only if this route is valid; that is, it exists in the edge router table (see Step 9 below).

To achieve this dynamic notification over the perimeter firewalls, eBGP is required to establish a connection across the firewall contexts (contexts do not support dynamic routing protocols). Because eBGP peering is required and this cannot be established between VRFs in a single box (the router ID would be the same for both VRFs, which would prevent the BGP adjacency from being established), a separate physical router is required for the fusion role (see Figure 4-11).

Figure 4-11 EBGP Peering for Dynamic Notification



The following steps must be completed to achieve the necessary BGP peering and inject the default routes conditionally:

Step 1 On the internal VRF, create a static route to the outside firewall subnet (209.165.201.0 /29):

```
7200-IGATE-DC1(config)#ip route vrf red-data 209.165.201.0 255.255.255.248 125.1.7.226
```

Step 2 On the inside firewall interface, create a static route to the internal VPN summary prefix:

```
np-fwsm/red-data(config)#route inside 125.1.0.0 255.255.0.0 125.1.7.225 1
```

Step 3 On the outside firewall interface, create a static default route to the Internet gateway:

```
np-fwsm/red-data(config)#route outside 0.0.0.0 0.0.0.0 209.165.201.2 1
```

Note The fusion router is directly connected to the outside firewall networks. No configuration is required

Step 4 On the fusion router, create a default route pointing at the Internet gateway (172.26.185.1 /32):

```
7200-IGATE-DC1(config)#ip route 0.0.0.0 0.0.0.0 172.26.185.1
```

Step 5 Configure static NAT entries for the internal VRF BGP peering address. These are necessary to establish the bi-directional TCP sessions for BGP peering. For any type of communication to be initiated from the outside of the firewall, a static NAT entry is required by the firewall; otherwise the connection is rejected.

```
static (inside,outside) 209.165.201.3 125.1.7.225 netmask 255.255.255.255 norandomseq
```

Step 6 Open the necessary firewall policies to permit BGP peering over the firewall:

```
access-list allow_any extended permit ip any any log debugging !Allows sessions initiated
from the inside of the firewall (i.e. the VPN).
access-list allow_any extended permit tcp host 125.1.7.225 eq bgp host 209.165.201.2 eq
bgp
access-list allow_bgp extended permit tcp host 209.165.201.2 eq bgp host 209.165.201.3 eq
bgp
!
access-group allow_any in interface inside
access-group allow_bgp in interface outside
```

Step 7 Configure the internal VRFs and the fusion router as BGP neighbors:

```
!Fusion Router!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
router bgp 10
 no synchronization
 bgp log-neighbor-changes
 redistribute static
 neighbor 209.165.201.3 remote-as 1
 neighbor 209.165.201.3 ebgp-multihop 255
 !
!PE router: Red-data VRF!!!!!!!!!!!!!!!!!!!!!!!!!
router bgp 1
 no synchronization
 bgp log-neighbor-changes
 neighbor 125.1.125.15 remote-as 1
 neighbor 125.1.125.15 update-source Loopback0
 neighbor 125.1.125.16 remote-as 1
 neighbor 125.1.125.16 update-source Loopback0
 neighbor 209.165.201.2 remote-as 10
 no auto-summary
 !
 address-family vpnv4
 neighbor 125.1.125.15 activate
 neighbor 125.1.125.15 send-community extended
 neighbor 125.1.125.16 activate
 neighbor 125.1.125.16 send-community extended
 exit-address-family
 !
 address-family ipv4 vrf red-data
 redistribute connected
 redistribute static
 neighbor 209.165.201.2 remote-as 10
 neighbor 209.165.201.2 ebgp-multihop 255
 neighbor 209.165.201.2 activate
 maximum-paths eibgp 2
 no auto-summary
 no synchronization
 exit-address-family
 !
```

Step 8 Originate a default route at the fusion router and send it over BGP to the internal VRFs. Use conditional statements so that the default route is advertised only if it is present in the local routing table (that is, if the Internet service is available).

```
router bgp 10
 no synchronization
 bgp log-neighbor-changes
 redistribute static
 neighbor 209.165.201.3 remote-as 1
 neighbor 209.165.201.3 ebgp-multihop 255
 neighbor 209.165.201.3 default-originate route-map SEND_DEFAULT
 neighbor 209.165.201.3 distribute-list 3 in
```

```
 no auto-summary
!
ip classless
ip route 0.0.0.0 0.0.0.0 172.26.185.1
no ip http server
!
!
access-list 1 permit 0.0.0.0
access-list 2 permit 172.26.185.1
access-list 3 deny   any
!
route-map SEND_DEFAULT permit 10
 match ip address 1
 match ip next-hop 2
 set metric 0
 set local-preference 100
```

Step 9 Prevent any BGP updates from the inside network coming onto the fusion router. If the fusion router is allowed to receive VPN routes via e-BGP, it replicates the received routes onto its other e-BGP peers. This would basically inject routes from one VPN into another, so these updates must be prevented.

```
router bgp 10
 no synchronization
 bgp log-neighbor-changes
 redistribute static
 neighbor 209.165.201.3 remote-as 1
 neighbor 209.165.201.3 ebgp-multihop 255
 neighbor 209.165.201.3 default-originate route-map SEND_DEFAULT
 neighbor 209.165.201.3 distribute-list 3 in
 no auto-summary
!
ip classless
ip route 0.0.0.0 0.0.0.0 172.26.185.1
no ip http server
!
!
access-list 1 permit 0.0.0.0
access-list 2 permit 172.26.185.1
access-list 3 deny   any
!
```

Routing Considerations

Advertising Multiple Routes into MP-iBGP

Advertising more than one default route or advertising multiple routes for the same prefix must be done with care. The default behavior of a route reflector is to make a decision based on metrics and attributes and to reflect only the best one of the advertised routes. The result is that all PEs always receive the route that is best for the route reflector, which is not necessarily the best route for the PE to reach the Internet.

To achieve load balancing and redundancy from injecting multiple routes for a common destination in this topology, it is important that the route reflector actually "reflects" all the routes it receives so that the route selection can actually be done at the PEs. To achieve this, the routes must be advertised with different RDs. For example, the default route advertised by Common Services Site 1 (SS1) is sent with an RD of 10:103, while the default route sent by Common Services Site 2 (SS2) is sent with an RD of 101:103. In this manner, some sites prefer SS2 while others prefer SS1.

Load balancing across the MAN core can be achieved by instructing BGP to install multiple paths in the routing table (**ibgp multipath**). Although it is tempting to use unequal cost paths and to load balance across all possible paths, this may affect the way traffic to the Internet is handled and may cause the use of suboptimal paths to access the Internet. In the proposed scenario, the requirement is for certain portions of the network to prefer on Common Services Site over another. Thus the load balancing is done per site rather than per flow. For example, site SC1 always tries to use SS1 first because it is the closest Internet access site. If unequal paths are allowed to be installed in the routing table, SC1 sends some flows over SS1 and others over SS2, potentially congesting low speed links that would not have been used if only one path had been installed on the routing table.

However, the solution is not to turn **bgp multipath** off, but to set the bgp multipath capability to install only multiple equal cost paths. This is important because equal cost load balancing is desirable between sites. Because only equal cost paths can be installed in the table, the Internet is accessed consistently via either SS1 or SS2, depending on the proximity of the site. If a failure is detected, the routing protocols must determine which sites are still available and recalculate the paths to the Common Services Sites to make a decision on where to exit the Internet.

Asymmetric Return Paths

This is a classic problem faced when multi-homing to the Internet, in which traffic exits the enterprise out of one gateway and the return traffic is received over a different gateway. The implications are many, but the main one is that the return traffic is normally not able to get through the firewall; no session has been established at the return firewall because the traffic originally left the network through a different firewall.

In the proposed scenario, the asymmetry of the return path is handled by using different global NAT address pools outside the different Internet gateways. Each Internet gateway advertises a unique address pool, thus eliminating any ambiguity in the return path. For example, the source address of traffic leaving SS1 is rewritten to a prefix advertised only by SS1. Therefore the return traffic for a stream that entered the Internet through SS1 must be through SS1 because the Internet has routes to the SS1 address pool only through SS1.

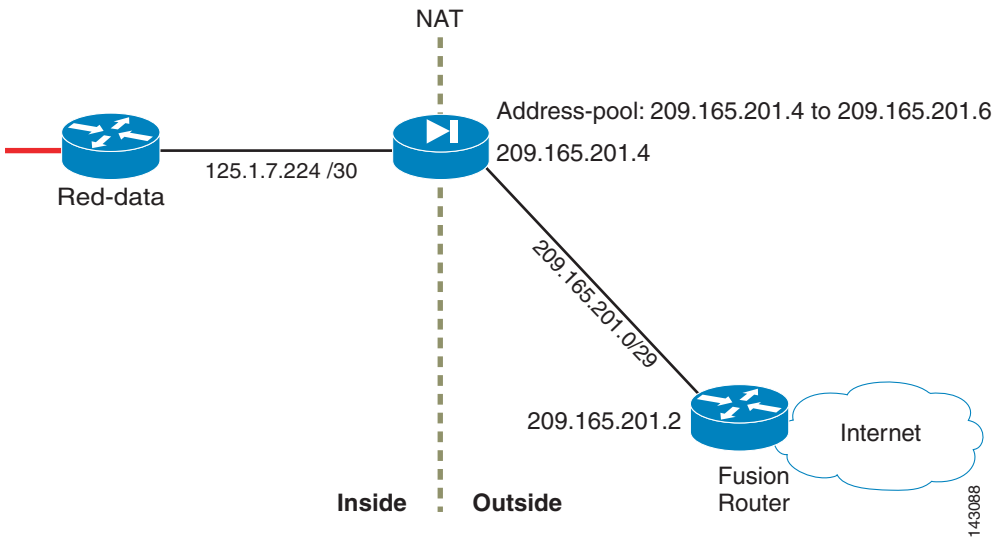
NAT in the MPLS MAN

A combination of dynamic and static NAT is required at the VPN perimeter:

- Dynamic NAT is used to allow connectivity for sessions established from the inside of the network.
- Static NAT is required to allow the following:
  - BGP peer establishment
  - Connectivity to resources shared from inside a service VPN

Dynamic NAT can be established by using either NAT or PAT. When using NAT, it is necessary to provide the outside interface of the firewall with an IP prefix that can accommodate the entire global address pool to be used in the translation. Figure 4-12 shows the scenario for the Red-data VPN in the Cisco test bed.

Figure 4-12 Red-data VPN Test Bed Scenario



Any connection from the Red-data VPN to the Internet creates one NAT entry and therefore uses one of the addresses in the address pool. Thus, the number of possible concurrent connections is limited to three in this specific scenario. Note that a 29-bit address mask (rather than 32 bits) has been used for the point-to-point connection to accommodate the NAT address pool.

The following commands configure the Red-data firewall context to allow this kind of connectivity:

```
! Create the dynamic NAT Pool
global (outside) 1 209.165.201.4-209.165.201.6 netmask 255.255.255.248
nat (inside) 1 125.1.0.0 255.255.0.0
```

The following commands allow outbound connectivity:

```
!Allow sessions initiated from the inside of the firewall (i.e. the VPN).
access-list allow_any extended permit ip any any log debugging
access-group allow_any in interface inside
```

Alternatively, PAT can provide dynamic translation without the limitation of the exhaustion of the global address pool. Configuring PAT is almost identical to configuring NAT, except that instead of defining a global range, a single IP is configured:

```
! Create the dynamic PAT Pool
np-fwsm/red-data(config)# nat (inside) 1 125.1.0.0 255.255.0.0
np-fwsm/red-data(config)# global (outside) 1 209.165.201.4
Global 209.165.201.4 will be Port Address Translated
```

A static NAT entry is required to allow BGP peering between the fusion router and the internal VRF as described in Firewalling for Common Services.

The necessary access lists must be configured to allow this type of connectivity as well. Care must be taken to open the firewall exclusively to the relevant BGP traffic, as shown in the following configuration:

```
! Create the static translation for the inside (125.1.7.225) peer
static (inside,outside) 209.165.201.3 125.1.7.225 netmask 255.255.255.255 norandomseq
! Allow bgp tcp session between the neighbors only and in both directions
access-list allow_any extended permit tcp host 125.1.7.225 eq bgp host 209.165.201.2 eq
bgp
access-list allow_bgp extended permit tcp host 209.165.201.2 eq bgp host 209.165.201.3 eq
bgp
! Apply policies in both directions
```

```
access-group allow_any in interface inside
access-group allow_bgp in interface outside
```

Other static NAT entries may be required if there are servers inside the VPN that are made available outside the VPN. As the number of servers to publish increases, the use of static PAT may be useful. The use of static PAT is beyond the scope of this document; for information on static PAT as well as more details on NAT in general, see the FWSM configuration guide at: http://www.cisco.com/univercd/cc/td/doc/product/lan/cat6000/mod_icn/fwsm/fwsm_2_2/fwsm_cfg/index.htm

Convergence

Redundancy within the MPLS network—power supplies, links, routers, etc.—provides protection against any such losses. But more and more applications require faster convergence which from a network perspective involves:

- Detecting the failure
- Finding an alternate resource or restoration
- Propagation of the change to the rest of the network, if required

The commonly used mechanisms in IP-environments dictates that an IGP extended for Fast Convergence together with convergence enhancements for BGP provides the overall protection at restoration function. Traditional MPLS Protection and Restoration mechanisms, such as Traffic Engineering Fast ReRoute (TE FRR), provide capabilities to circumvent node or link failures.

- **Link failure detection**—Various mechanisms are in place to provide a fast detection of link failure, both generic and media dependent. The fastest mechanism by far is the integrated OAM mechanism of SONET/SDH framing. Other mechanisms include Loss of optical Signal (LOS), PPP keepalives, and various LMI mechanisms. Bidirectional Forwarding Detection (BFD) is a generic lightweight hello-based mechanism that can be used in conjunction with any type of media.
- **Failure propagation**—Depending on the Protection and Restoration mechanism being used, there may not be an associated propagation delay before the backup for a failed facility is installed. This is the case with TE FRR. If an IGP or BGP is used then the updated network information has to be flooded throughout the network.

Additional processing time is required for the IGP to compute a new network view by performing an SPF operation. Once that operation is completed, updated routing information is installed in the RIB. In an MPLS network protected by TE FRR, this operation still takes place, but the service restoration is not dependent on its completion. In the case of BGP, an update or bestpath operation has to be performed and the time this operation takes is a direct consequence of the BGP table size. After the RIB has been updated the associated FIB also has to be updated so that the forwarding plane can make use of the updated information.

Traffic Engineering Fast ReRoute (TE FRR)

FRR is a mechanism for protecting MPLS TE LSPs from link and node failures by locally repairing the LSPs at the point of failure, allowing data to continue to flow on them while their headend routers attempt to establish new end-to-end LSPs to replace them. FRR locally repairs the protected LSPs by rerouting them over backup tunnels that bypass failed links or nodes.

- **Link Protection**—Backup tunnels that bypass only a single link of the LSP's path provide link protection. They protect LSPs if a link along their path fails by rerouting the LSP's traffic to the next hop (bypassing the failed link). These are referred to as next-hop (NHOP) backup tunnels because they terminate at the LSP's next hop beyond the point of failure.
- **Node Protection**—Backup tunnels that bypass next-hop nodes along LSP paths are called next-next-hop (NNHOP) backup tunnels because they terminate at the node following the next-hop node of the LSP paths, thereby bypassing the next-hop node. They protect LSPs if a node along their path fails by enabling the node upstream of the failure to reroute the LSPs and their traffic around the failed node to the next-next hop.
- **RSVP Hellos**—RSVP Hello enables RSVP nodes to detect when a neighboring node is not reachable. This provides node-to-node failure detection. When such a failure is detected, it is handled in a similar manner as a link-layer communication failure.

RSVP Hello can be used by FRR when notification of link-layer failures is not available (for example, with Ethernet) or when the failure detection mechanisms provided by the link layer are not sufficient for the timely detection of node failures.

A node running Hello sends a Hello Request to a neighboring node every interval. If the receiving node is running Hello, it responds with Hello Ack. If four intervals pass and the sending node has not received an Ack or it receives a bad message, the sending node declares that the neighbor is down and notifies FRR. There are two configurable parameters:

- Hello interval—Use the **ip rsvp signalling hello refresh interval** command.
- Number of acknowledgment messages that are missed before the sending node declares that the neighbor is down—Use the **ip rsvp signalling hello refresh misses** command.

Fast Reroute Activation

There are two mechanisms that cause routers to switch LSPs onto their backup tunnels:

- Interface down notification
- RSVP Hello neighbor down notification

When a router's link or neighboring node fails, the router often detects this failure by an interface down notification. For example, on a POS interface this notification is very fast. When a router notices that an interface has gone down, it switches LPSs going out that interface onto their respective backup tunnels.

RSVP Hellos can also be used to trigger FRR. If RSVP Hellos are configured on an interface, messages are periodically sent to the neighboring router. If no response is received, Hellos declare that the neighbor is down. This causes any LSPs going out that interface to be switched to their respective backup tunnels.

An additional mechanism that will be available in the future would be BFD. BFD is a detection protocol that is designed to provide fast forwarding path failure detection times for all media types, encapsulations, topologies, and routing protocols. In addition to fast forwarding path failure detection, BFD provides a consistent failure detection method for network administrators. Because the network administrator can use BFD to detect forwarding path failures at a uniform rate, rather than the variable rates for different routing protocol hello mechanisms, network profiling and planning will be easier, and reconvergence time will be consistent and predictable.

As long as each BFD peer receives a BFD control packet within the detect-timer period, the BFD session remains up and any routing protocol associated with BFD maintains its adjacencies. If a BFD peer does not receive a control packet within the detect interval, it informs any clients of that BFD session about the failure.

Note As of August, 2006, BFD support for FRR triggering is planned to 12.0(33)S for GSR and the "cobra" release for 7600.

Backup Tunnel Selection Procedure

When an LSP is signaled, each node along the LSP path that provides FRR protection for the LSP selects a backup tunnel for the LSP to use if either of the following events occurs:

- The link to the next hop fails.
- The next hop fails.

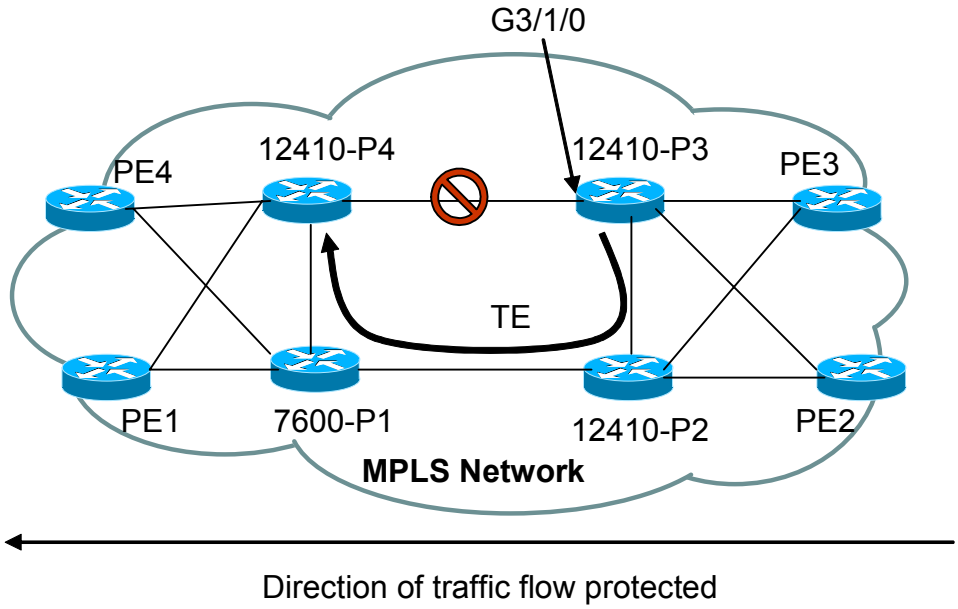
By having the node select the backup tunnel for an LSP before a failure occurs, the LSP can be rerouted onto the backup tunnel quickly if there is a failure.

For an LSP to be mapped to a backup tunnel, all of the following conditions must exist:

- The LSP is protected by FRR; that is, the LSP is configured with the **tunnel mpls traffic-eng fast-reroute** command.
- The backup tunnel is up.
- The backup tunnel is configured to have an IP address, typically a loopback address.
- The backup tunnel is configured to protect this LSP's outgoing interface; that is, the interface is configured with the **mpls traffic-eng backup-path** command.
- The backup tunnel does not traverse the LSP's protected interface.
- The backup tunnel terminates at the LSP's NHOP or NNHOP. If it is an NNHOP tunnel, it does not traverse the LSP's NHOP.
- The bandwidth protection requirements and constraints, if any, for the LSP and backup tunnel are met.

Protecting the Core Links

Figure 4-13 Protecting the Core Links



In Figure 4-13, the link between P3 and P4 is protected (for traffic going from P3 to P4) by a explicitly configured TE tunnel P3-P2-P1-P4. Since the tunnels are unidirectional, a reverse path would need to be created for traffic going from P4 to P3. All the links are POS and rely on underlying SONET alarms for link failure detection. Explicit tunnels can be setup to provide fast re-route capabilities in case of core failures.

```
interface Tunnel11
 ip unnumbered Loopback0
 no ip directed-broadcast
 tunnel destination 100.0.250.14
 tunnel mode mpls traffic-eng
 tunnel mpls traffic-eng priority 0 0
 tunnel mpls traffic-eng bandwidth 10000
 tunnel mpls traffic-eng path-option 5 explicit name backup_of_10
!
interface POS3/1/0
 ip address 100.0.4.1 255.255.255.0
 no ip directed-broadcast
 ip pim sparse-mode
 no keepalive
 mpls traffic-eng tunnels
 mpls traffic-eng backup-path Tunnel11
 tag-switching ip
 crc 32
 clock source internal
 pos ais-shut
 pos report lrdi
 ip rsvp bandwidth 1866000 1866000
!
ip explicit-path name backup_of_10 enable
 next-address 100.0.3.1<P2>
 next-address 100.0.2.11<P1>
 next-address 100.0.1.2<P4>
```

Performance

Failure detection plays an important role in determining the switchover performance. Depending on the platform/linecard/IOS, different detection mechanisms can be deployed, such as SONET alarms for POS, RSVP hellos for other interface types, and potentially BFD in the future. Another factor is how quickly the database get updated once the failure has been signaled, which is dependent on the number of routes being protected.

In the above example with 2000 IGP routes in the network, the failover testing was done to compare the responses with and without TE/FRR. A single stream of traffic was observed end-to-end. When the POS link between P3-P4 was shutdown, the packet loss was measured and the downtime was calculated based on that. It was found that with FRR a POS failure created a failure of 4-5s, but with FRR this was measured to be less then 10ms.