>THIS IS **THE WAY**

>THIS IS NORTEL

Nortel Multiservice Switch 7400/15000/20000

IP VPN Technology Fundamentals

NN10600-802

Document status: Standard
Document issue: 7.2S1
Document date: March 2006
Product release: PCR7.2 and up
Job function: Product Fundamentals
Type: NTP
Language type: U.S. English

Contents

What's new

The following feature was added to this document:

- Secure shell (Ssh) on Nortel Multiservice Switch 15000 and 20000 (page 7)

Other changes to this document include the following:

Attention: To ensure that you are using the most current version of an NTP, check the current NTP list in NN10600-000 *Nortel Multiservice Switch 7400/15000/20000 What's New*.

- This document was renumbered from NN10600-581 to NN10600-802 as part of an ongoing effort to improve the organization of Nortel Multiservice Switch documentation.
- Replaced obsolete request for comments (RFC) references with current RFC references.

Secure shell (Ssh) on Nortel Multiservice Switch 15000 and 20000

The following section was updated for this feature:

- Management VR (page 16)

Virtual private networking conceptual overview

With the exponential growth of the Internet, carriers face increasing demands from their enterprise customers for IP-based services that provide facilities equivalent to a private network. Nortel Multiservice Switch IP virtual private network (VPN) service allows carriers to provide site-to-site intranet connectivity to its enterprise customers, with greater flexibility and reduced costs.

Navigation

What is an IP VPN?

An IP VPN is a managed IP service offered by a carrier to an enterprise customer. The IP VPN service provides secure and reliable connectivity, management, and addressing (equivalent to that available on a private network) over a shared public network infrastructure.

See the following sections for more information:

IP VPN applications

IP VPNs enable the following types of applications:

- site-to-site intranet connectivity

  Site-to-site intranet VPNs provide scalable, secure connectivity among multiple enterprise sites over a public IP network.

- corporate extranet connectivity

  Extranet VPNs provide scalable, secure connectivity between an enterprise and its business partners.

- remote access

  Remote access VPNs provide an enterprise's remote employees with reliable access to the enterprise network.

- Internet access

  Internet access VPNs provide reliable connectivity to the Internet with Network Address Translation (NAT) for private to public IP address translation and firewalls for security.

Nortel Multiservice Switch IP VPN service supports site-to-site intranet connectivity between multiple enterprise sites.

IP VPN management

Either the enterprise customer or the carrier can implement IP VPNs.

VPNs implemented by the customer are based on equipment that the enterprise owns and operates. The customer premises equipment (CPE) uses standards-based tunnels over a public IP network as the basis of the VPN. The enterprise customer is responsible for all configuration and maintenance of the VPN. CPE-based VPNs can operate over the Internet or a carrier's IP network, and are often based on encryption and Network Address Translation (NAT).

VPNs implemented by the carrier are based on equipment that the carrier owns and operates. The equipment can be located on the customer's premises (for CLE- based VPNs) or at the carrier's point-of-presence (for POP-based VPNs). Both CLE-based and POP-based VPNs typically operate over the carrier's public IP network.

Nortel Multiservice Switch IP VPN service allows carriers to offer a POP-based VPN solution to their enterprise customers.

Why use Multiservice Switch IP VPN service?

Nortel Multiservice Switch IP VPN service is a standards-based solution that co-exists with legacy WAN services and meet carriers' requirements for addressing, forwarding mechanisms, routing information distribution, and quality of service.

Carriers can offer a separate IP VPN service to each of their enterprise customers. From a carrier perspective, each Multiservice Switch IP VPN appears as an independent routed network, and uses separate, independent virtual routers linked together through point-to-multipoint IP tunnels across a backbone.

With Multiservice Switch IP VPN service, data from each customer remains separate from all other traffic flows, guaranteeing a high degree of security within the carrier network. In addition, Multiservice Switch IP VPN service provides the following key benefits:

- Security (page 10)
- Reliability (page 10)
- Flexibility (page 10)
- Core independence (page 10)
- Scalability (page 11)

Security

Each IP VPN uses separate, independent virtual routers. With the help of IP tunneling encapsulation, data from each customer or each VPN always stays separate from other traffic flows. This traffic isolation provides a high degree of data privacy within the carrier network for customer sites in the same address domain.

Reliability

Nortel Multiservice Switch IP VPN service offers carriers the highest level of reliability. Every component in the Multiservice Switch node supports sparing, including control processors (CP) and fabrics. Fast switchovers to backups ensure minimal service impacts.

For high-speed optical lines, SONET/SDH one-for-one protection switching is available, and for electrical DS-3/E3 interfaces, one-for-n sparing is available. Carriers can spare all common equipment and replace all modules while the unit remains in service.

Flexibility

Carriers can tailor Nortel Multiservice Switch IP VPN services to meet a wide range of enterprise customer needs. Multiservice Switch systems uses DiffServ IP class of service, allowing carriers to offer their customers different classes of service for different types of traffic. The carrier can then map specific IP CoS requirements to ATM and frame relay QoS requirements as customer traffic traverses the ATM or frame relay backbone network.

Core independence

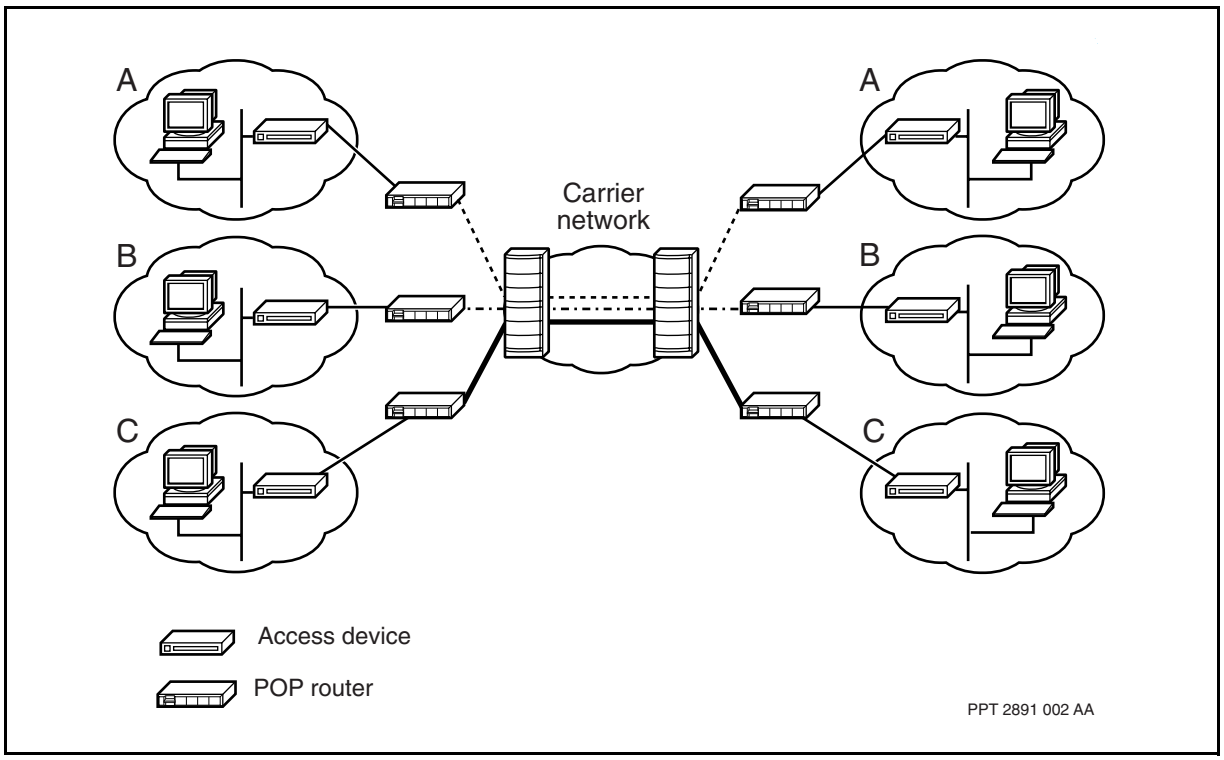
Nortel Multiservice Switch IP VPN service presents an IP interface to the subscriber, separating the VPN from the underlying link-layer technology (for example, frame relay or ATM).

Multiservice Switch IP VPN service operates independently of the core network infrastructure. Core independence allows carriers to implement IP VPN services in their existing frame relay or ATM infrastructures, and allows for smooth migration to future core technologies with minimal disruption to service.

Scalability

The traditional approach to POP-based IP VPNs involved leased lines from the customer site to separate POP routers for each customer network connection. The POP routers in turn connected to the WAN node through separate links. This meant using multiple ports and deploying more hardware, as well as PVC and SPVC meshing between backbone nodes. See the figure Traditional VPN configuration (page 11).

Traditional VPN configuration

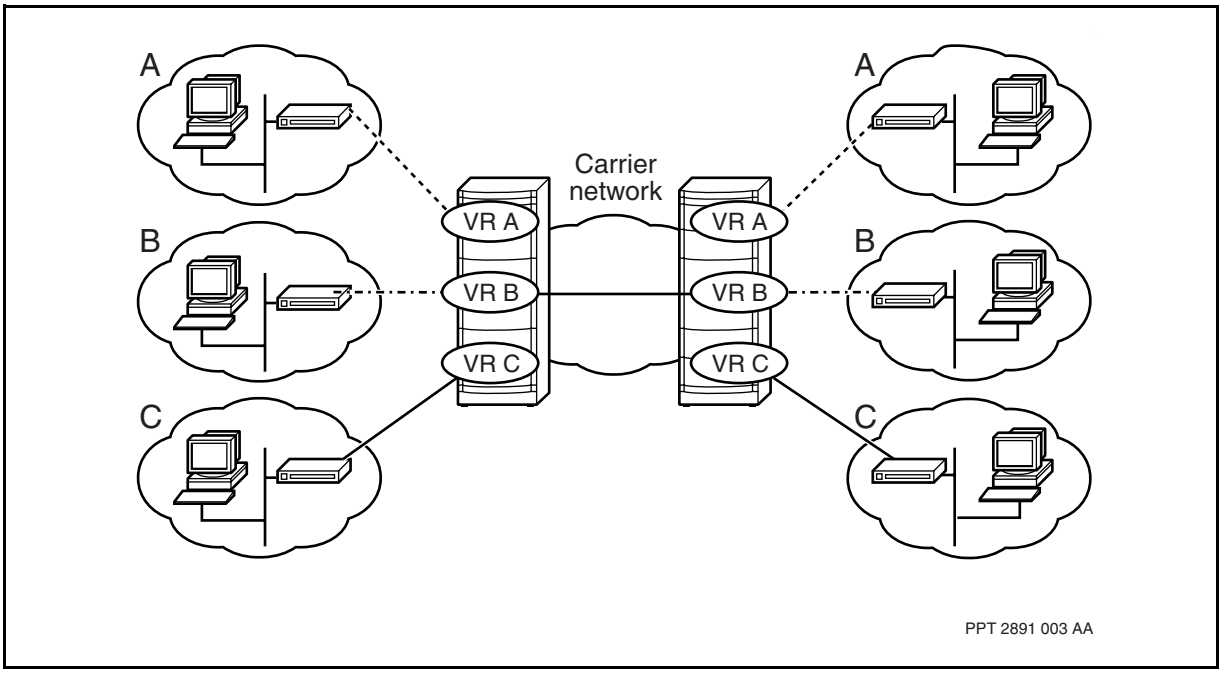


With Nortel Multiservice Switch IP VPN service, carriers can provide POP-based VPNs through virtual routers on a single node. CPE routers connect to the virtual routers on the carrier's Multiservice Switch node, eliminating the need for multiple POP routers. This architecture reduces the number of ports required, as well as the degree of PVC meshing in the backbone. See the figure Multiservice Switch IP VPN configuration (page 12).

Multiservice Switch supports multiple VRs on a single node. This high degree of scalability allows carriers to offer site-to-site intranet service to a large number of enterprise customers of all sizes. In addition, BGP-4 route reflectors reduce the need for BGP-4 peer meshing as carriers add new customer sites to an existing VPN.

Scaling of the backbone reduces layer 2 meshing and configuration. Only a single set of connections from each node is required, and the addition of new VPN sites does not impact the existing backbone configuration. In addition, the aggregation of all traffic over common backbone links simplifies the engineering of backbone connections.

Multiservice Switch IP VPN configuration



Internet Requests for Comments

The following Requests for Comments (RFCs) containing information related to IP are available from numerous sources, including Internet Network Information Center (NIC) servers:

- RFC 791, *Internet Protocol*
- RFC 793, *Transmission Control Protocol*
- RFC 950, *Internet Standard Subnetting Procedure*
- RFC 1745, *BGP4/IDRP for IP-OSFP Interaction*
- RFC 1771, *Border Gateway Protocol 4 (BGP-4*)
- RFC 1772, *Application of the Border Gateway Protocol in the Internet*

- RFC 2003, *IP Encapsulation within IP*
- RFC 2178, *OSPF Version 2*
- RFC 2453, *RIP Version 2*
- RFC 2858, *Multiprotocol extensions for BGP-4*

Multiservice Switch IP VPN architecture

Nortel Multiservice Switch IP virtual private network (VPN) service uses a combination of existing Multiservice Switch network routing capabilities and new IP tunneling functionality to allow a carrier to offer site-to-site intranet connectivity to its enterprise customers.

An IP VPN consists of multiple VPN sites, each representing a private customer network. Enterprise customers access the IP VPN service at each network site by connecting to a virtual router on the Multiservice Switch node over an access link. Multiservice Switch virtual routers and access devices on the customer premises exchange routing information for individual VPN sites. Routing tables associated with each virtual router define the site-to-site connectivity for that enterprise VPN.

Navigation

Access media

Nortel Multiservice Switch systems can provide customer access to the carrier network using the media listed in the table Multiservice Switch-supported access media for IP VPN (page 15).

Multiservice Switch-supported access media for IP VPN

Multiservice Switch	Access media
Multiservice Switch 7400	ATM, frame relay using IP-optimized DLCIs, frame relay using FRDTE, PPP (single-link or multi-link), 10BaseT Ethernet, 100BaseT Ethernet
Multiservice Switch 15000 and Multiservice Switch 20000	ATM, frame relay using IP-optimized DLCIs, frame relay using FRDTE, PPP (single-link)

The ATM multiprotocol encapsulation (MPE) interface is an access service that allows IP encapsulation over ATM, in accordance with RFC 2684. Use the ATM MPE service to transmit IP traffic to interconnected external routers and other Multiservice Switch virtual routers over an ATM network.

IP-optimized DLCIs and the Multiservice Switch frame relay data termination equipment (DTE) interface are access services that allows IP encapsulation over frame relay, in accordance with RFC 2427. IP-optimized DLCIs directly bind to a protocol port and are the recommended method for frame relay access to an IP VPN. Use the frame relay DTE service (in conjunction with a FR UNI and virtual framer) to transmit IP traffic to interconnected external routers and other Multiservice Switch VRs over a frame relay network.

Multiservice Switch point-to-point protocol (PPP) interface is an access service that provides a standard method for encapsulating IP packets over serial lines. PPP provides full-duplex simultaneous packet transfer between two dedicated WAN peers, and provides a standard method for routing IP packets over both low-speed asynchronous and high-speed synchronous link connections.

For more information on these access services, see NN10600-800 *Nortel Multiservice Switch 7400/15000/20000 IP Technology Fundamentals*.

Virtual routers

Nortel Multiservice Switch virtual routers (VR) provide a software emulation of physical routers. Through VRs, the Multiservice Switch node forwards packets to the correct destination, and isolates each customer's traffic by maintaining separate routing tables for each customer. From the carrier's perspective, each IP VPN consists of a set of VRs. In addition, a single customer VR can support multiple VPN customer sites.

Multiservice Switch IP VPN solution uses the following implementations of the VR for site-to-site intranet connectivity:

- Customer VR (page 16)

- Management VR (page 16)
- Virtual connection gateway (page 17)

For detailed information about VRs, see NN10600-800 *Nortel Multiservice Switch 7400/15000/20000 IP Technology Fundamentals*.

Customer VR

An IP VPN contains one or more customer VRs and can span multiple Nortel Multiservice Switch nodes. The carrier assigns a customer VR to every customer site that connects to its network. Thus, each customer VR on a Multiservice Switch node represents a separate VPN site.

Customer VRs provide separate routing functions for each enterprise customer network that connects to them. Since customer traffic is only processed by customer VRs dedicated to the enterprise customer who owns the VPN, separate routing capabilities guarantee traffic isolation from other customers while running on shared switching and transmission resources.

CPE access devices in the enterprise customer's network connect to the customer VR through access links; for more information, see Access media (page 14). To the CPE access device, the customer VR appears as a neighbor router in the customer's network, to which it sends all traffic for non-local VPN destinations.

Each CPE access device must learn the set of destinations reachable through its connection to the customer VR on the Multiservice Switch node; this can be as simple as a default route. The customer VRs within a single IP VPN are responsible for learning and disseminating reachability information among themselves.

Management VR

The management VR is a Nortel Multiservice Switch virtual router that provides a single point of entry into the Multiservice Switch node and allows the management of all VRs that reside on the node.

A single TCP agent running under the management VR allows external access to the node (for example, through Telnet, Ssh, FTP or FMIP). You can also manage all VRs on the node through a single SNMP agent running under the management VR.

By default, the first VR that you create on a node is the management VR. Once you activate your provisioning view, you cannot designate any other VR as the management VR.

Telnet services are supported on CP-based and Vpn Xc-based virtual routers.

Virtual connection gateway

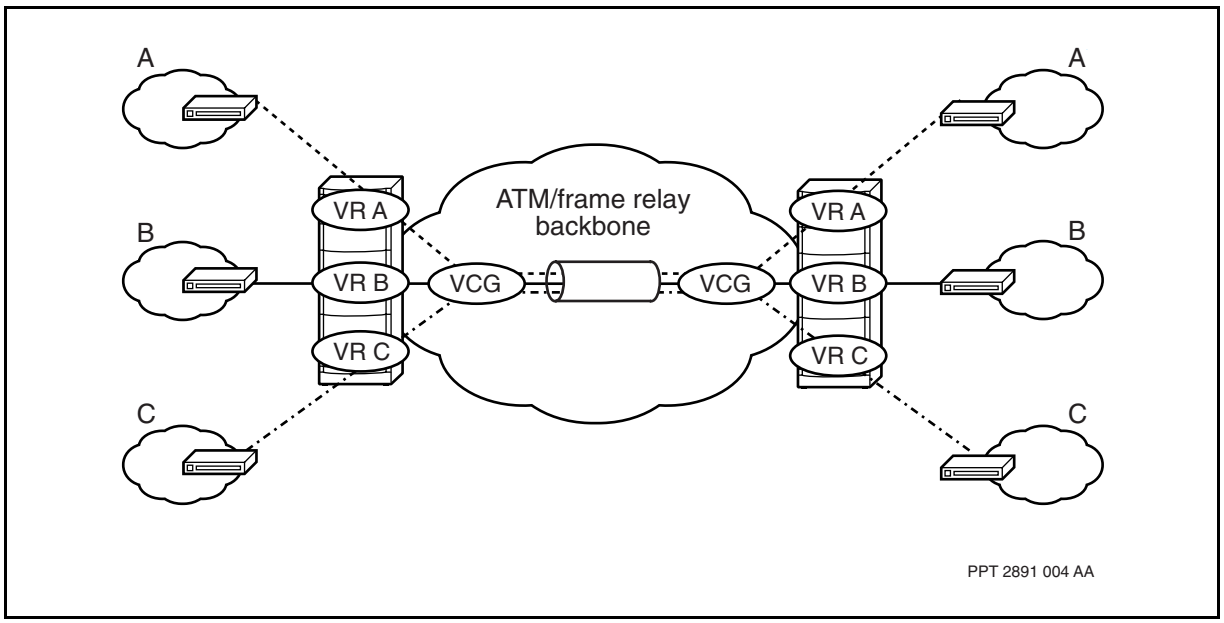
In a typical Nortel Multiservice Switch IP VPN implementation, CPE routers connect to a customer VR assigned to that enterprise. Each customer VR on the node connects to a common VR for the device, called the virtual connection gateway (VCG).

The VCG aggregates traffic from the customer VRs and provides a single outbound connection into the wide area network (WAN) for all individual customer traffic on the node. The VCGs link all Multiservice Switch nodes that provide IP VPN functionality, and provide connectivity between customer VRs in the same IP VPN through point-to-multipoint (PTMP) IP tunnels.

The carrier connects VCGs on each node to achieve full connectivity in the backbone. Since the VCG allows the aggregation of layer 2 connections over the network core, backbone configuration remains unaffected as the carrier adds new customer VRs to the network. See the figure Aggregation of VR traffic in the network backbone (page 17).

Carriers can configure more than one VCG on a node, assigning a subset of customer VRs to each to reduce memory utilization and to enhance throughput on the FPs linked to specific VCGs.

Aggregation of VR traffic in the network backbone



IP routing protocols

Nortel Multiservice Switch customer VRs and VCGs run a separate forwarding table to ensure separation of VPNs. Each VR's IP forwarding table and routing database remains separate from every other VR on the node.

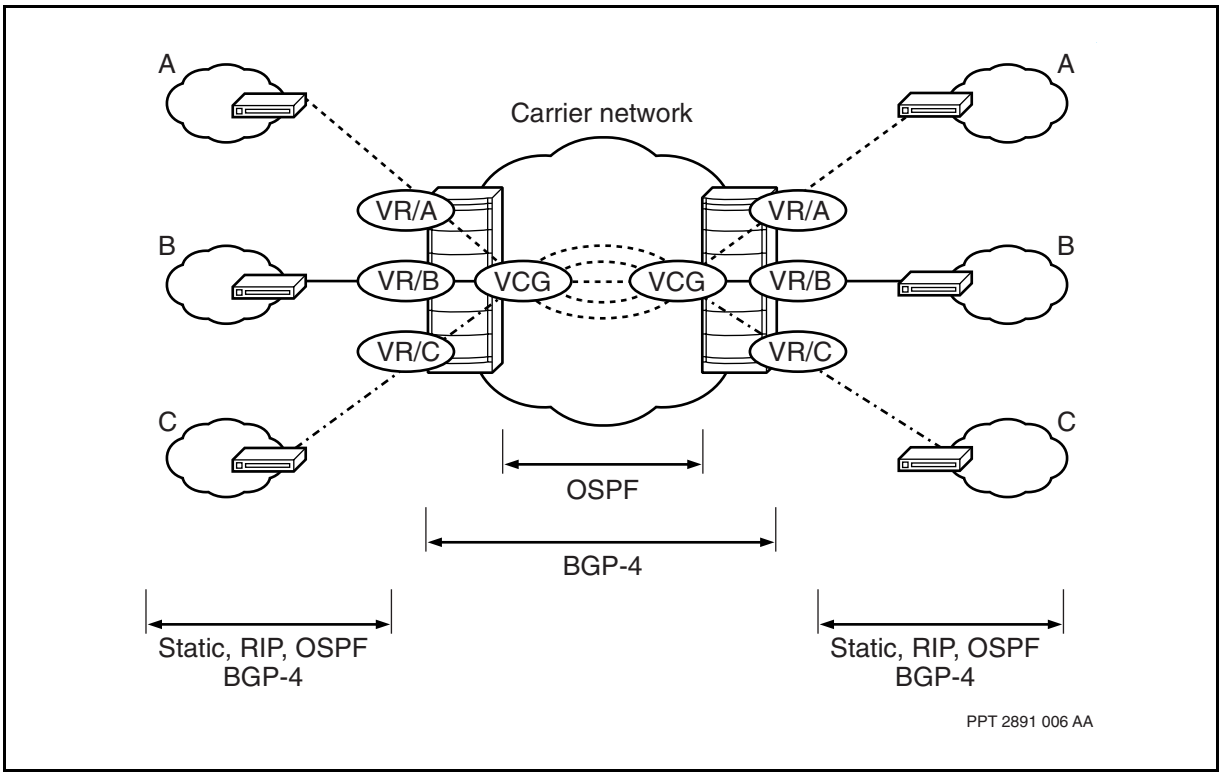
Multiservice Switch VRs support both static routes and dynamic routing protocols, such as RIP v1 and v2, OSPF, and BGP-4. See the figure Routing protocols in the IP VPN service (page 18).

Each customer VR learns routing information from the attached CPE device through static routes or internal gateway protocols (IGP). Dynamic routing is supported through RIP, OSPF, and BGP-4 from the enterprise. Carriers must configure routing protocols on the customer VR to receive this information.

BGP-4 peering provides reachability information within the IP VPN. The customer VRs export the routing information learned from CPE devices through IBGP peering in the backbone. IBGP peering is based on IP tunnel end points, through which routing information is propagated to all other customer VRs that belong to the same VPN. To provide scaling, BGP-4 route reflectors minimize BGP peer configurations as the number of VPN sites increase. VCGs use IGPs, such as OSPF, to exchange routing information in the backbone.

Attention: You can also use OSPF and RIP in NBMA mode to exchange reachability information within the IP VPN, with some engineering restrictions.

Routing protocols in the IP VPN service



Network backbone

Carriers can deploy Nortel Multiservice Switch IP VPN service over a frame relay or ATM infrastructure.

Each VCG on a Nortel Multiservice Switch node can connect to other nodes through its own core technology. With multiple VCGs, a carrier can migrate its backbone from one core technology to another with minimal disruption to service. In addition, multiple logical connections on the VCG allow for translation of IP class of service (CoS) to backbone quality of service (QoS).

VCG-based connectivity

An IP virtual private network (VPN) consists of multiple customer virtual routers (VR), each representing a private customer VPN site. In addition, a single customer VR can support multiple VPN customer sites.

In a VCG-based IP VPN, carriers connect customer VRs using point-to-multipoint (PTMP) IP tunnels through the virtual connection gateway (VCG). VCGs on different Nortel Multiservice Switch nodes connect to each other through logical backbone connections. See the figure VCG-based IP VPN with point-to-multipoint IP tunnels (page 24).

Navigation

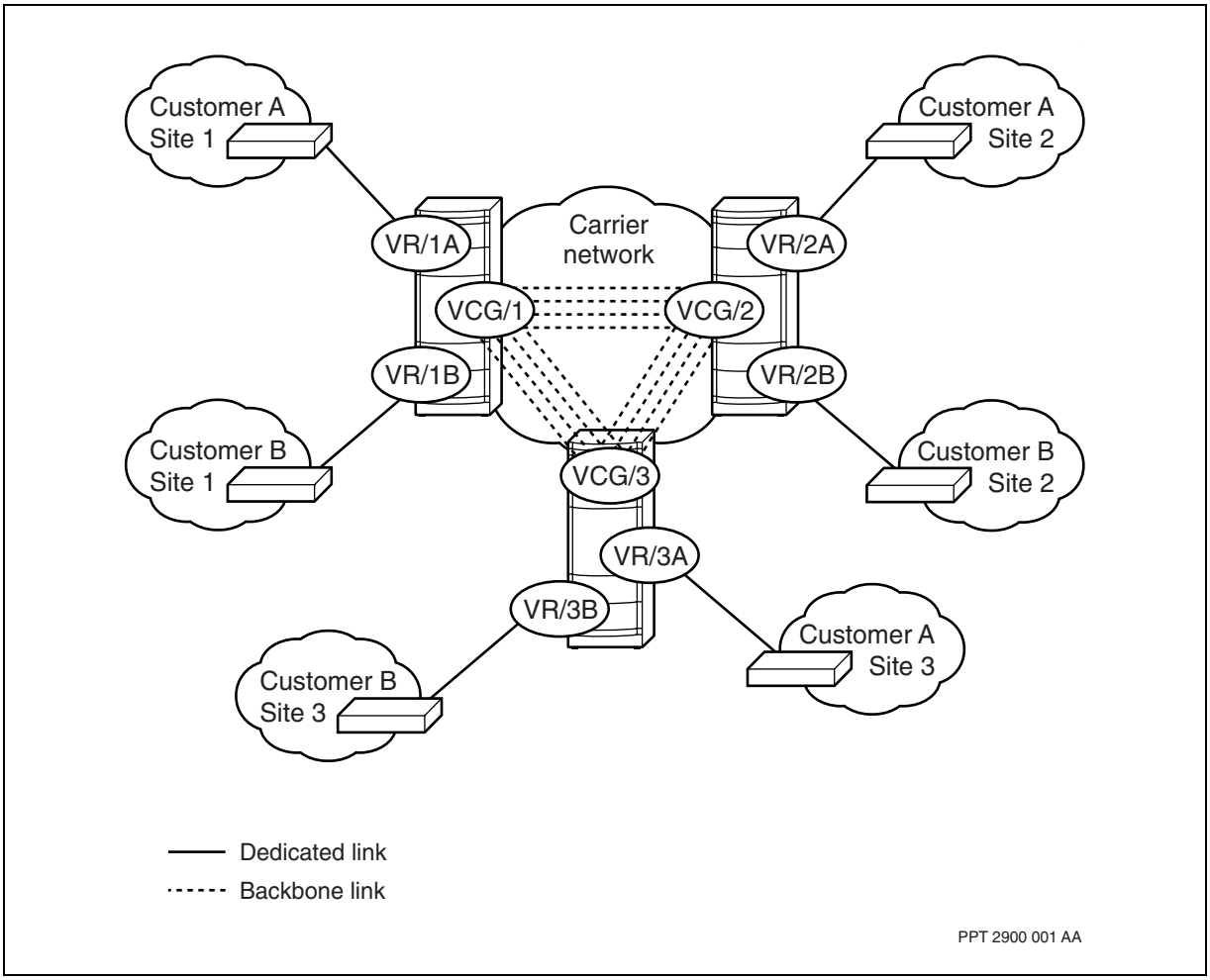
Backbone VC mesh between VCGs

The VCG provides a single outbound connection for all customer VRs on the Nortel Multiservice Switch node, aggregating all customer VR connections. Carriers can connect VCGs across the carrier's public network through ATM VCCs and frame relay DLCIs, resulting in the aggregation of all customer traffic over common backbone links and a reduction in VC meshing. See the figure VCG connectivity in the backbone (page 21).

Each VCG connects to other nodes through its own core technology. With multiple VCGs, carriers can migrate between core technologies with minimal disruption to service.

For information about configuring ATM or frame relay connections between VRs, see NN10600-801 *Nortel Multiservice Switch 7400/15000/20000 IP Configuration Management*.

VCG connectivity in the backbone



Dynamic and static VPNs

You can configure dynamic and static VPNs. Dynamic VPNs are the recommended configuration.

A static VPN requires more complex manual configuration, for example, adding the static Address Resolution Protocol (ARP) entries for each remote IP tunnel end point defined on the customer VR.

When you add a static ARP entry with a CoS index, all existing dynamic ARP entries with the same IP address are deleted upon activation, regardless of the CoS index.

Dynamic VPNs are configured by enabling the auto discovery feature and are much simpler to provision. Auto discovery enables the customer VRs to dynamically learn the public and private addresses of the tunnel end points across the backbone. You do not need to provision PTMP destination

addresses or the tunnel end points in the static ARP table as you do in a static VPN. Instead, the tunnel end point information is automatically exchanged among the customer VRs.

If your VPN is connected to a device that does not support inverse ARP, you must manually configure the ARP entries for the IP tunnel end points. Nortel Multiservice Switch systems support inverse ARP.

In a dynamic VPN, BGP allows the creation of dynamic IBGP peers in various modes. The value of the mode is set by the *Vr Ip Bgp vpnPeeringTopology* attribute. Set this attribute to none, hub, or spoke as follows:

- Use none to tell BGP to never create dynamic IBGP peers. This value is useful when an IP routing protocol other than BGP is used to distribute VPN site information.
- Use hub to establish a fully-meshed peering topology. In this case, all discovered remote tunnel end points need to have their *vpnPeeringTopology* attribute set to hub. You can also use the hub setting to establish a star topology. In this case, dynamic IBGP peers are created only for discovered remote tunnel end points whose *vpnPeeringTopology* is set to hub or spoke.
- Use spoke to represent a hub client. In this case, dynamic IBGP peers are created only for discovered remote tunnel end points whose *vpnPeeringTopology* is set to hub.

In order to create a dynamic VPN, you need to include the following steps when you are provisioning the VPN:

- BGP must be running on the VCG and the attributes *Vr Ip Bgp locals* and *Vr Ip Bgp Peer Desc peeras* must be equal. Also, the *Vr Ip Bgp Peer Desc addressFamily* attribute must include ipv4vpn. See the section on BGP in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.
- The virtual private network identifier (VPN ID) must be set to the same value for all the VPN sites that are using auto discovery in the VPN. See the section on configuring customer VRs in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.
- The *Vr Ip Tunnel Msep autoDiscovery* attribute must be enabled for VPN sites using auto discovery. See the section on configuring VCGs in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.
- The customer VRs in VPN sites using auto discovery must have their peering topology set and, optionally, be set up as route reflectors. This involves provisioning the *Vr Ip Bgp vpnPeeringTopology*, *Vr Ip Bgp rr*, and *Vr Ip Bgp rrCluster* attributes as required. See the section on configuring

route distribution in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.

You can override a dynamic peer with a static peer in a dynamic VPN if the system-provided values of the dynamic peer are not the desired configuration. See the example in the section on configuring route distribution in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*. In a dynamic VPN, all static peers must be in the same AS and all VPN sites running auto discovery must be in the same AS.

Attention: Because auto discovery does not support EBGP peering in the core, no auto discovery information is sent to external BGP peers. To distribute auto discovery information, you must configure all BGP instances on each VCG to be in the same AS (i.e., IBGP).

Point-to-multipoint IP tunnels

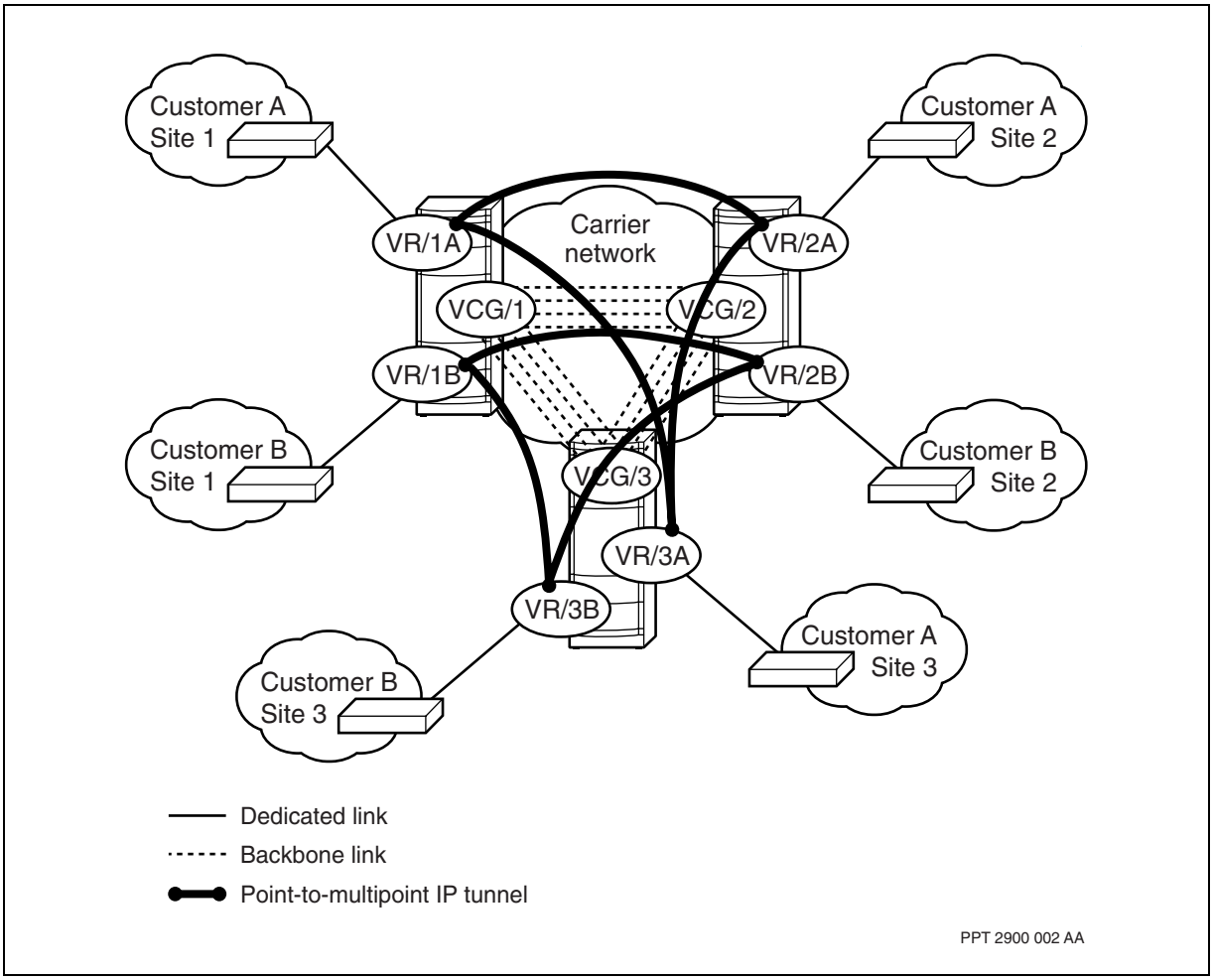
Nortel Multiservice Switch IP VPN service uses PTMP IP tunnels to provide connectivity between customer VRs that reside on different Multiservice Switch nodes. For full site-to-site connectivity, the carrier must configure the source address, and in static VPNS only, the multiple destination addresses of the PTMP IP tunnel on every customer VR in the IP VPN. The customer VR performs IP in IP tunnel encapsulation (as defined in RFC 2003) at the ingress. The VCG performs decapsulation at the egress.

Attention: IP in IP tunnel encapsulation is not compatible with RFC 2003 when using an ATM IP FP as a backbone FP.

For more information, see the following sections:

- PTMP IP tunnel end points (page 24)
- Tunnel source and destination addresses (page 25)
- Tunnel end point address resolution (page 25)
- Tunnel optimization (page 26)
- Path MTU discovery (page 31)
- IP VPN accounting statistics for PTMP tunnels (page 32)

VCG-based IP VPN with point-to-multipoint IP tunnels



PTMP IP tunnel end points

Each end point of a PTMP IP tunnel has a source address and multiple destination addresses. Nortel Multiservice Switch systems support two PTMP IP tunnel instances on each customer VR. Two PTMP IP tunnel instances allow you to migrate a VPN site from one VCG to another VCG configured on the same Multiservice Switch node.

PTMP IP tunnel end points stretch across the customer VR to the VCG on each Multiservice Switch node. Each PTMP IP tunnel end point maps to a private address on the customer VR. The source and destination addresses of the PTMP IP tunnel represent public addresses on the VCG. The private addresses belong to the same subnet, and must be unique only within the IP VPN. Each public tunnel address must be unique across the network.

Multiservice Switch systems theoretically supports an unlimited number of end points (that is, destination addresses) on every PTMP IP tunnel. There must be a corresponding number of remote VCGs for each end point. To achieve optimum performance, the actual number of tunnel end points must fall within the design limits specified by the engineering guidelines.

Tunnel source and destination addresses

The PTMP IP tunnel source address is a public address that maps to a logical interface under a virtual media protocol port on the VCG. The destination addresses are public addresses that map to virtual media logical interfaces on remote VCGs.

A virtual media application is not associated with a physical port. Since logical IP interfaces under the virtual media application are defined independently of any physical media, they remain up even if individual links to the Nortel Multiservice Switch node experience loss of connectivity. An IP address associated with a virtual media protocol port is always reachable as long as the node itself remains connected to the network.

Each VCG supports multiple logical interfaces under a single virtual media protocol port. Each logical interface under the associated virtual media software component represents an aggregate of the source addresses of all PTMP IP tunnels, where each source address is in the subnet range of the logical interface. Be careful not to provision duplicate source addresses.

In order to support source address aggregation, set attribute *Vm Interface mode* to alwaysUpSummary. If you have used alwaysUpInterface, you can migrate to alwaysUpSummary by following the procedure Migrating to the alwaysUpSummary mode for virtual media in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.

All PTMP IP tunnel configuration occurs on the customer VR. The carrier configures the source address of the PTMP IP tunnel on the customer VR by associating it with a logical interface under the VCG's virtual media protocol port. In a static VPN, the carrier configures the PTMP IP tunnel destination addresses statically on the customer VR. In a dynamic VPN, the auto discovery feature allows the customer VRs to dynamically discover the destination addresses of other customer VRs in the same VPN.

Tunnel end point address resolution

The ARP table on the customer VR provides address resolution between a tunnel end point's private address on the customer VR and its corresponding (public) destination address on the VCG. In a dynamic VPN, the ARP table is dynamically updated with the tunnel end point's private address and destination address. In a static VPN, you must manually configure a static ARP entry for every tunnel end point on every customer VR within the same VPN.

The private PTMP IP tunnel addresses that belong to the same VPN must be in the same subnet. The public PTMP IP tunnel addresses need not be in the same subnet. Since the public PTMP IP tunnel address resides on the VCG, enterprise customers cannot see the carrier's address, and private PTMP IP tunnel addresses can overlap between different IP VPNs.

Tunnel optimization

Tunnel optimization moves the processing of IP VPN tunnel end points from the trunk card (where the VCG resides) to the access cards, which completely removes the corresponding customer VRs from the trunk card. The resources formerly used by the customer VRs and their interfaces on the trunk card are distributed across the access cards, increasing the overall capacity of the shelf.

Attention: If you want to deploy tunnel optimization, contact your Nortel Networks technical representative for information on the scalability and engineering aspects of this feature.

Background

In Nortel Multiservice Switch IP VPN, customer traffic is aggregated over IP PTMP tunnels by encapsulating the packets in a second IP header at the tunnel entry. By default, IP decapsulation of these tunnels occurs on the trunk card where the remote tunnel end points reside, specifically, on the ingress forwarding engine processor on the trunk card.

Tunnel optimization description

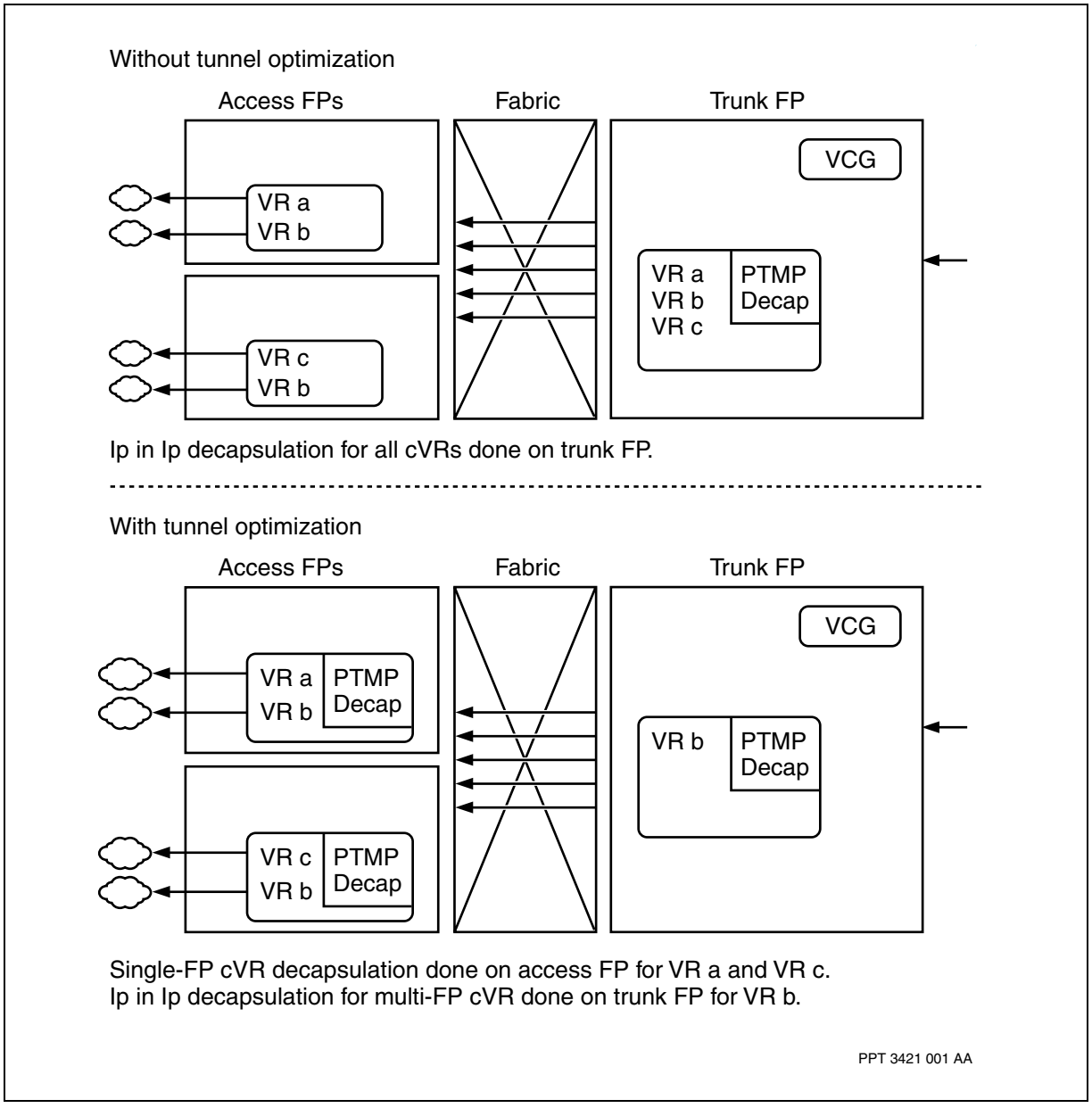
Tunnel optimization moves IP tunnel decapsulation processing onto the egress forwarding engine of the access card, where the customer VR resides. Decapsulation is then spread across multiple access cards instead of all decapsulation occurring on the trunk card. See the figure Tunnel optimization (page 27).

Benefits of tunnel optimization include:

- Significant reduction in the amount of memory required on the trunk card and an increase in its forwarding throughput.
- For deployments using the VPN extender card, an increase in the total number of CPE routers that can access a Multiservice Switch node for IP VPN services.
- For service providers with IP VPN customers with widely distributed VPNs (many sites per VPN), an increase in the number of VPNs that can be supported on a node. This increase is accomplished by allowing an increase in the average number of sites per VPN that can be supported on a node without the need to reduce the total number of VRs on the node.

Tunnel optimization is only supported on PQC2 and PQC12 FPs. See NN10600-551 *Nortel Multiservice Switch 7400/15000/20000 FP Configuration Reference* for FP information. For performance reasons, it is recommended that you disable tunnel optimization on any customer VR where a small MTU has been provisioned, or if the PQC FP shares a tunnel port with a QGM-based FP. See the section in NN10600-800 *Nortel Multiservice Switch 7400/15000/20000 IP Technology Fundamentals* on provisioning MTU size for more information.

Tunnel optimization



Conditions preventing tunnel optimization

Any of the following conditions prevent a tunnel from being optimized:

- The customer VR has *ProtocolPort* components that are linked to components representing physical media on more than one access card. This includes the cases when one of the cards is a standby card, and when the physical media is linked to a SONET or SDH port that is protected by inter-card APS.
- SPVC *AtmMpe Ac*'s reside on more than one access card.
- The *VirtualMedia Interface* component corresponding to any of the *Msep* subcomponents of the *Tunnel* instance has its mode attribute set to a value other than alwaysUpSummary Interface.
- The *Tunnel* component has a *Sep* subcomponent.
- The customer VR has a protocol port that is linked to an *AtmMpe* component whose *encapType* attribute is set to llcBridgeEncap.
- The customer VR has a protocol port that is linked to a *VirtualMedia Interface* component whose *mode* attribute is set to interVrConnection.
- The customer VR is not linked to any real media. See Tunnel optimization with no access card (page 30).
- There is insufficient *Lp Eng Arc connectionPoolAvailable* on the cVR access card. Each VirtualMedia Interface, with its mode attribute set to "alwaysUpSummary", requires four connections in addition to those already used by the media on the cVR access card.

Tunnel optimization configuration

Configure tunnel optimization on customer VRs, not VCGs. There are two related attributes: the provisionable attribute *Vr Ip Tunnel optimization* and the operational attribute *Vr Ip Tunnel optimizationStatus*. Set *optimization* to *enabled* to make a tunnel available for optimization and then check the value of *optimizationStatus* to see whether it is actually optimized. For more information on these attributes and their values, see NN10600-060 *Nortel Multiservice Switch 7400/15000/20000 Component Reference*.

The following two indicators are useful when checking for tunnel optimization:

- A non-optimized tunnel can be optimized only when *optimizationStatus* is either *eligible* or *optimizationOnHold*.
- A tunnel is actually optimized only when *optimizationStatus* is *optimized*.

The following tables describe the behavior of *optimizationStatus* when tunnel optimization is enabled or disabled. Note that an exceptional case occurs in Disabling tunnel optimization (page 29) and Disallowing tunnel optimization (page 30) when the customer VR is not linked to any real media. See Conditions preventing tunnel optimization (page 28).

- Enabling tunnel optimization (page 29) shows how *optimizationStatus* is affected when you enable the feature by setting *optimization* to *enabled*.

Enabling tunnel optimization

		Updated value of optimizationStatus
Initial value of optimizationStatus	**eligible**	optimized
	ineligible	ineligible

- Disabling tunnel optimization (page 29) shows how *optimizationStatus* is affected when you disable the feature by setting *optimization* to *disabled*.

Disabling tunnel optimization

		Updated value of optimizationStatus
Initial value of optimizationStatus	**ineligible**	ineligible
	optimized	eligible
	optimizationOnHold	eligible

- Disallowing tunnel optimization (page 30) shows how *optimizationStatus* is affected when conditions occur, either dynamically or via provisioning, that prevent tunnel optimization from occurring. See Conditions preventing tunnel optimization (page 28). The tunnel detects when these conditions occur and when necessary, automatically changes state.

  When this occurs, the tunnel automatically becomes non-optimized but the value of the provisionable attribute *Vr Ip Tunnel optimization* does not change.

Disallowing tunnel optimization

		Updated value of optimizationStatus
Initial value of optimizationStatus	**optimized**	ineligible
	eligible	ineligible
	ineligible	ineligible
	optimizationOnHold	ineligible

- Allowing tunnel optimization (page 30) shows how *optimizationStatus* is affected when the conditions listed in Conditions preventing tunnel optimization (page 28) are cleared and tunnel optimization is again allowed. For example, a VR/VPN that previously appeared on two access FPs is changed to appear on only one access FP as shown in the figure Tunnel optimization (page 27). The tunnel detects when these changes occur and when necessary, automatically changes state.

  When this occurs the tunnel does not automatically become optimized and the value of the provisionable attribute *Vr Ip Tunnel optimization* does not change. To optimize the tunnel you must do one of the following:

  — If *optimization* is *disabled*, set it to *enabled.*

  — If *optimization* is *enabled* and *optimizationStatus* is optimizationOnHold, set *optimization* to *disabled* and then set it back to *enabled*.

Allowing tunnel optimization

		Updated value of optimizationStatus	
		optimization is enabled	**optimization is disabled**
Initial value of optimizationStatus	**ineligible**	optimizationOnHold	eligible

Tunnel optimization with no access card

When the customer VR has no access card, its tunnel is ineligible for optimization. However, if a customer VR with an optimized tunnel loses its access card, the tunnel remains in the optimized state. This prevents the customer VR from having to be present on the trunk card, which protects trunk card resources. Under normal circumstances, the state changes from optimized to ineligible as shown in table Disallowing tunnel optimization (page 30).

If you later disable tunnel optimization by setting *optimization* to *disabled*, the tunnel changes state from optimized to ineligible. Under normal circumstances, the state changes from optimized to eligible as shown in table Disabling tunnel optimization (page 29).

Migrating multiple tunnels to/from the optimized state

When migrating multiple tunnels to or from the optimized state, the following is recommended in order to minimize the traffic outage per customer VR:

- When enabling tunnel optimization, enable no more than fifty customer VRs at a time until they all become enabled.
- When disabling tunnel optimization, disable a single customer VR at a time.

Migrating tunnels places demands on the VPN extender card and trunk cards that can cause a protracted traffic outage if more than the recommended number of tunnels are migrated simultaneously. All transitions take effect within ten minutes.

Path MTU discovery

In a PTMP IP tunneling configuration, IP datagrams occasionally need to travel across dissimilar network mediums to reach a destination address (DA). Each network medium requires the IP datagram to conform to a specific byte length. The maximum transmission unit (MTU) is the largest unit of data that a network's physical medium can transmit. As an IP datagram passes from one network medium to another, the datagram can undergo fragmentation due to the network's MTU requirements. Ethernet, Frame relay, and ATM networks transmit information using different size data units.

- Ethernet networks transmit IP datagrams of default size 1500 bytes (MTU=1500)
- Frame relay networks transmit IP datagrams of default size 1600 bytes (MTU=1600)
- ATM networks transmit IP datagrams of default size 9180 bytes (MTU=9180)

Path MTU discovery allows the PTMP IP tunnel port to detect the entire network path and adjust the IP datagrams to the optimum length for delivery across all related networks. This feature reduces excessive IP data fragmentation at each network node and reassembly at the destination.

When path MTU discovery is enabled, the PTMP IP tunnel port stores all destination addresses (DAs) and associated MTU values. When data enters the PTMP IP tunnel, the PTMP IP tunnel port selects the lowest common MTU

value based on the DA and the related routing networks. The port creates the appropriate size data fragments and sends the data across all relevant networks.

For more information on how to configure the MTU for a PTMP IP tunnel, see the table Configuring PTMP IP tunnels for customer A's dynamic IP VPN in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.

IP VPN accounting statistics for PTMP tunnels

For VCG-based configurations the IP VPN accounting service allows you to collect, record and report layer 3 usage measurements for each customer VR that is part of an IP VPN. Site-to-site accounting information is available for VRs within a VPN connected by PTMP tunnels.

IP tunnel encapsulation and decapsulation counts are collected for PTMP IP tunnel configurations on ATM IP functional processors. The statistics are measured by CoS on the ingress traffic of a PTMP IP tunnel. Source and destination address counts are provided for the egress traffic. For further information on CoS, see NN10600-808 *Nortel Multiservice Switch 7400/15000/20000 Layer 3 Traffic Management Fundamentals*.

For PTMP IP tunnel configurations, an accounting record is generated for each source address and destination address pair. If accounting is enabled at both ends of the tunnel, then two accounting records are generated for each tunnel (double-ended accounting).

For more information on IP VPN accounting, see NN10600-560 *Nortel Multiservice Switch 7400/15000/20000 Accounting*.

Round-trip delay measurements

The round-trip delay (RTD) measurements feature, which is based on RFC 2681, lets you measure the time it takes for an ICMP packet to travel from a VCG, across the backbone to any of the remote VCGs, and back to the original VCG. You can make this measurement for each of the four available class of service (CoS) indexes that can be defined on the connection.

RTD measurement is available on VCGs and not on customer VRs. The list of destination addresses for which RTD measurement is calculated is automatically derived from the BGP peer list on the VCG. If auto discovery is enabled, the BGP peer list contains both the static BGP peer set and the dynamically learned BGP peer set, which is provided by all the route reflectors in the network. For more information on auto discovery, see the section on configuring the management VR and VCG on the Nortel Multiservice Switch node in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management* and Dynamic and static VPNs (page 21).

A timestamp is stored on the originating VCG when the ICMP packet is sent out. Once the ICMP packet travels back to the originating VCG, a new timestamp is taken. The RTD calculation is based on the two timestamps.

Enable RTD measurement by adding component *Vr Ip Rtd*. This component has attributes that allow you to customize related parameters, such as the CoS indexes and the destination addresses for which RTD can be measured. For more information on these attributes, see NN10600-060 *Nortel Multiservice Switch 7400/15000/20000 Component Reference*.

Update the list of CoS indexes (attribute *Vr Ip Rtd rtdCosList*) if you have not defined all four available CoS indexes on your connection. Otherwise, the Multiservice Switch node uses resources to try to calculate a non-existent RTD measurement.

Update the list of destination addresses (attribute *Vr Ip Rtd rtdDstAddrList*) to create a customized list of destination addresses that overrides the default BGP peer list. You may want to do this under either of the following conditions:

- To extend the destination address list to include nodes that are not part of the automatically derived BGP peer list
- To reduce the destination address list to a subset of the automatically derived BGP peer list

Display RTD measurement results by displaying the attributes of component *Vr Ip Rtd VcgDestAddr*.

For information on provisioning RTD measurement and displaying the results, see the section on configuring round-trip delay measurements in NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.

IP over ATM soft PVCs

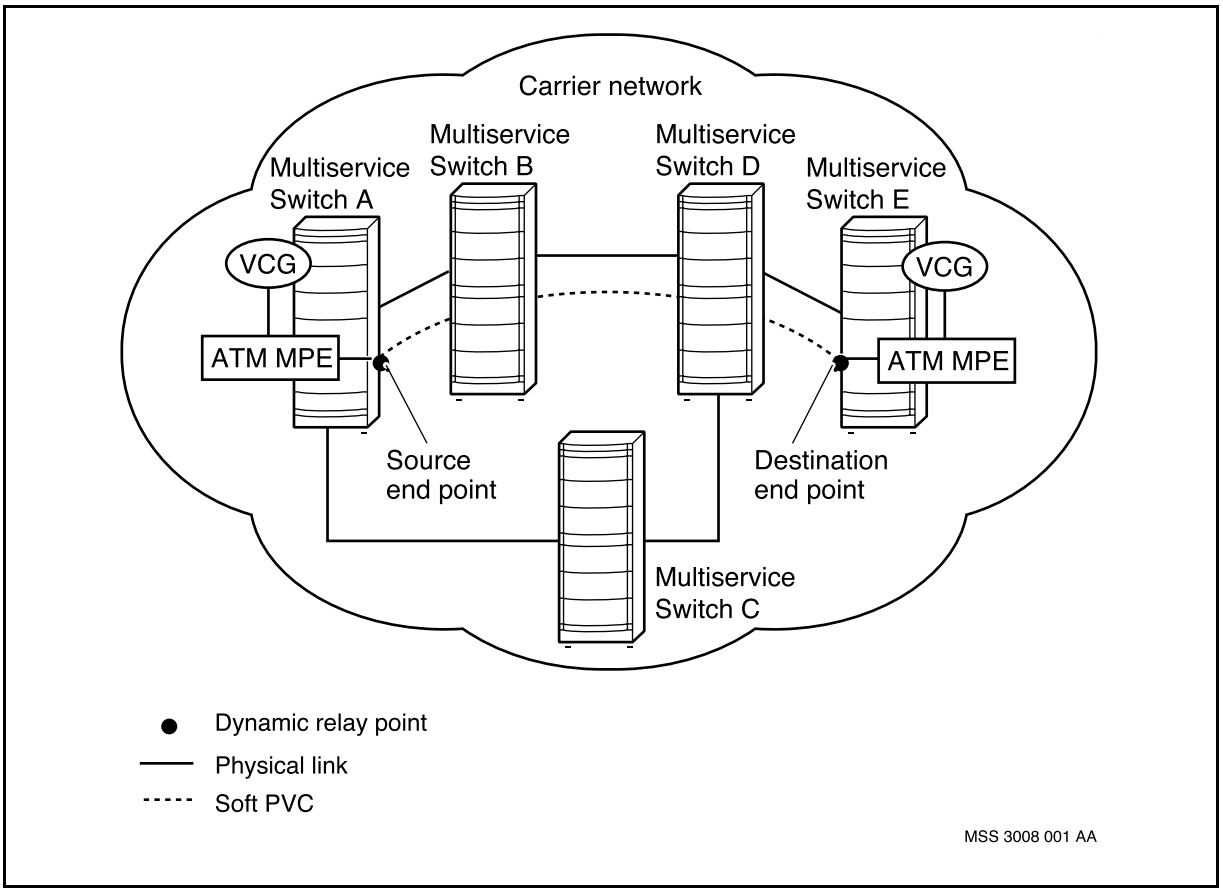
Carriers can connect VCGs across the ATM backbone using either permanent virtual circuits (PVC) or soft PVCs. In a PVC, all the connection points through the network are defined, or nailed up. In a soft PVC, only the end points are defined. ATM PNNI routing provides route selection through the network between the end points. (For information on PNNI, see NN10600-702 *Nortel Multiservice Switch 7400/15000/20000 ATM Routing and Signalling Fundamentals*.)

Attention: ATM MPE soft PVCs are only supported in a PNNI network.

It is possible to configure a soft PVC between an ATM UNI interface and ATM MPE. See NN10600-800 *Nortel Multiservice Switch 7400/15000/20000 IP Technology Fundamentals* for more information about this configuration.

The figure IP over ATM soft PVC (page 34) shows a simplified network in which IP traffic is transmitted from Multiservice Switch A to Multiservice Switch E over a soft PVC. One end point is assigned through provisioning as the source and the other as the destination. Each end point is identified with an NSAP address. (For information on NSAP addressing, see NN10600-702 *Nortel Multiservice Switch 7400/15000/20000 ATM Routing and Signalling Fundamentals*.)

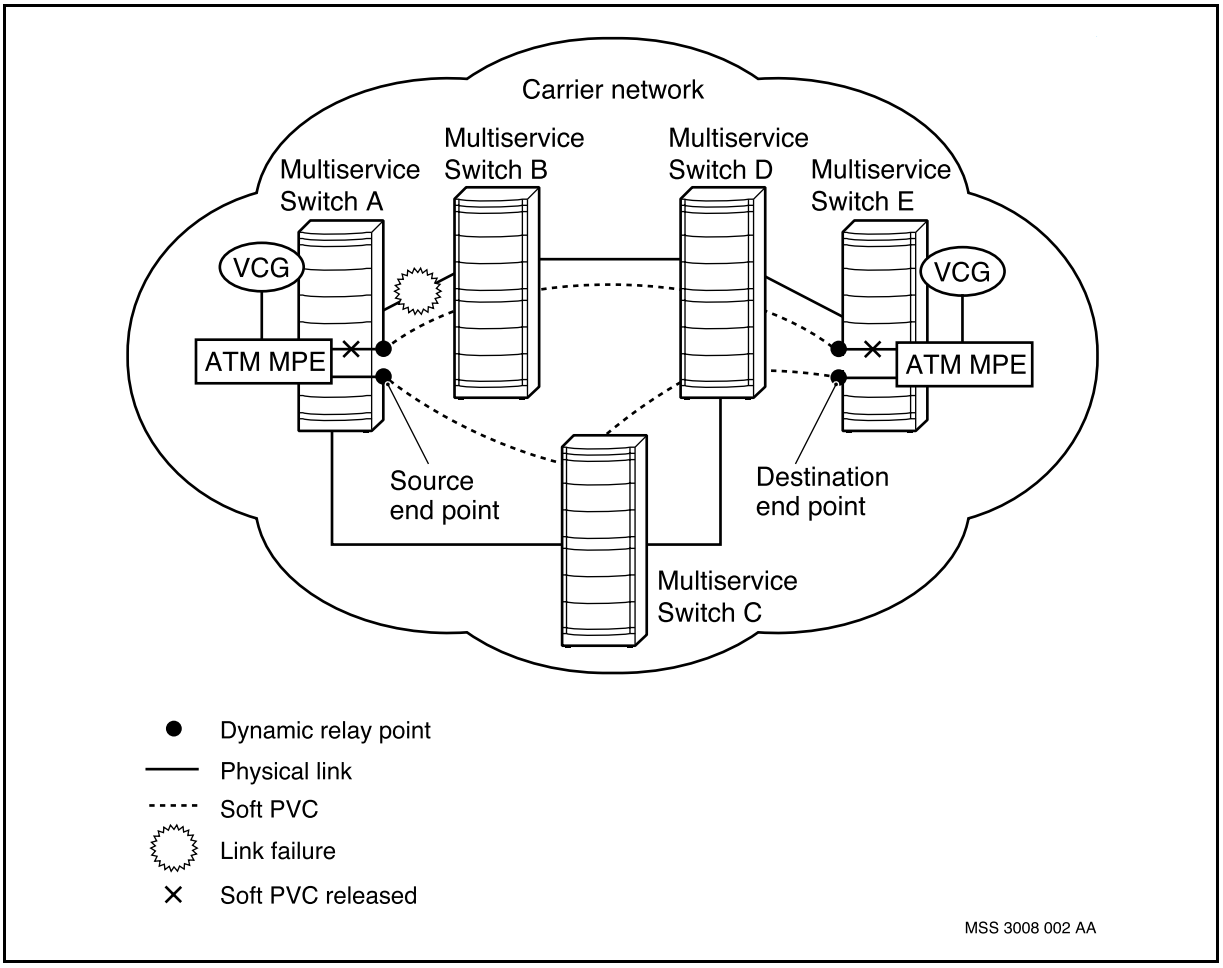
IP over ATM soft PVC



The route that the PVC takes between the end points is established through private network-to-network interface (PNNI) routing and signaling procedures. The source, or calling, end point owns the soft PVC, and initiates the signaling for the connection. The calling end point is configured with the information needed to reach the remote end point, and the soft PVC is set up dynamically through the backbone at activation time. The PNNI nodes use dynamic routing and signaling to establish an ATM connection with a pre-defined traffic contract according to the provisioned ATM service category and peak cell rate (PCR). (For information on ATM traffic management, see NN10600-705 *Nortel Multiservice Switch 7400/15000/20000 ATM Traffic Management Fundamentals*.)

Soft PVCs provide path resiliency through dynamic ATM rerouting in cases of failure. For example, in the figure IP over ATM soft PVC resiliency (page 35), a link failure has occurred between Multiservice Switch A and Multiservice Switch B. In this case, the existing soft PVC is released at both end points. Then the PNNI system reroutes the soft PVC dynamically through Multiservice Switch C.

IP over ATM soft PVC resiliency



Routing information between VPN sites

In a VCG-based configuration, carriers can use BGP-4 to distribute IP routing information across the IP VPN. For more information, see the following sections:

- IBGP at PTMP IP tunnel end points (page 36)
- BGP-4 route reflectors (page 36)
- Passive OSPF interfaces (page 36)

IBGP at PTMP IP tunnel end points

In a VCG-based configuration, carriers can use BGP-4 to distribute IP routing information across the IP VPN. Carriers configure internal border gateway protocol (IBGP) peers at each PTMP IP tunnel end point, so that IP routing information transmits through the IP tunnel to all other VPN sites.

IBGP peers can export routing information from IGPs such as OSPF, RIPv1, and RIPv2. In addition, carriers can use OSPF and RIP running in non-broadcast multi-access (NBMA) mode at PTMP IP tunnel end points to exchange routing information, with some engineering considerations.

BGP-4 route reflectors

Carriers can also use Nortel Multiservice Switch BGP-4 route reflector functionality to provide scaling and redundancy. Since BGP-4 route reflectors advertise only the best IBGP routes to their clients, the number of routes processed by client peers is far less than in a fully meshed peering configuration.

A BGP router configured as a route reflector allows a 1:n peer relationship, instead of a fully-meshed n:n-1 peer relationship. As the number of VPN sites (and therefore, the number of IBGP speakers) increase, configuration changes are restricted to the route reflector and new IBGP speaker only.

In addition, by configuring multiple BGP-4 route reflectors in a given cluster, carriers can provide redundancy by having critical nodes peer to more than one route reflector simultaneously.

For more information about BGP-4 route reflection in Multiservice Switch systems, see NN10600-800 *Nortel Multiservice Switch 7400/15000/20000 IP Technology Fundamentals*.

Passive OSPF interfaces

Each VCG has a virtual media (VM) protocol port through which to configure a logical interface, which aggregates the source addresses (SAs) of all PTMP IP tunnels. If you configure the aggregated logical interface as an OSPF interface, set attribute *Vr Pp IpPort LogicalIf OspfIf ifType* to passive. Then, OSPF propagates this logical interface subnet address into the OSPF autonomous system.

Multi-protocol BGP route distribution

The information in this section only pertains to multiprotocol BGP for the RFC 2764 VCG solution. For information about multiprotocol BGP for RFC 2547, see VPN route distribution and routing policy using BGP (page 52).

Navigation

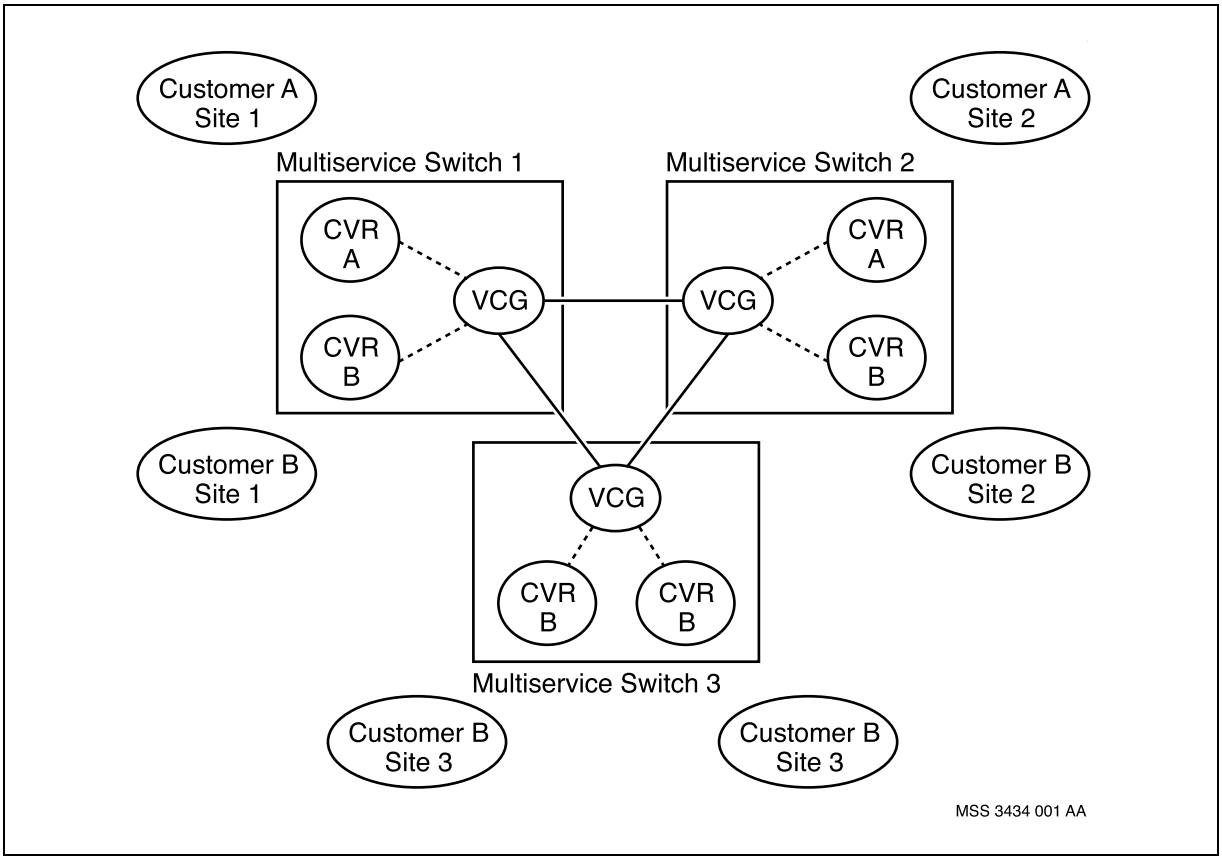
MBGP route distribution overview

By using BGP Multi-protocol extension messages (mBGP), we eliminate the need to run routing protocols directly among customer Virtual Routers (cVRs) belonging to the same VPN. Instead, the routing information for each cVR is distributed by mBGP via the Virtual Connection Gateway (VCG) Internal BGP (IBGP) peer connections on behalf of the cVRs. This significantly improves the scalability of service providers' (SPs) VPN network. For an illustration, see BGP distribution of VPN routing information (page 38). You can also add import and export policies for VPN routes on both the customer VR and the VCG level. This allows the VPN customer and carrier to have better control over:

- selecting which local cVR routes to distribute to the carrier VCG
- selecting which remote cVR routes to accept at the local cVR
- selecting which local VCG routes to distribute to the remote VCG peer
- selecting which remote VCG routes to accept from the remote VCG peer

MBGP route distribution is enabled when the *mBgp* component is configured on the customer VR and the *addressFamily* attribute of the VCG IBGP peers is set to include *mbgpVpn*. For information about configuring mBGP route distribution, see NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.

BGP distribution of VPN routing information



Automatic filtering

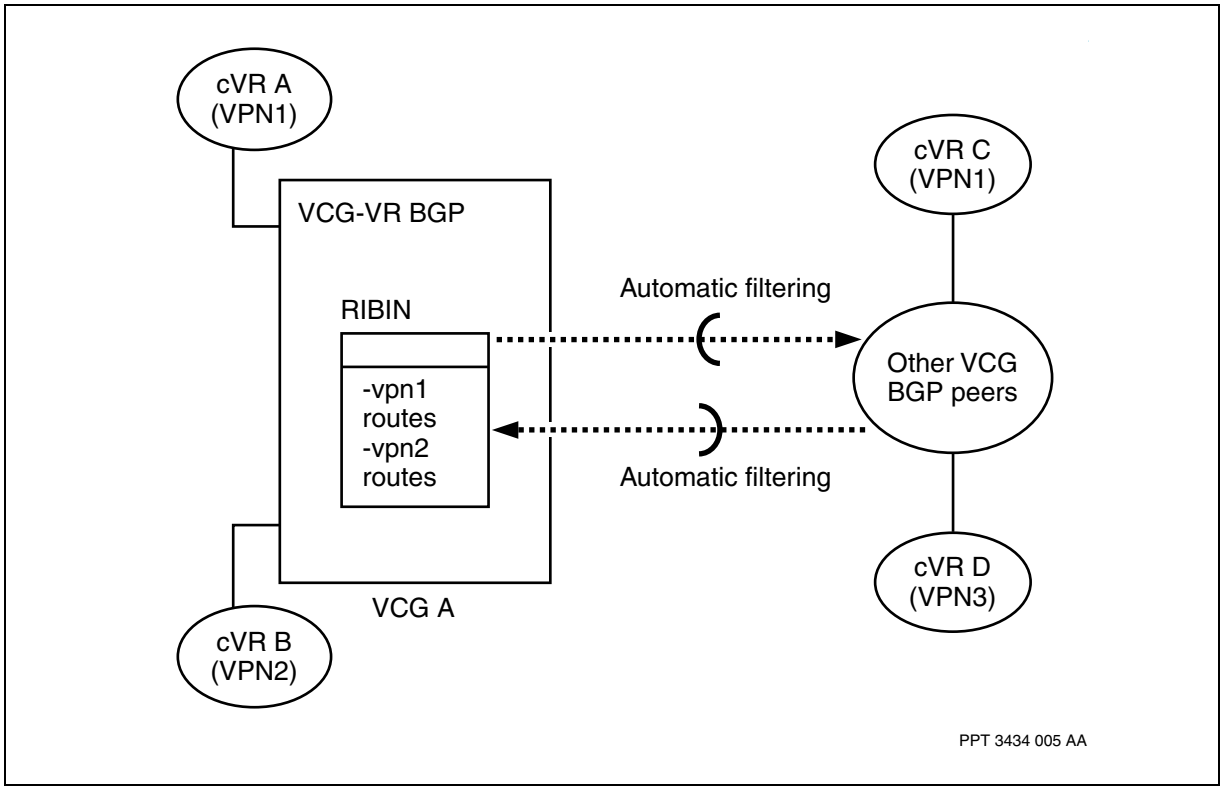
Automatic filtering does not require any additional configuration and is automatically enabled on the VCG for distribution of MBGP routes. Routes are distributed between VCGs on a need-to-know basis, based on the VPNs with which they are associated. This reduces unnecessary messaging and keeps the Routing Information Base (RIBIN) in the VCG as small as possible. If import and export policies are provisioned, automatic filtering works in conjunction with them. For information about import and export policies, see Import and export policies (page 41).

When importing routes learned from other VCG-BGP peers into the VCG-BGP's RIBIN, the VCG-BGP will apply automatic filtering by each of its peers via the mBGP extension messages to only import those routes whose VPN

are supported by the VCG. For example, MBGP automatic filtering (page 39) shows VCG A only supporting routes that belong to VPN1 and VPN2 and will only import routes that belong to those VPNs. When exporting mBGP VPN routes, the VCG-BGP will apply automatic filtering so that only those routes whose VPN is supported by the peer VCG are sent. The net effect is that only routes whose VPNs are supported by the VCG are saved in the RIBIN, thus reducing the storage requirements of the VCGs.

When a VPN is removed from one VCG side, it will inform the peers about the change so that VPN routes learned from the VCG can be withdrawn altogether.

MBGP automatic filtering



The VCG learns which VPNs are supported by each of its peers by messaging. This allows the selective announcement of routes to its peers based on which VPNs the peer VCG supports. This reduces the number of messages that needs to be sent during route distribution.

Similarly, the VCG learns which VPNs are no longer supported by each of its peers by messaging. All routes belonging to the unsupported VPN that were learned from the peer are automatically withdrawn from the RIBIN database. This eliminates the need for any route withdrawal messages from that peer for the actual individual routes, thereby reducing messaging.

Route selection

When a route is received by BGP running mBGP route distribution, the route is immediately subject to route selection. If the same prefix from the same VPN is received from more than one VCG peer, only one “best” route is selected and installed. The other route is kept in the RIBIN and, in the case that the “best” route goes away, the second best route will be promoted to be the best route.

There are various route selection criteria; but, in the case where all the criteria are the same between the two routes under comparison, the route with the lowest BGP router ID is selected as the best route.

The route selection order is determined as follows:

- higher local preference
- lower AS weight
- lower MED
- shorter cluster list
- lower peer router ID
- lower remote peer address

If the first criteria for selection is the same for both routes, the second criteria for selection is considered.

Route selection is always performed on the cVR. After applying the VCG and, if required, the cVR import policy, VCG will not make a selection on behalf of the cVR. In the case where the VCG or cVR import policy is blocking the specific route, the import policy will never make it to the cVR to participate in the cVR best route selection process.

Mbgp route preference

You can configure the route preference for routes imported from VCG-BGP’s RIBIN on each cVR. For routes learned by mBGP route distribution, the protocol is set to mBgp. The preference value will be used for tie-breaking when there are multiple routes from different protocols to the same destination. During the migration of existing Nortel Multiservice Switch IP VPN configuration to mBGP route distribution enabled IP VPN configuration, the preference value is used to choose the mBGP route over the ipv4 unicast Bgp route.

The default value of the mBgpRtePreference is 123. (Ibgp is 122). Ibgp routes are preferred over mBGP routes by default.

When the mBGP route preference value is changed, the new route preference value takes effect without affecting mBGP route distribution or traffic flow.

Import and export policies

The following import and export policies are discussed in this section:

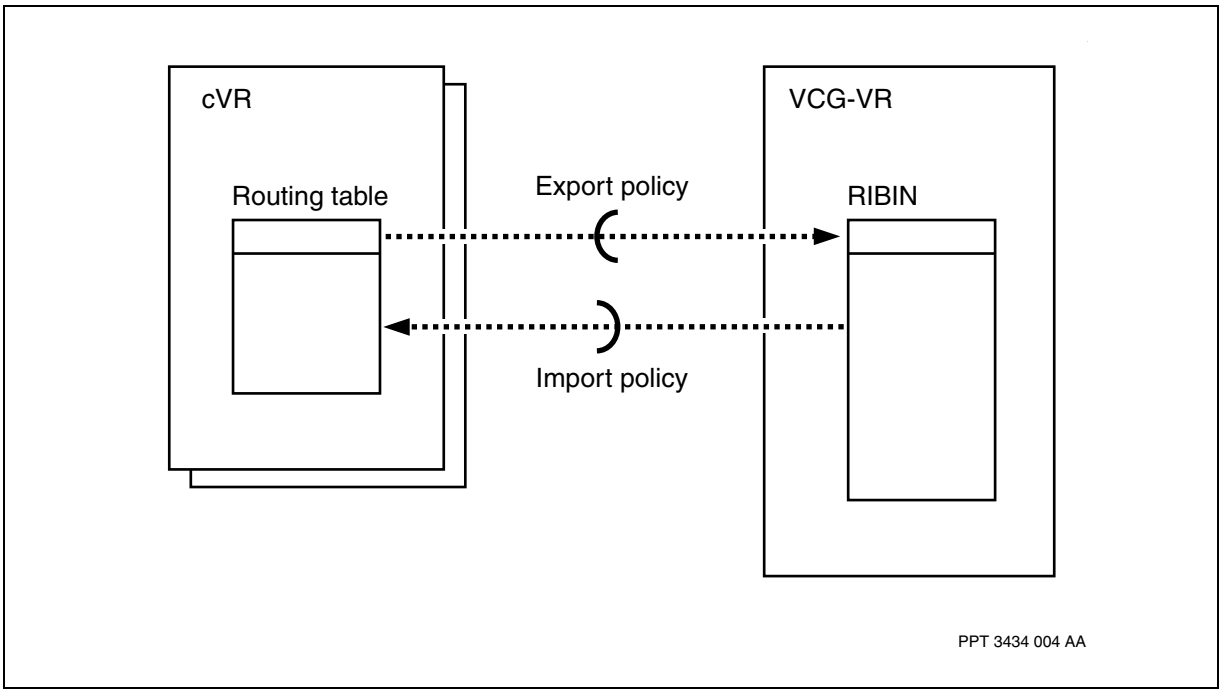
Customer VR import policy

Import policy will allow you to accept or reject mBGP routes based on VCG peer IP address (vcgPeerIpAddress), origin AS number (originAs), ORIGIN (originProtocol), AS_PATH (asPathExpression), COMMUNITIES (communityExpression), network prefix and length (network), and will allow you to modify localPreference and mbgpRtePreference.

Customer VR export policy

Export policy will allow you to export routes from various protocols. Export policy can filter mBGP routes based on the adjacent AS number (bgpAsId), AS_PATH (asPathExpression), COMMUNITIES (communityExpression) RIP interface IP address (ripInterface), RIP neighbor IP address (ripNeighbor), OSPF external route tag (ospfTag), network prefix and length (network), and will allow you to modify localPreference, insertDummyAs, and sendCommunity.

Customer VR MBGP import and export policy



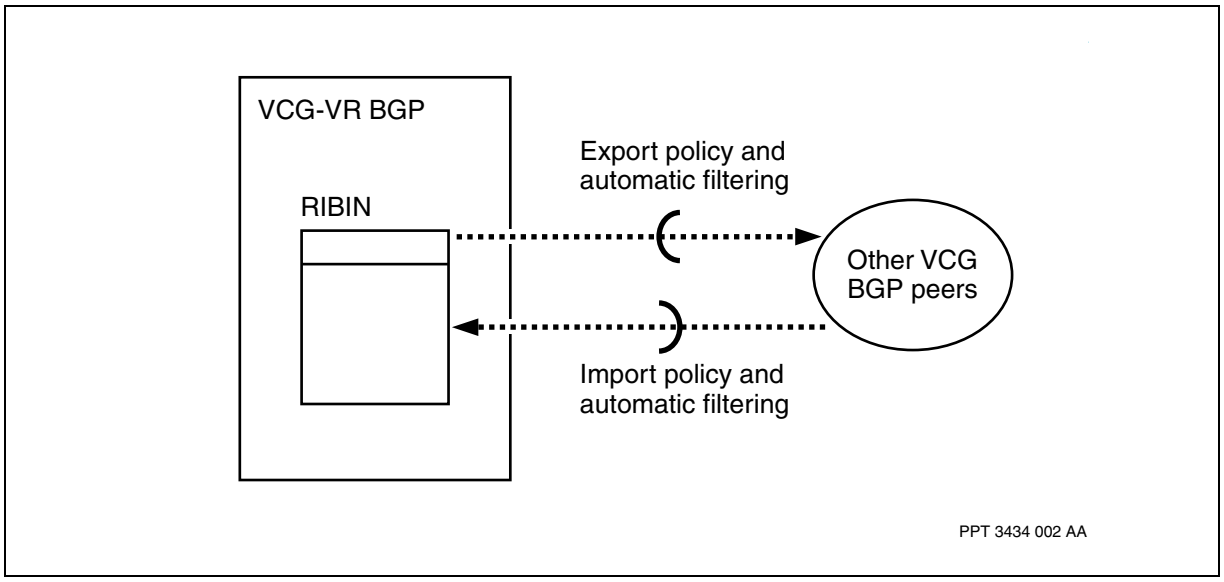
VCG BGP import policy

Import policy will allow you to accept or reject mBGP routes based on address family (addressFamily), peer AS number (peerAs), origin AS number (originAs), peer IP address (peerIpAddress), ORIGIN (originProtocol), AS_PATH (asPathExpression), COMMUNITIES (communityExpression), network prefix and length (network), and will allow you to modify localPreference and appendCommunity.

VCG BGP export policy

Export policy will allow you to export routes from various protocols. Export policy can filter mBGP routes based on protocol type (protocol), peer AS number (peerAs), peer IP address (peerIpAddress), VPN ID (vpnid), adjacent AS number (bgpAsId), AS_PATH (asPathExpression), COMMUNITIES (communityExpression) RIP interface IP address (ripInterface), RIP neighbor IP address (ripNeighbor), OSPF external route tag (ospfTag), network prefix and length (network), and will allow you to modify localPreference, multiExitDisc, insertDummyAs, and sendCommunity.

VCG BGP import and export policy



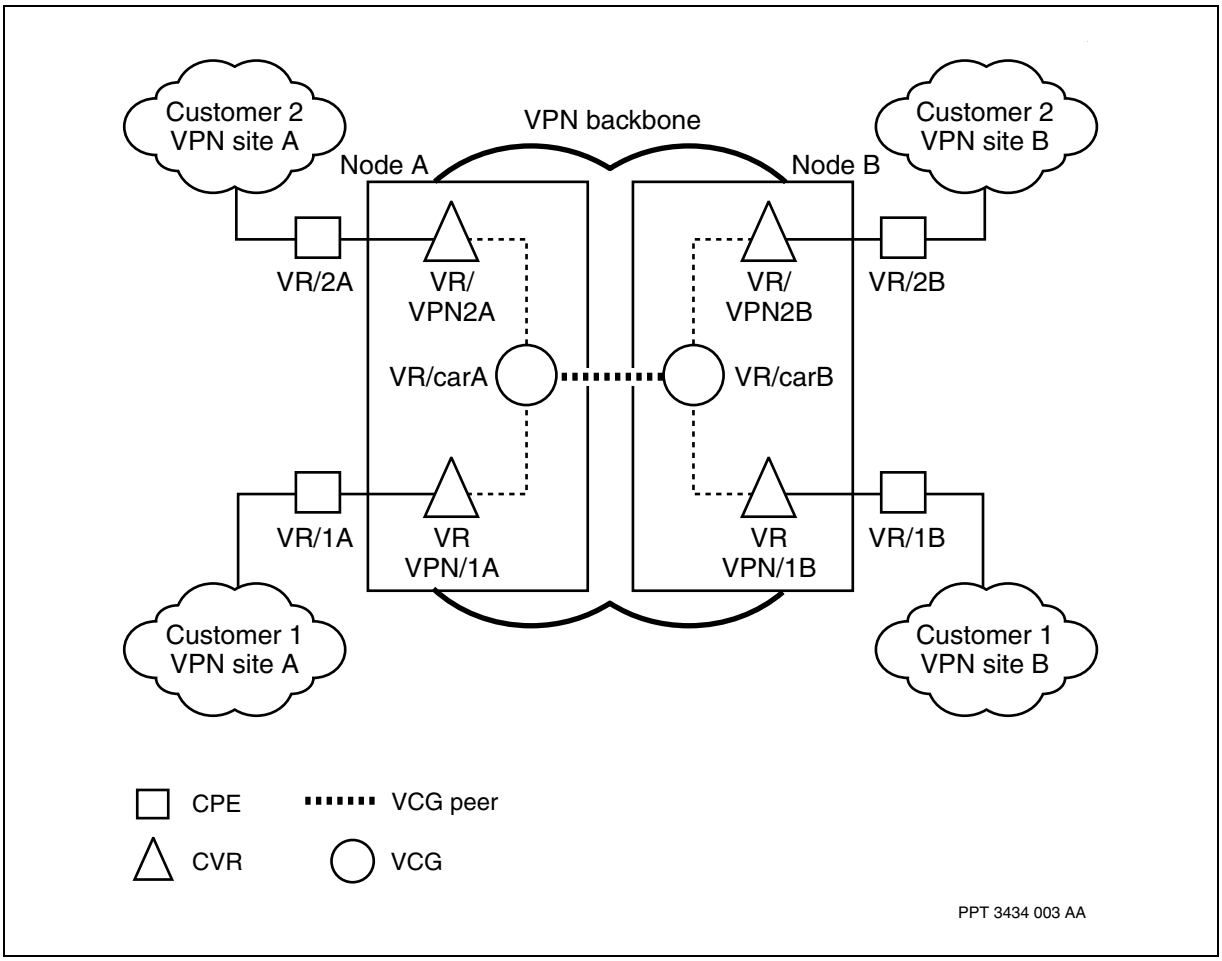
Policy control in multihoming scenario

The following topics are discussed in this section:

Singly-homed stub VPN customer site

A singly-homed stub VPN customer site is a VPN site that only directly connects to one single cVR. See Singly-homed stub VPN customer site (page 43).

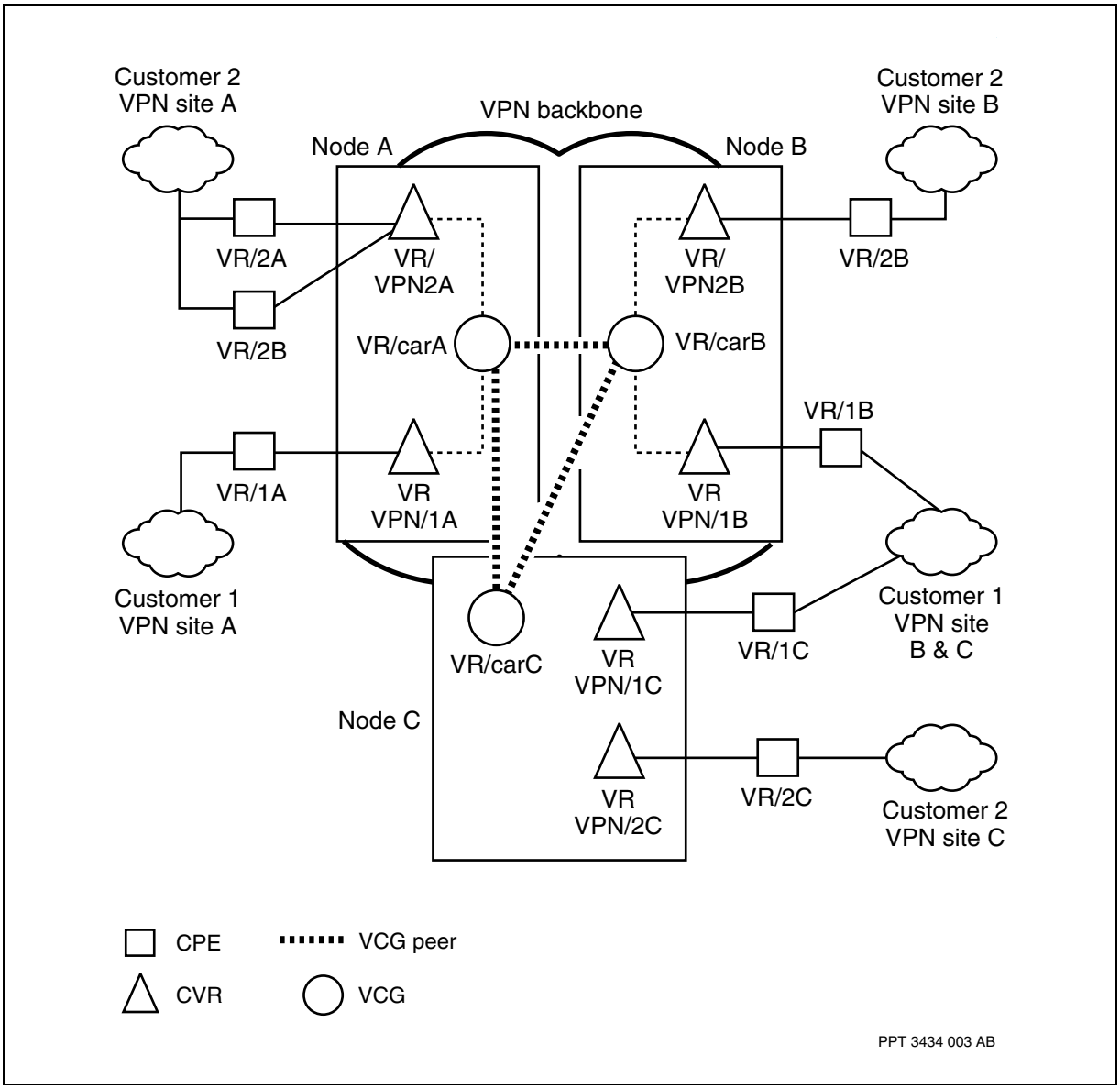
Singly-homed stub VPN customer site



Multi-homed VPN customer site

A multi-homed VPN customer site is a VPN customer site that connects to a VPN backbone via multiple connections. There are two types of multi-homed VPN customer sites: multiple connections to a single customer VR or multiple connections to different customer VRs under different VCGs.

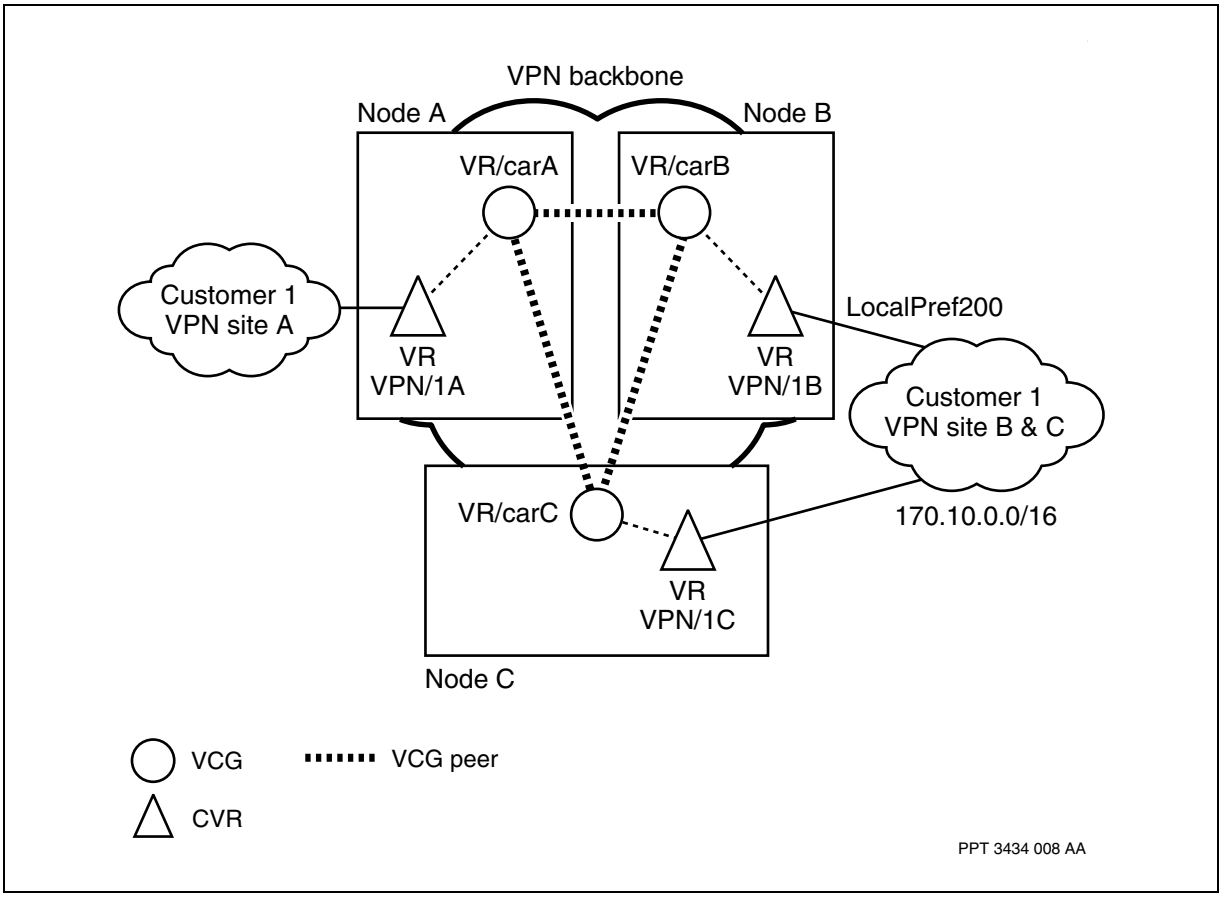
Multi-honed VPN customer site



Local preference

Local preference is a preference value that can be carried within an autonomous system (AS) to indicate which path is preferred to exit the AS to reach the destination network. A path with a higher local preference value is preferred.

Local preference usage to select the best exit point



Remove private AS numbers

A VPN customer site that speaks BGP with its cVR may be assigned a private AS number, which is different from the cVR or VCG backbone AS numbers. When the mBGP routes announce that they have crossed the VCG backbone to the other VPN sites, the private AS numbers are carried in the AS_PATH attribute. The receiving VPN customer site can use the AS_PATH to detect any AS loops.

Attention: Under certain multi-homed VPN scenarios, this is not desirable. The remove private AS numbers feature needs to be configured accordingly.

Route refresh

Route refresh is a mechanism for a BGP speaker to request the re-sending of previously advertised routes from its peers. When route refresh capability is supported between peers, each BGP peer does not need to store the copy of all routes from its peers at all times. A peer can keep only routes that are accepted by its import policy and discard the rest of routes learned from the peer. When an import policy for a peer changes, a peer can send a route

refresh message to receive the re-advertisement of routes, then to evaluate all routes against the new policy. This reduces the memory requirement and the CPU usage, since fewer routes need to be maintained in the BGP's routing database.

A route refresh is a BGP capability that is negotiated during the peer establishment. The capability code 2 and the length 0 is used in OPEN messages to negotiate this capability. In Nortel Multiservice Switch systems, this capability is always negotiated and there is no need to configure the route refresh. When this capability is successfully negotiated between the peers, the value routeRefresh shows up under the Peer component *capabilityNegotiated* attribute.

BGP/MPLS VPN overview

Nortel Multiservice Switch Border Gateway Protocol/Multiprotocol Label Switching (BGP/MPLS) Virtual Private Network (VPN) solution allows Service Providers (SPs) to offer standards-based, low-cost, managed IP VPN services to customers over their existing Multiservice Switch network. By using BGP extensions to distribute VPN routing information, and MPLS to transport data between VPN sites, a SP backbone may be used to provide IP services to multiple VPN sites and customers.

This chapter describes the general operation of the RFC 2547 service. In addition to the general RFC 2547 deployment, there are more specific deployments referred to as Carrier's Carrier and Hub and Spoke, which have unique data paths. See the following sections: BGP/MPLS VPN over Carrier's Carrier MPLS networking overview (page 81) and Hub and Spoke networking overview (page 100).

For information on how to configure a BGP/MPLS VPN, see NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.

Navigation

Main BGP/MPLS VPN components

The five major components required to enable a BGP/MPLS VPN are:

- Provider Edge (PE) node (page 48)
- VPN Routing and Forwarding table (VRF) (page 49)
- Provider (P) router (page 49)

BGP/MPLS VPN network topology (page 50) shows the five components.

VPN site

A VPN site is a set of devices or routers that share IP connectivity. These devices are connected through the same set of Customer Edge (CE) devices to the Provider Edge (PE) node. A VPN connects remote locations of an organization, which share the same policies, over a public network. Each VPN maintains separate routing and addressing information.

Customer Edge (CE) device

A CE device resides in a VPN site and connects to a Provider Edge (PE) node. A CE device allows a local VPN site access to remote VPN sites that belong to the same VPN.

A CE device can connect to a PE node using any number of routing protocols, including RIP, OSPF, BGP, or static routing. Also, different routing protocols can be configured on separate links when two or more CEs connect to the same PE.

Provider Edge (PE) node

A PE node is a router that attaches to one or more CE devices and peers using IBGP with at least one other PE node. The PE node allows remote access to other VPNs that are locally supported by this PE. The PE node keeps track of all VPN routing information, which it learns both locally and remotely. It also acts as a Label Edge Router (LER) device that terminates a Label Switched Path (LSP) tunnel, which is used to forward traffic to other PE nodes. PE node functionality is provided on the CP and on the VPN Extender Card on Nortel Multiservice Switch platforms.

The PE node only knows information about the VPNs to which it is directly attached and it learns local routes from those VPNs. It also learns routes to remote sites that belong to one or more VPN sites to which the PE subscribes.

Attention: In the Multiservice Switch implementation of BGP/MPLS VPNs, a PE node includes two router types: a customer Router that provides VPN Routing and Forwarding table (VRF) functionality and direct connectivity with a CE device, and a Router that aggregates IP traffic from its associated VRFs onto a Gigabit Ethernet or ATM link for transport on the MPLS/IP network. This document refers to a VRF (unless otherwise explicitly stated) as a customer Router instance equivalent operating in the RFC 2547 VPN mode, which includes managing the VPN Routing and Forwarding table. This customer Router is directly connected to a CE router providing a customer site with a VPN connection.

VPN Routing and Forwarding table (VRF)

One or more VRFs reside on a PE node. A VRF table stores and manages VPN routing information.

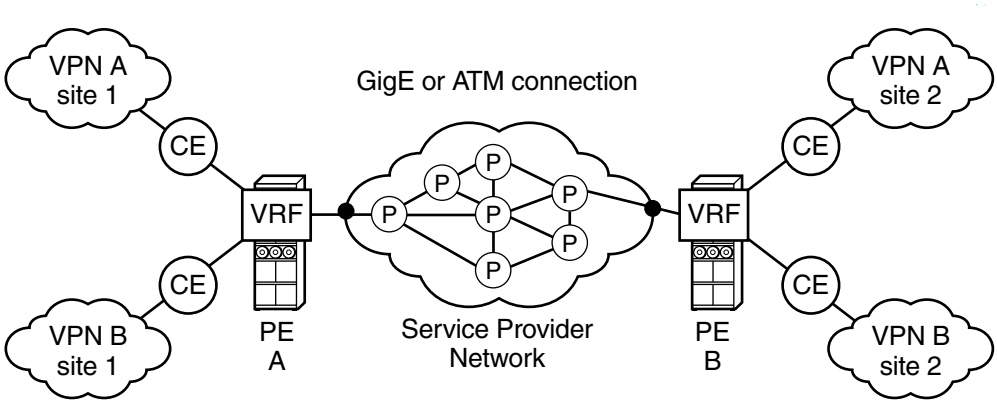
Attention: In the Nortel Multiservice Switch implementation of BGP/MPLS VPNs, a VRF is a customer Router instance that contains VRF tables as specified in RFC 2547 and provides connectivity to one or more CE devices.

Provider (P) router

The P router is a backbone router that provides Interior Gateway Protocol (IGP) connectivity between ingress and egress PE nodes. The P router is not connected to any CE devices and has no knowledge of VPN routing information.

Attention: If a P router is a Route Reflector, it has knowledge of VPN routing information.

BGP/MPLS VPN network topology



CE	Customer Edge
PE	Provider Edge
P	Provider
VPN	Virtual Private Network
VRF	VPN Routing and Forwarding table

PPT 3402 001 AA

Why use Multiservice Switch BGP/MPLS VPN?

Nortel Multiservice Switch BGP/MPLS VPN allows an SP to use an existing Multiservice Switch/IP network infrastructure to transport both public and private data traffic.

See the following sections for more details about the benefits of the Multiservice Switch BGP/MPLS VPN solution:

- Interoperability (page 50)
- Reduced costs (page 51)
- Easy to provision (page 51)
- Scalable and flexible (page 51)

Interoperability

Nortel Multiservice Switch BGP/MPLS VPN solution complements Nortel Networks' existing RFC 2764 Virtual Router VPN solution. For more information, see Direct VR-to-VR connectivity (page 129). By supporting both VPN implementations, Multiservice Switch systems can interoperate with other vendors' VPNs and act as a gateway between RFC 2764 and RFC 2547 networks.

Reduced costs

Nortel Multiservice Switch BGP/MPLS VPN solution allows SPs to leverage existing IP backbone network equipment, or to migrate their existing ATM and frame relay services to Layer 3 BGP/MPLS VPN services. This allows SPs to add value to their customer offerings while keeping capital costs down.

Easy to provision

Nortel Multiservice Switch BGP/MPLS VPN falls into the category of a Provider-Provisioned VPN (PP VPN). The SP's network is transparent to the customer and all provisioning is done by the SP. From a customer perspective, there is no network to provision and maintain as all WAN operations are shifted from the customer to the SP.

BGP/MPLS VPNs leverage dynamic routing protocols. Dynamic routing protocols simplify provisioning as new sites are added to the network. For example, MPLS is used to automatically establish connections between remote VPN sites without any of the sites in the VPN having direct connections to the other sites. With Layer 2 networks such as Frame Relay and ATM, new end-to-end circuits must be provisioned for each new site that is added. Essentially, SPs with BGP/MPLS VPN technology can provide any-to-any VPN site connectivity.

Scalable and flexible

The architecture of a BGP/MPLS VPN provides an efficient way to scale a network, and is flexible enough to accommodate large-scale deployment of VPN services to multiple customers.

As well, Nortel Multiservice Switch BGP/MPLS VPN gives SPs more flexibility in terms of providing another technology choice when planning their network evolution.

How is VPN traffic transported in BGP/MPLS VPN?

From a high-level control perspective, the following two processes occur to allow for data to flow between VPN sites, across the SP's backbone network:

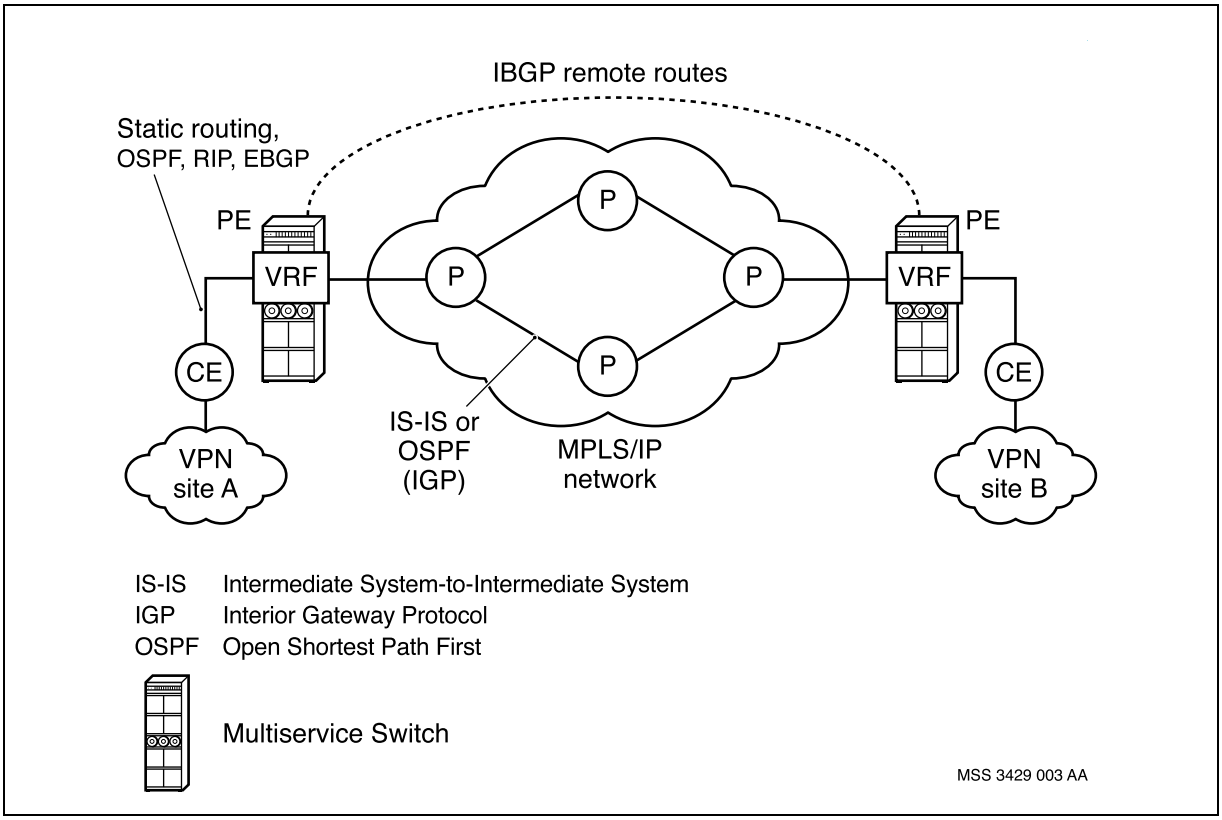
1 Using the access protocol (BGP, OSPF, RIP, or static), routing information is exchanged at the edges of the network between PE nodes and CE devices, and between PE nodes across the SP's backbone network.

2 Using MPLS, LSPs are established between PE nodes across the SP's backbone network.

Once the LSP is established, a host at one VPN site can access a server at a remote VPN site. IP routing between P nodes is handled by the Interior Gateway Protocol (IGP), either Intermediate System-to-Intermediate System (IS-IS) or Open Shortest Path First (OSPF).

BGP/MPLS VPN routing protocols (page 52) shows a high-level view of Nortel Multiservice Switch BGP/MPLS VPN routing protocols.

BGP/MPLS VPN routing protocols



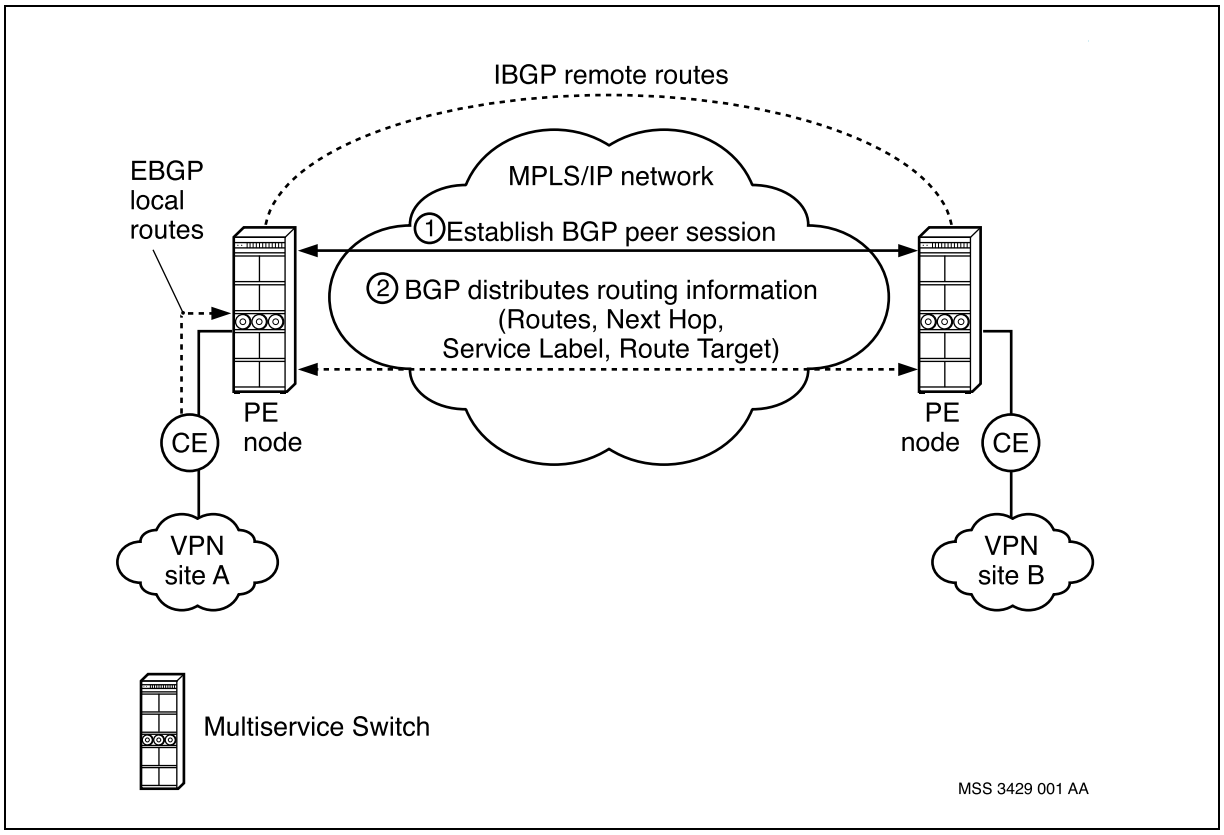
VPN route distribution and routing policy using BGP

BGP is a routing protocol used between and within autonomous systems (AS). An AS is a network or group of networks under a common administration and with common routing policies.

When BGP is used between ASs, the protocol is referred to as External BGP (EBGP). In Nortel Multiservice Switch BGP/MPLS VPN, EBGP can be used to exchange routes between a customer VPN's CE device and the SP's PE node to which it is connected. If an SP uses BGP to exchange routes within an AS, then the protocol is referred to as Interior BGP (IBGP). In a Multiservice Switch BGP/MPLS VPN, IBGP is used by PE nodes to exchange routes across the SP's network. BGP route distribution (page 53) describes the BGP route distribution process.

The following sections contain more information about key BGP route distribution and routing policy functionality in Multiservice Switch BGP/MPLS VPN:

BGP route distribution



About BGP attributes

BGP is a robust and scalable routing protocol. To achieve scalability, BGP uses route parameters, called attributes, to define routing policies and maintain a stable routing environment. BGP peers with other PE nodes in the SP network to advertise routing information and routing updates using information encoded in these attributes.

Route distinguisher (RD)

VPNs can use private addresses, public addresses, or both. Because BGP assumes all addresses it advertises and receives are globally unique addresses, RDs are used to differentiate between identical IPv4 addresses received from different VPNs.

An RD is an eight-byte value that prefixes an IPv4 address. When the RD is added to an IPv4 address, a VPN - IPv4 address is formed. The VPN - IPv4 address, even though it may share the same IPv4 address with a different VPN, is now a unique address in the context of a BGP/MPLS VPN. PE nodes use the RD to convert a received IPv4 address to a unique VPN - IPv4 address before BGP announces it to peer PE nodes and CE devices. The first two bytes encode the Type field and the other six bytes encode the Value field.

An RD consists of an administration field and an assigned number field. The RD is encoded in the NLRI field of the MP_REACH_NLRI path attribute of the UPDATE message. An RD can have two different formats, as follows:

- Type 0 format: administration field contains a 2 byte AS number and the assigned number field contains a 4 byte number assigned by the SP. It is recommended that you use an IANA-assigned non-private AS number. Preferably, the VPN customer's own AS number or the SP's AS number. For example, for an AS number of 100 and an assigned number of 1, the RD would be 100:1.
- Type 1 format: administration field contains a 4 byte ipv4 address and the assigned number field contains a 2 byte number assigned by the SP. It is recommended that you use a globally unicast address, such as the PE's router id or an interface address. For example, for an IP address of 24.24.1.1 and an assigned number of 3, the RD would be 24.24.1.1:3.

One RD must be configured for the VRF and must be unique within the PE for all VRFs. Conceptually, the RD can be configured network wide to be unique per VPN or unique per VRF as long as the above criteria holds.

Attention: When an RD value is changed on a VRF, BGP withdraws all the routes learned by that VRF from its peers and flushes the routes from the Routing Information Base (RIB). BGP then re-installs the routes in the VRF and into the RIB, and re-announces the routes with the new RD value. BGP also installs any already learned routes destined to that VRF in the VRF's routing database.

Route target

Route targets are a form of routing policy used to identify a set of sites within a VPN. BGP uses routes targets to control the distribution of VPN ipv4 routes.Two types of route target exist. When a route is learned from another PE, the "import" route target is used to identify to which VRF that route is

destined. When a route is to be announced to another PE, the “export” route target, associated with the VRF from which the route was learned, is encoded with the route.BGP uses route targets to control the distribution of VPN - IPv4 routes.

A route target is an eight-byte field encoded in a BGP path attribute and communicated to other BGP peers. A route target can have two different formats, as follows:

- Type 0 format: administration field contains a 2 byte AS number and the assigned number field contains a 4 byte number assigned by the SP. It is recommended that you use an IANA-assigned non-private AS number. Preferably, the VPN customer’s own AS number or the SP’s own AS number.
- Type 1 format: administration field contains a 4 byte ipv4 address and the assigned number field contains a 2 byte number assigned by the SP. It is recommended that you use a globally unique unicast address, such as the PE’s router id or an interface address.

Optionally, import and export route targets can be configured for a VRF. Without a route target provisioned for the VRF, no remotely learned routes are installed in the VRF and no locally-learned routes in the VRF are advertised across the backbone.

Attention: As with RD values, when a change is made to the route target value BGP withdraws from its peers all the routes learned from that VRF and flushes the routes from the RIB. BGP then re-installs the routes in the VRF and into the RIB, and re-announces the routes with the new export route target.

Routing policy

A routing policy consists of a set of values that is used to match specific criteria when making routing decisions. Policy can be used, for example, to balance VPN traffic among different VPN sites across the SP network, control the route selection of CEs, or both.

The following BGP routing policies are supported:

- Inbound route filtering (page 56)
- VRF export policies (page 56)
- BGP export policies (page 57)

Attention: Only VRF export—not import—policies are supported. The default behavior is to accept all routes from the RIB destined for the VRF based on the VRF's import route targets. No additional filtering is provided.

Inbound route filtering

When a PE receives a BGP update message containing remotely-learned routes, the routes are subject to inbound filtering. During the filtering process, if a learned route is found not to contain at least one locally supported route target, it is discarded. In other words, if the PE node does not contain a VRF instance that has at least one import route target in common with the route targets found in the BGP update message, the route is discarded.

VRF export policies

The VRF export policy is a set of rules to be applied to decide if certain locally learned routes should be installed in the BGP Mpls Vpn Routing Information Database (MvRib) and further advertised to other PE nodes as well as other VRFs on the same PE node. A policy consists of attributes which are used as matching criteria, attributes which are used as set criteria (or outputs), and an action (send or block). The set criteria attributes are only applied if a "send" policy is matched.

The following is a list of attributes available as match criteria:

- protocol
- asPathExpression
- communityExpression
- expressPreference
- network

The following is the only attribute available as set criteria:

- localPreference

The following attribute specifies the action to take on a policy match:

- advertiseStatus (send or block)

Export policies can be used to filter routes. When a policy mode is set to block, a matching route is not advertised. Given that the default behavior of the VRF is to allow everything, a policy can be used to block specific subnets and to advertise everything else.

If an Export policy component is not configured under the VRF component, the default behavior is to install all locally learned IPv4 routes into the BGP MvRib to be further advertised to the remote peers and other local VRFs. The routes will only be advertised to other local VRFs if the Import Route Target of those VRFs match the Route Targets of the routes.

BGP export policies

A BGP export policy can be used to decide if certain VPN IPv4 routing information is blocked or advertised to its peers. If no export policy is provisioned, BGP distributes all VPN IPv4 routes to its peers. You can decide to filter VPN IPv4 routes based on the values of the network address, the remote peer IP address, and the route target in order to eliminate unnecessary advertisement of VPN IPv4 routes by BGP speakers.

If multiple export policies are provisioned, the one with the most specific match applies. Here is the list of attributes or components from the highest precedence to the lowest precedence:

- *RouteTarget* attribute
- *peerIpAddress* attribute
- *Network* component

Loopback address

The loopback address resides on a virtual or non-physical interface on the RTR. Both BGP and MPLS use the loopback address to set up the LSP required to carry VPN traffic across the network.

Route selection

No route selection is performed at the Router by BGP for BGP/MPLS VPN. Routes with the same RD and prefix learned from different peers are installed in the RIB and distributed to the appropriate VRFs.

When the VRF learns more than one VPN-IPv4 route to the same destination, a route selection algorithm is run to determine the best VPN-IPv4 route. Route selection is used to determine the best route within the same routing protocol. The routing protocol is bgpMplsInternal for both remote and local VPN-IPv4 routes. Route preference is used as one of the criteria in the route selection algorithm to determine which VPN-IPv4 route is more preferred.

The route selection algorithm for VPN-IPv4 learned routes is as follows:

1 Choose the route with the highest local preference.

2 If local preference is the same, choose the route with the shortest AS PATH.

3 If the AS PATH length is the same, choose the route with the lower MED value.

4 If the MED value is the same, choose the VPN-IPv4 route with the lower route preference value.

5 If the route preference value for the VPN-IPv4 route is the same, choose the route with the lower peer Router ID.

6 If the peer Router ID is the same, choose the route with the lower peer IP Address.

7 If the peer IP Address is the same, choose the route with the lower decimal RD value.

It is possible that the same route was learned by different routing protocols. Once the best VPN-IPv4 route is chosen based on route selection, the route preference value of the chosen VPN-IPv4 route is used to compare against the route preference value of the route learned via another protocol to determine the overall “best” route.

All routes to the same destination that are learned are kept in the VRF routing table in case the best route becomes unreachable. If that happens, then a new best route is chosen and forwarding continues based on the new route.

Route preference

When multiple routes to the same destination are learned via different protocols, the route preference value is used to determine the “best” route to install in the VRF forwarding table. For VPN-IPv4 routes, the route preference is also used as one of the criteria to choose the best VPN-IPv4 route as explained in Route selection (page 57).

The route preference for routes imported from the BGP MvRib can be configured for each VRF individually. When the route preference value is changed, the VPN-IPv4 routes installed in the VRF routing table will be re-evaluated based on the new value.

For remote VPN-IPv4 routes learned by MP-BGP, the protocol is set to ‘bgpMplsInternal’. The protocol is also set to ‘bgpMplsInternal’ for routes learned from other local VRFs on the same PE.

The default value for VPN-IPv4 routes learned from other PE nodes is set to 127, making these routes least preferred over other protocols and over VPN-IPv4 routes learned from other VRFs on the same PE node. The default value for VPN-IPv4 routes learned from other VRFs on the same PE node is set to 125, making these routes least preferred over other protocols, but more preferred over VPN-IPv4 routes learned from other PE nodes.

The default route preference values for some of the protocols are as follows: local (0), OspfInternal (30), bgpExternal (70), staticRemote (72), OspfExternal Type1 (80), Rip (82), OspfExternal Type2 (120), bgpInternal (122), mbgp (123), bgpMplsInternal (125 for locally learned VPN-IPv4 routes or 127 for remotely learned VPN-IPv4 routes).

The lower the preference value, the more preferable the route.

Attention: If the same route is learned via aggregate policy and via BGP/MPLS VPN, the aggregate route will be preferred as it has a lower preference value.

Default route preference values

Default route	Preference value
Local, static discard	0
MPLS	10
Reserved for RIP migration	20
Reserved for RIP migration	21
OSPF internal	30
OSPF external	70
Static remote	72
OSPF external type 1	80
RIP	82
OSPD external type 2	120
BGP internal	122
mBGP	123
BGP aggregate	124
bgpMplsInternal (local)	125
bgpMplsInternal (remote)	127
Reserved for internal use	254
UN_PREF	255

Forwarding classes and route preferences

Each route preference belongs to one of three forwarding classes. The lower the value of the forwarding class, the more important the routing information. The table Forwarding classes (page 60) shows how a route preference is assigned a forwarding class. It is recommended that whenever you change a route preference, you keep the new value within its existing forwarding class.

The forwarding class has an effect on routing. Specifically, a new more specific route can only be inserted into the forwarding table as long as less specific routes do not exist in a more important forwarding class.

For example: Assume an OSPF internal route (forwarding class 0) exists for address 10.1.0.0/16 and BGP internal (forwarding class 1) finds a route for address 10.1.1.0/24. The BGP route, which is more specific, is not allowed in the forwarding table because its forwarding class is less important than the forwarding class of the OSPF route.

Forwarding classes

Route preference range	Forwarding class
0-63	0
64-127	1
128-255	2

Changing default route preference values

Although route preference values are configurable by the user, it is important to be aware of the implications of modifying a route preference which in turn result in a change in forwarding class.

For example: consider the scenario where Local Static route address is 10.0.0.0, and BGP_MPLS_INTERNAL routes addresses are 10.56.133.0 and 10.224.133.0. Additionally, the user has changed the BGP_MPLS_INTERNAL pref value to 130 from its default of 127.

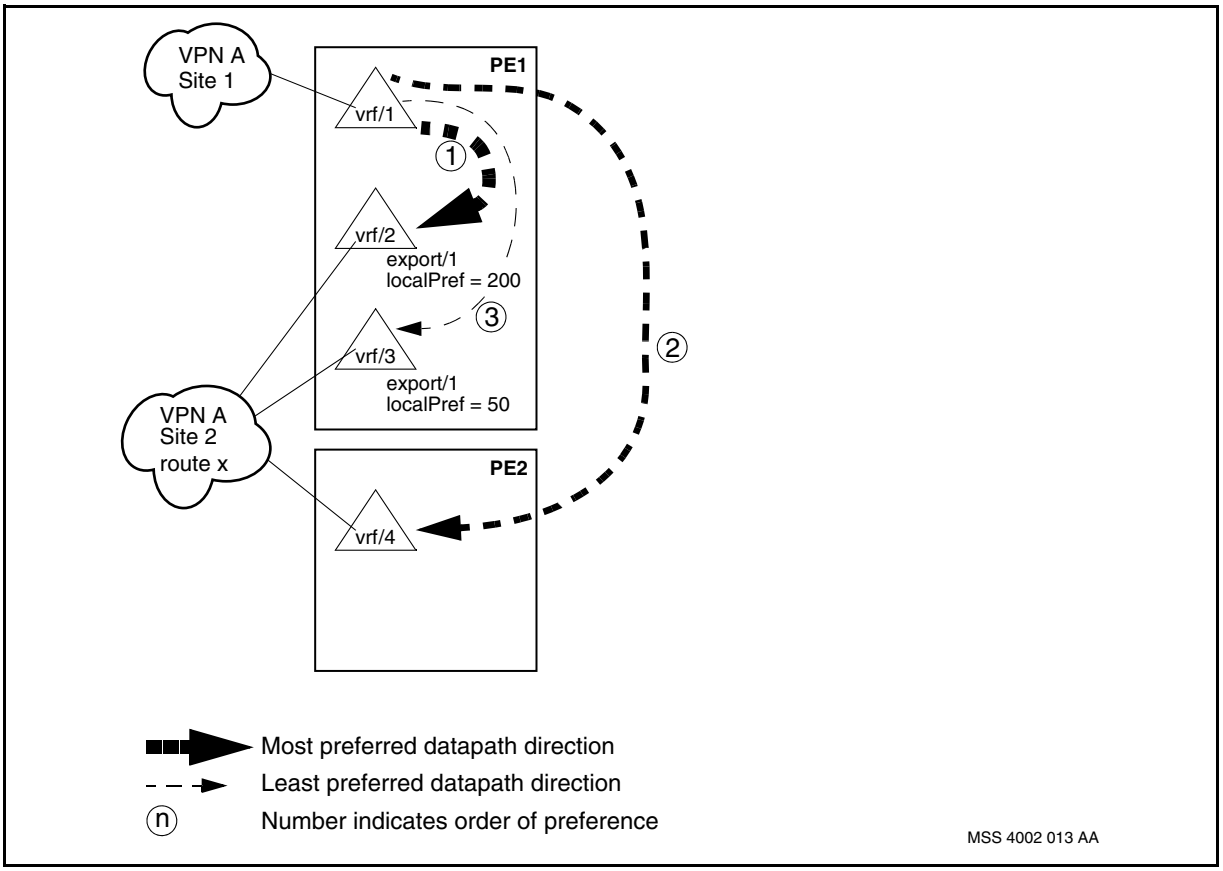
When remote BGP_MPLS_INTERNAL routes are added to the VRF RDB, they have a preference value of 130, while the local STATIC route has a preference value of 72. The BGP_MPLS_INTERNALS route, which is more specific, is not allowed in the forwarding table because it belongs to a less important forwarding class than that of the Local Static route. As a result, you will be left only with the less-specific static route in your forwarding table.

How to manage VPN-IPv4 preferences via export policy

By default, VPN-IPv4 routes learned from local VRFs are always preferred over VPN-IPv4 routes learned from a remote PE node.

To manage the preference of bgpMplsInternal routes that are learned between different VRFs on the same PE node, the local preference can be used in the VRF export policy. Local preference is a preference value that can be carried within an AS to indicate which path is preferred to exit the AS in order to reach the destination network. A path with a higher local preference value is more preferred. For example, consider the scenario depicted in the figure Managing VPN-IPv4 preferences via export policy (page 61). VRF/1 learns route X from VRF/2, VRF/3, and VRF/4. If the same route is advertised by two different local VRFs, VRF/2 and VRF/3, and one remote PE2 node, and it is desirable to prefer the path through VRF/2, then provision a VRF/2 and a VRF/3 export policy and set the localPreference to be higher for VRF/2. Note that since the default local preference value is 144 and that VRF/4 does not have an export policy provisioned, then the route learned from VRF/3 is more preferred over the route learned from VRF/3.

Managing VPN-IPv4 preferences via export policy



How to manage VPN-IPv4 preferences via preference values

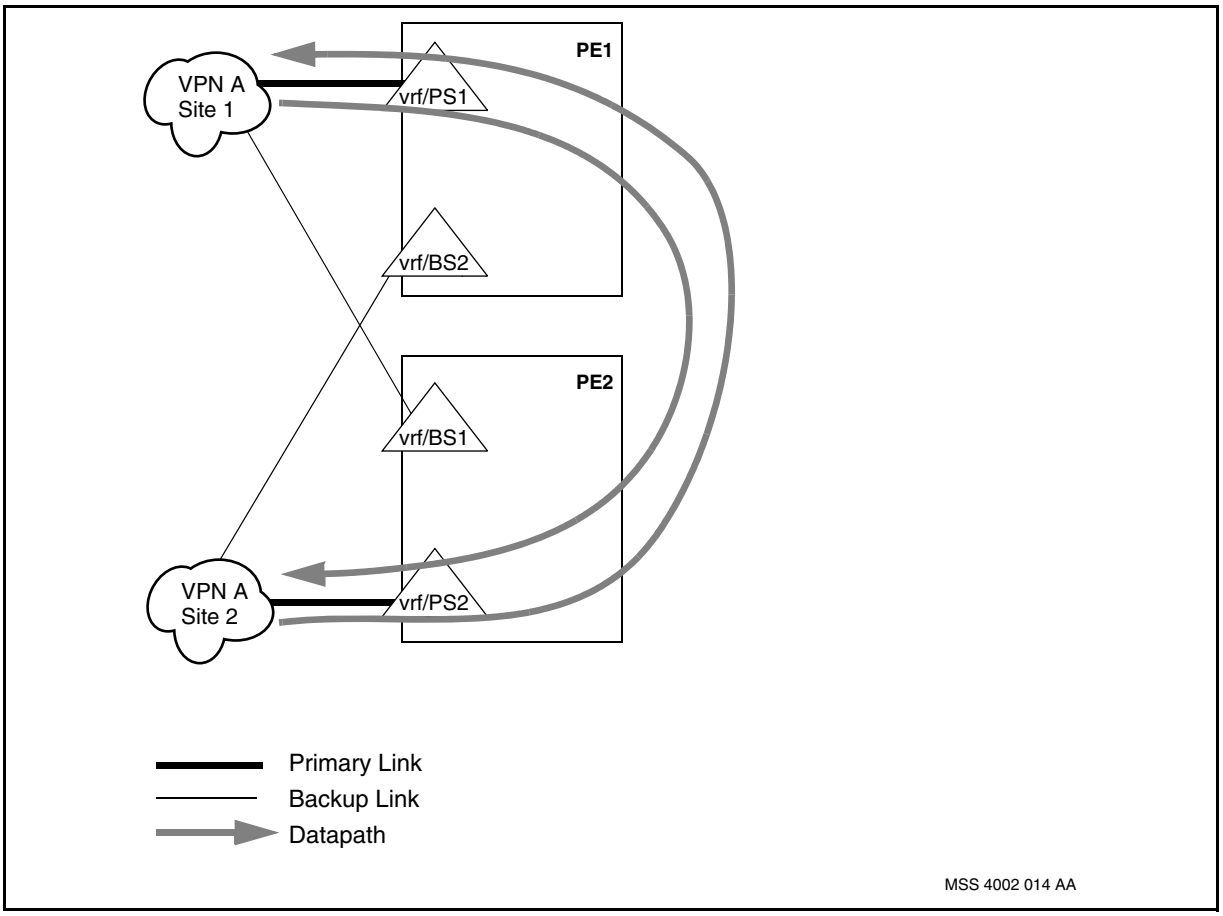
By default, VPN-IPv4 routes learned from local VRFs are always preferred over VPN-IPv4 routes learned from a remote PE node.

To manage the preference of bgpMplsInternal routes learned from another PE node over those bgpMplsInternal routes learned via a local VRF, the route preference can be used.

The ability to set a route preference value allows more flexible service deployment in situations where multiple paths exist to the CE router. For example, consider the scenario depicted in figure Managing VPN-IPv4 preferences via route preferences values (page 63). The route advertised by Site1 is learned by VRF/PS2 both via the local VRF/BS1 and the remote PE node. It is desirable that traffic originating from Site2 and destined to Site1 use Site1's primary link. In order for this to happen, VRF/PS2 needs to prefer the route learned from the remote PE node over the route learned from the local VRF/BS1. This is achieved by setting the route preference value of remotely learned VPN-IPv4 routes under VRF/PS2 to a value lower than that of the route preference value of locally learned VPN-IPv4 routes in order to make VPN-IPv4 routes learned from the remote PE node more preferred over those learned from the local VRF.

Similarly, the route advertised by Site2 is learned by VRF/PS1 both via the local VRF/BS2 and the remote PE node. It is desirable that traffic originating from Site1 and destined to Site2 use Site2's primary link. In order for this to happen, VRF/PS1 needs to prefer the route learned from the remote PE node over the route learned from the local VRF/BS2. This is achieved by setting the route preference value of remotely learned VPN-IPv4 routes under VRF/PS1 to a value lower than that of route preference value of locally learned VPN-IPv4 routes in order to make VPN-IPv4 routes learned from the remote PE node more preferred over those learned from the local VRF.

Managing VPN-IPv4 preferences via route preferences values



Route distribution between BGP/MPLS network elements

The following sections contain the steps involved in distributing routes among the network elements in a BGP/MPLS VPN:

- Route distribution: CE to PE (page 63)
- Route distribution: between PE nodes (page 64)
- Route distribution: PE to CE (page 64)

Route distribution: CE to PE

The following are the events that occur when a PE learns a route from a CE:

1 The CE advertises its routes to the PE. These routes are learned through an access routing protocol such as OSPF, RIP, or EBGP, or by static routing.

2 If an import policy for the access routing protocol is configured (in the case of RIP and EBGP), some routes may be filtered before they reach the PE.

3 Routes that pass the import policy, if one exists, get installed in the VRF's routing table as routes learned by means of that particular access routing protocol.

4 As a route to the same destination may be learned from different protocols (in other words, through different interfaces), only one of these routes is chosen based on the route preference. The best route chosen is installed in the VRF's forwarding table. For information about route preference, see Route preference (page 58).

5 A newly-learned IPv4 route that passes the VRF export policy is installed in the RTR's RIB as a VPN - IPv4 route, by adding the VRF's RD as a prefix to the route.

6 BGP advertises the newly-learned VPN - IPv4 routes to its peers using the multi-protocol extensions for BGP and associating the appropriate VRF's export route target(s), MPLS Service label, and RD with those routes.

Route distribution: between PE nodes

The following are the events that occur between PE nodes:

1 When a BGP speaker distributes a VPN - IPv4 route to its peers, it assigns an MPLS Service label (see About MPLS labels (page 67) for more information) to the route. This route is referred to as a "labeled" VPN - IPv4 route.

2 BGP-4 with Multiprotocol Extensions (MP - BGP) is used to negotiate and distribute labeled VPN - IPv4 routes across the backbone between RTRs. RTRs exchange BGP messages containing encoded attributes, which announce routing information, including reachable, labeled VPN - IPv4 routes. For information, see BGP peer session establishment and capabilities negotiation (page 70).

3 The ingress PE would then use the encoded route target(s) to match against all the import route target(s) to determine which VRFs should learn these routes.

Route distribution: PE to CE

The following are the events that occur when a PE receives a route update by means of IBGP from another PE:

1 A labeled VPN - IPv4 route is learned by BGP on the RTR interface.

2 The BGP message containing the encoded attributes is parsed and subjected to inbound route filtering. The PE's VRFs are scanned to see if any of their import route target(s) match the learned route's export route target(s). If there is no match, the route is discarded; otherwise, the route(s) is installed in the RIB as a remotely-learned route.

3 For each VRF that has at least one of the remotely-learned routes' route targets configured as an import route target, the VPN - IPv4 route is installed in that VRF's routing table as a IPv4 route.

4 If needed, route selection at the VRF is performed (see Route selection (page 57) for more information).

5 VRF uses its access protocol's export policy to redistribute learned routes to the CE using the appropriate routing protocol running between the PE and CE.

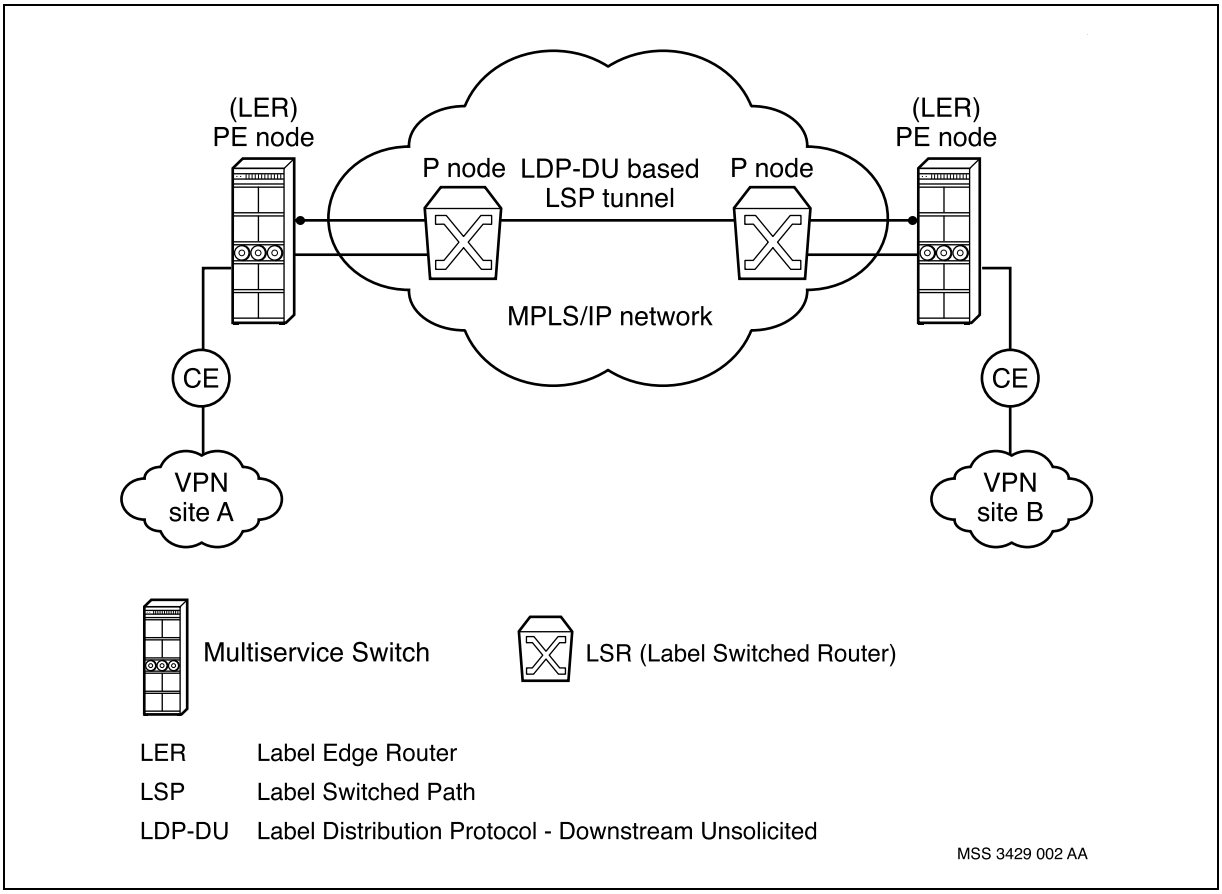
Forwarding VPN traffic using MPLS

As shown in MPLS LSP establishment (page 66), MPLS employs the Label Distribution Protocol, operating in Downstream Unsolicited mode or LDP-DU, to create LSP tunnels. MPLS-based LSP tunnels provide a secure means by which VPN traffic can be transported across the SP network. LSP tunnels associate sets of packets to an MPLS label. LSP tunnels interconnect ingress and egress PE nodes through one or more P nodes.

See the following sections for more details about MPLS functionality in Nortel Multiservice Switch BGP/MPLS VPN:

- About Label Distribution Protocol - Downstream Unsolicited (LDP-DU) (page 66)
- About MPLS labels (page 67)
- Traffic forwarding: ingress PE node (page 67)
- Traffic forwarding: P node (page 67)
- Traffic forwarding: egress PE node (page 67)

MPLS LSP establishment



About Label Distribution Protocol - Downstream Unsolicited (LDP-DU)

LDP-DU is used for LSP set-up, maintenance, and tear-down. LDP-DU handles these functions through a series of protocol-specific messages exchanged between LDP peers.

During LSP set-up, LDP-DU is used by P nodes (known as Label Switched Routers or LSRs in MPLS terminology), to exchange information about the meaning of labels used to forward traffic. In Nortel Multiservice Switch implementation of BGP/MPLS VPNs, LDP-DU assigns and distributes labels before VPN traffic is transported across the network. That is, LDP-DU sets up transport LSPs first so that when VPN traffic arrives at a backbone node, it can be label-swapped immediately and mapped onto the appropriate LSP.

To map packets to the appropriate LSP, LDP-DU also associates a Forwarding Equivalence Class (FEC) with each LSP during the LSP set-up process. LSPs are extended across the network as each LSR in the backbone associates incoming labels for a FEC to the outgoing label assigned to the next hop for a given FEC.

About MPLS labels

MPLS provides label switching between VRFs on ingress and egress PE nodes. There are two MPLS labels—Transport and Service—and each performs a different role.

The Transport label is the outer label. It directs the packet to the correct PE router. It is associated with a BGP next hop and uniquely identifies an LSP.

The Service label is the inner label. It determines how the PE router should forward the packet to the CE router. A Service label is associated with a VRF and is unique for each PE node. This label is used to advertise routes learned by the particular VRF.

Labels are “pushed” onto, or “popped” from, packets as the various network elements in a BGP/MPLS VPN forward and receive, respectively, VPN traffic.

Traffic forwarding: ingress PE node

The following occurs when an ingress PE node receives a packet from a VPN on one of its VRF instances that is destined for a remote VPN site:

1 The PE obtains the Service label (advertised by the VRF on the egress PE and associated with the particular route to the remote VPN site), the BGP next hop (the primary loopback address of the egress PE), the LSP to use, and the Transport label associated with that LSP.

2 The PE pushes the Service labels onto the packet.

3 The PE forwards the packet through its RTR interface to the first P router along the LSP from the ingress to the egress PE.

Traffic forwarding: P node

The P nodes label-switch received traffic from an incoming label to an outgoing label, and forward packets across the backbone based on the Transport label. When a P node receives a packet through an LSP, the following occurs:

1 If the P node is not performing penultimate label popping (PLP), it pops the Transport label and pushes a new Transport label (associated with the LSP) onto the packet. If the P node performs PLP, it pops the Transport label and exposes the Service label.

2 The P node forwards the packet to the egress PE, which may or may not involve another P node.

Traffic forwarding: egress PE node

The following occurs when an egress PE receives a VPN packet:

1 The egress PE pops the MPLS Service label. (There is at most one Service label.)

2 The associated VRF does a route lookup and forwards the packet to its destination.

Attention: To support the equal cost multi-path (ECMP) on egress, all access links on the associated VRF must reside on the same functional processor (FP).

Control flow

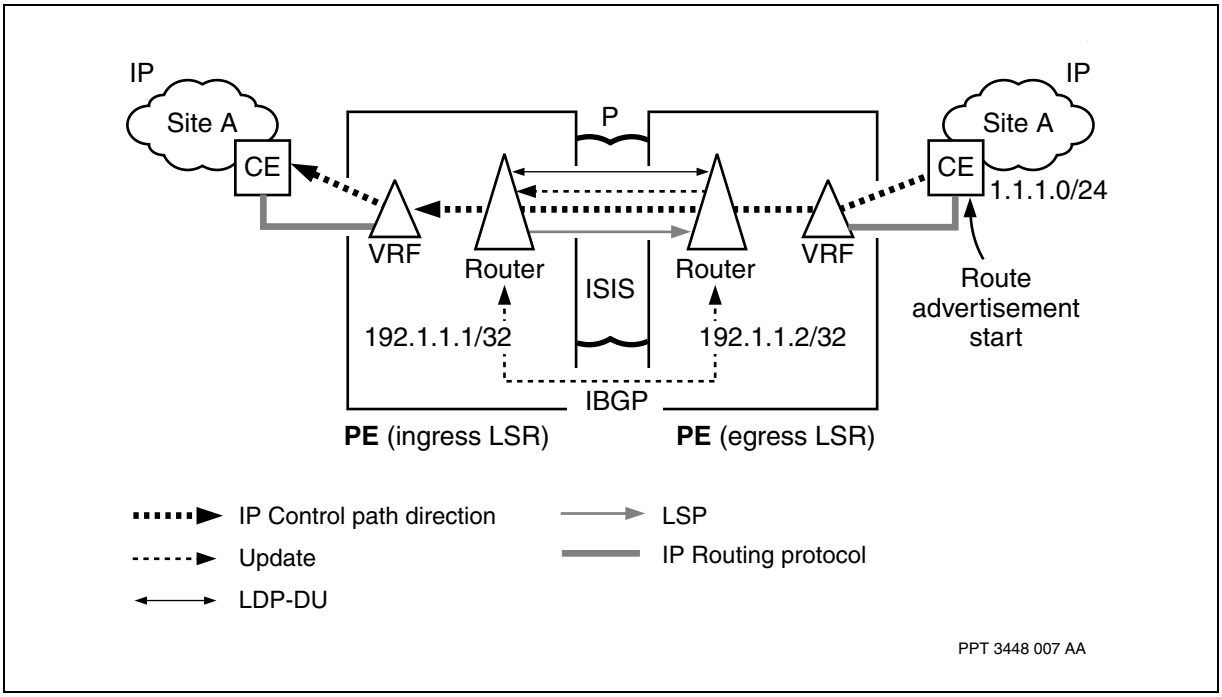
The control flow for RFC 2547 VPNs involves both MPLS control plane and IP control plane. In the scope of IP control plane, BGP is used to distribute VPN routing information between PE nodes. In the scope of MPLS plane, Label Distribution Protocol in Downstream Unsolicited mode (LDP-DU) is used to establish LSP tunnels between PE nodes that can be used to transport VPN traffic.

This section describes the operation of the IP control plane and indicates how the MPLS control plane operation fits into the BGP/MPLS VPN solution.

To achieve the setup of a datapath between CE nodes, the LSP must use the same information BGP is using when new VPN routes are learned. In particular, the BGP next hop is used as the transport layer next hop for VPN data traffic. To achieve this cooperation, both MPLS and BGP use the same always up virtual media interface on the router. These virtual media always up interfaces are referred to as primary loopback addresses. The primary loopback address is an always up interface on a Nortel Multiservice Switch node. It is referred to as primary since it represents a single instance of a loopback address on a Multiservice Switch Router component for use by applications on that router.

High level control flow diagram (page 69) shows a high level view of the control flow in setting up the datapath. In this figure, there is an ingress PE, an egress PE, and the CEs that are attached to them. IBGP is running a session between the ingress PE Router and the egress PE Router. The primary loopback address for the ingress PE is 192.1.1.1/32 and the primary loopback address for the egress PE is 192.1.1.2/32. The IBGP session is using these primary loopback addresses.

High level control flow diagram



Additionally, LDP-DU runs between the two Router nodes. LSP tunnels are unidirectional connections. To enable a datapath in the direction from ingress PE to egress PE, LDP-DU uses primary loopback address value of the egress PE as a FEC of an LSP. To enable bidirectional communication, the process is repeated in the opposite direction. High level control flow diagram (page 69) shows an LSP tunnel established from ingress PE to egress PE.

High level control flow diagram (page 69) shows a route 1.1.1.0/24 being learned by the egress PE VRF through some IP routing protocol. This route is then distributed using IBGP across the SP backbone and passed to the ingress PE node. Here, the route is distributed to any VRF, which is in the same VPN. When the route is installed into a VRF, it is subject to route distribution by means of the supported access routing protocol. In the above diagram, the route is distributed into Site A.

Once the transport LSP is available, and the remote site's address has been learned by the CE on the ingress PE side, the customer traffic from Site A may start flowing, being tunneled over the LSP, toward the advertised address in Site B.

At any given time, there may be more than one LSP to the BGP Next Hop. Only one of these LSPs is primary (capable of transmitting customer traffic) at any one time. Others, referred to as standby LSPs, are established but not used for the transfer of dataflow at the moment. It is transparent to BGP which LSP is currently chosen to be primary. BGP updates are only sent and

forwarded to the appropriate VRFs when the corresponding LSP is up. When the LSP is not available for the BGP Next Hop, the routes associated with the BGP Next Hop are unreachable. All unreachable routes are retained in the RIB. When the LSP does come up, these routes are distributed as detailed above.

Handling backbone network topology changes

The CE devices communicate with other CE devices through the VRF. The VRF is configured to handle dataflow across the backbone through a transport LPS. When a backbone topology change occurs, the IGP running in the core may decide the best path to a PE peer has changed. In this case, the IGP notifies MPLS that any LSPs that may be using this information should be reevaluated. At this time, MPLS may decide to change the active LSP. The liberal retention mode that MPLS uses in the BGP/MPLS VPN solution allows it to retain multiple transport labels to any given FEC. This capability combined with the use of IGP's "shortest path" algorithm allows MPLS to select which transport LSP to use.

The control path updates the transport LSP to use by all VRFs which are currently using the affected LSP. Depending on the extent of the IGP change, some data loss may be present. If core connectivity is interrupted, CE to CE dataflow is briefly interrupted also.

BGP peer session establishment and capabilities negotiation

When a BGP peer is configured with an addressFamily of ipv4MplsVpn, BGP initiates a TCP session with the peer using the *localAddressConfigured* (LAC) as the source IP address and the *peerIpAddress* as the destination IP address for the TCP connection. The LAC is defaulted to the primary loopback address of the PE router.

Once the TCP connection is established, BGP proceeds to establish a connection with the peer. This is done by negotiating the BGP speaker's capabilities in the OPEN message. To support BGP/MPLS VPNs, the capability to support multi-protocol extensions for BGP is negotiated. The BGP router ID used in the OPEN message is based on the same algorithm used for regular BGP. The only other address family that can be simultaneously supported, and hence negotiated, with the ipv4MplsVpn address family is ipv4Unicast.

The peer session is considered established once an agreement on the capabilities is reached. At this point, BGP can perform a database exchange between the peers using the UPDATE message. VPN reachable routes are encoded in the MP_REACH_NLRI path attribute of the UPDATE message. The BGP next hop used in the message is the primary loopback address of the PE router and is encoded with RD equal to zero plus the ipv4 loopback

address. Once all routes are exchanged among all BGP peers, the network is said to have converged. BGP now only sends updates to any new or removed routes.

Attention: The underlying MPLS LSP tunnel is not required to be up to enable the exchange of routing information between BGP peers. The MPLS LSP tunnel, however, must be up before any routes learned by the BGP peers can be installed in the VRF routing table and before any routes in the VRF forwarding table can be installed in the Router RIB.

Route distribution between PE routers

When a BGP speaker distributes a vpn-ipv4 route, it assigns an MPLS service label to the route where the MPLS service label is associated with the VRF. This route is referred to as a labeled vpn-ipv4 route.

BGP4 with Multi-Protocol extensions (MP-BGP) is used to distribute labeled vpn-ipv4 routes across the backbone between RTRs. There is a single Router instance on a PE supporting BGP/MPLS VPNs. Once BGP peers negotiate the address family of 1/128, they begin to exchange BGP/MPLS VPN UPDATE messages.

The routing information is encoded in the MP_REACH_NLRI, MP_UNREACH_NLRI, and EXTENDED_COMMUNITY path attributes of the UPDATE message. The MP_REACH_NLRI attribute is used to announce reachable labeled vpn-ipv4 routes. The MP_UNREACH_NLRI attribute is used to withdraw previously announced labeled vpn-ipv4 routes that are no longer reachable. The EXTENDED_COMMUNITY attribute is used to encode the export route target(s) common to the set of labeled vpn-ipv4 routes. The ingress PE would then use the encoded route target(s) to match against all the import route targets to determine which VRFs should learn these routes.

The label referred to above is called the service label. The service label is used to identify the VRF which advertises this routing information. Since the VRF may span several access cards, a unique service label is created per VRF per access card on which it resides. This label is used to advertise routes learned by that VRF on the corresponding card.

Sending an UPDATE message

Sending an UPDATE message is triggered in the following cases:

- when the VRF instance on the PE learns a new route from the CE, provided the export policy allows the route advertisement
- when it is determined that a previously learned route from the CE is no longer available

- when the export policies are modified and routes are either no longer valid or new routes need to be advertised
- when a VRF is created or deleted
- there is an OSPF change
- a route refresh request occurs

In the case that a new route is learned and installed in the associated VRF as an ipv4 address, if the export policy allows the route, the route is redistributed into the RIB as a vpn-ipv4 address. This is done by prefixing the ipv4 address route with the RD associated with the appropriate VRF. BGP encodes the MP_REACH_NLRI path attribute, among other things, with AFI/SAFI 1/128, the label associated with the VRF, the RD, and the ipv4 address. BGP also encodes the EXTENDED_COMMUNITY path attribute with the export routes targets associated with the VRF. Multiple VPN routes with common path attributes (other than MP_REACH_NLRI) can be carried in a single UPDATE message. The route information is announced to the peers by means of the UPDATE message.

In the case that a previously learned route is no longer reachable, BGP withdraws the route by encoding the MP_UNREACH_NLRI attribute with the vpn-ipv4 address and sending the UPDATE message. The route is removed from both the VRF's routing table and the RIB.

Receiving an UPDATE message

Assuming the route target is supported, receiving an UPDATE message can either trigger a newly learned route to be distributed to the appropriate CEs or can trigger a previously learned route to be withdrawn from the appropriate CEs and/or VRFs.

In the former case, the new routes are received in the MP_REACH_NLRI attribute and a lookup in the RIB is performed to see if this route information was previously received. If a match is found (same RD, prefix from the same peer), the route is considered to be a duplicate and is discarded. If no match is found, the vpn-ipv4 route is subject to inbound route filtering based on the encoded route targets (encoded as EXTENDED_COMMUNITY attributes). If there is no locally configured VRF that contains at least one of these route targets, the update message is discarded.

In the event that the route target is supported, the route is installed in the RIB and the route targets are used to determine the appropriate VRFs to install the route to. For each VRF that contains a matching route target, the route is installed into the VRF's routing table as a bgpMplsInternal route. If the same ipv4 address is already in the VRF's routing table with a different next hop (that is learned from a different PE node), the route is installed in the routing table and a route selection is performed on the two routes. For more information,

see Route selection (page 57). Installing both routes allows for fast recovery if the “best” route becomes unreachable and we need to change to the next available route.

In the latter case, the non-reachable route is received in the MP_UNREACH_NLRI attribute. A lookup is performed in the RIB and the route is flushed from the RIB. The route is also flushed from all appropriate VRFs. If the same route was previously learned from a different peer, that route gets promoted to the “best” route in the VRF routing table.

At the PE to CE link, normal routing protocol operation occurs to withdraw the “old” route from the routers in the site and advertise the “new” route to the routers in the site.

Route refresh capability

For scalability reasons, only routes that belong to a locally connected VPN are retained by the PE. As a result, when a configuration change is done to an import route target or a new VRF is provisioned, a route refresh request message (containing AFI/SAFI 1/128) will be sent to all connected BGP peers. This will cause the PE peers to reannounce their databases.

Similarly, you may receive a route refresh request message from a remote peer. Upon receiving this message for AFI/SAFI 1/128, all routes that should be sent to that PE peer are readvertised. This may be several BGP UPDATE messages.

Attention: The route refresh capability is required to establish a peer supporting AFI/SAFI 1/128. This capability will be automatically sent during session establishment when the BGP peer addressFamily contains ipv4MplsVpn.

Default route

A default route can be set up to handle all traffic that does not match any route in the RDB. The default route can originate either at the CEs (the VPN sites) or at the PEs. If it originates at the VPN site, the PEs distribute it like any other route. If the default route originates at the VRF, a static default route must be provisioned on the VRF that points to the CE as the next hop.

Data flow

There are two MPLS labels that are used in different roles:

- Transport label (page 74)
- Service label (page 74)

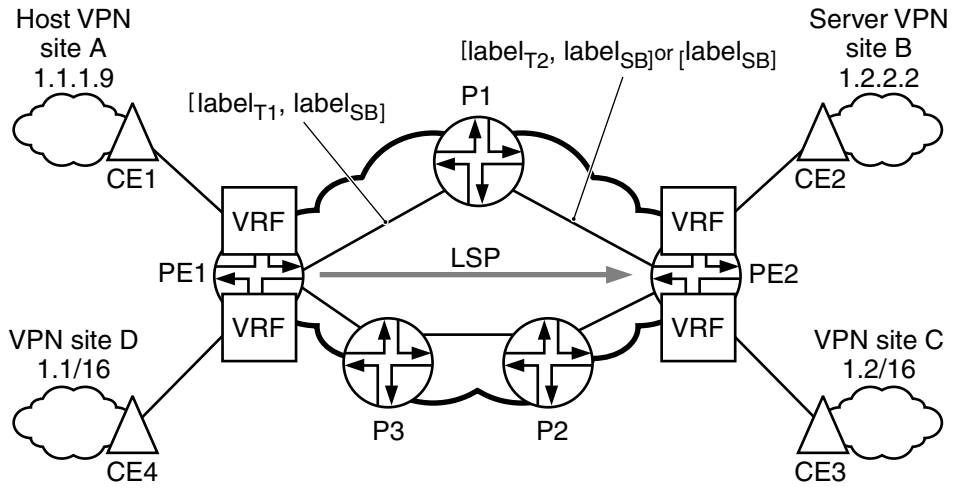
Transport label

The transport label is the outer label. It directs the packet to the correct PE router. It is associated with a BGP next hop and uniquely identifies an LSP.

Service label

The service label is the inner label. It determines how the PE router should forward the packet to the CE router. A service label is associated with a VRF. Each service label is unique per node.

BGP/MPLS VPN data flow



BGP/MPLS VPN data flow (page 74) shows Host 1.1.1.9 from Site A forwarding all data packets for Server 1.2.2.2 to its default gateway (CE1). CE1 receives the VPN traffic destined to VPN Site B, does an IP lookup, and forwards to PE1.

PE1 receives the VPN traffic through an interface associated with the VRF. PE1 performs an IP DA lookup on the VRF and obtains the following information:

- service label SB advertised by VRF PE2 associated with route 1.2.2.2
- BGP next hop for the route (the primary loopback address of PE2)

- the outgoing subinterface for the LSP from PE1 to PE2
- the transport label T1 associated with that LSP

PE 1 adds label SB to the IP packet, then adds label T1 and forwards on the outgoing interface to the first P router along the LSP from PE 1 to PE2.

The packet arrives at P router P1 through the LSP. P1 switches packets across the core of the provider's backbone network based on the outer label T1. Depending on whether P1 does penultimate label hopping or not, the protocol stack is different when P1 forwards the packet to PE2. If no penultimate label popping is done, P1 pops the transport label T1, pushes label T2 associated with that LSP, and forwards the packet to PE2. If P1 does penultimate label popping, it pops label T1 (exposing the service label) and forwards the packet to PE2.

When PE2 receives the VPN traffic, it pops the MPLS label [TL,SL] or [SL]. VRF does a route lookup and forwards the packet to CE2, which forwards the packet to Server 1.2.2.2 at Site 2.

Packet fragmentation

Data traffic with a frame size greater than the label switched path - maximum transmission unit (LSP-MTU) is discarded at the ingress access functional processor (FP). For every packet discarded, the VRF *fragFails* attribute is incremented by one. The table, Maximum IP frame size (page 75), can help you with engineering details to avoid fragmentation related to frame size.

Maximum IP frame size

Provisionable MTU-related attribute	Provisioned value (X)	LSP-MTU (Y)	RFC2547 maximum IP frame size (Z)	IP LER maximum IP frame size (Y)
AtmMpe MTU for trunk FP	X	Y = X - 4	Z = Y - 4	Y
LP Eth maxFrameSize for trunk FP	X	Y = X - 18 - 4	Z = Y - 4	Y
The value of 4 is either the MPLS transport or service shim header. The value of 18 is the 4-byte frame check sequence added to the 14-byte ethernet address.				

Service label scalability

The current implementation of RFC 2547 on a Multiservice Switch uses service label aggregation, which means it assigns only one Service Label per VRF per access card. This limits the number of service labels generated by a Multiservice Switch node. However, this is not always true for PE nodes implemented by other vendors. For example, a third party vendor can use one service label per route. Consequently, the huge number of non-aggregated

service labels generated by the third party vendor nodes can potentially consume all resources (VROs) on a Multiservice Switch. To avoid this, the Service Label Scalability feature is available to protect the VROs against non SL-aggregated nodes. This feature provides support for a non-aggregated mode in a scaled network, interworking with third party vendors in a scaled network. and traffic forwarding beyond a system's hardware capacity.

ServiceLabelUsage, a dynamic subcomponent of the VRF, is automatically generated whenever a VRF interface is provisioned on a new access card. This subcomponent holds the remote service label and the associated VRO usage information on a per card basis.

In a RFC 2547 VPN network of multi vendors, a Multiservice Switch configured as a PE node raises alarms when the resource for hardware data paths on an access card is going to be exhausted. The alarm can be one of three severity levels: minor, major, and critical corresponding to the VRO usage of 85, 95, and 99 percent respectively.

Dynamic forwarding type

The system supports the VRF of dynamic forwarding type as the default and only type.

The system takes the following action when the consumption of hardware resource on an access card reaches 95% (by all applications):

- The system demotes the hardware data paths on the card belonging to the VRF that used the most hardware resources (VROs consumed), to software data paths. An alarm is raised to indicate such an event.

Hardware resource exhaustion

The alarms can warn the operations of VRO exhaustion. The operator can use the CAS commands given in Monitoring remote service labels usage and associated hardware resources (page 78) to determine the access card that is going to exceed its engineering limit.

Hardware resource exhaustion is typically caused by one or more remote PE nodes running in a non-aggregated mode, i.e., using a unique service label per prefix or per host. At this point, the operator can choose one of the following two preventive actions:

- Do nothing - upon exceeding the card's engineering limit, the system automatically demotes the hardware data paths on the card belonging to the VRF with the most number of VRO installed.
- Re-engineer the card - the operator can select one of the following two options:
  - — off load the card by moving some VRFs to other card(s); this is done by moving some or all interfaces on the VRF to the other card(s)

— operationally, move the hardware data paths on the card belonging to one or more VRFs to software in order to free up the hardware resource. Refer to for the CAS commands in Setting the datapath forwarding mode based on ServiceLabelUsage (page 79).

Service label hardware resource usage

The maximum VROs pool on PQC6v2 and PQC12v1-based access cards are 4K and 8K respectively. Hardware resource usage per service label (page 77) summarizes the service label hardware resource usage on access card for the Multiservice Switch platforms, and access-trunk cards combinations.

Hardware resource usage per service label

Platform	Access Cards	Trunk Cards	PHP (T-Label = 3) ON	PHP (T-Label = 3) OFF
Multiservice Switch 7400	PQC6v2	PQC6v2	4 VROs/SL	8 VROs/SL
Multiservice Switch 15000/ Multiservice Switch 20000	PQC6v2	PQC12v1, MS3	4 VROs/SL	4 VROs/SL
Multiservice Switch 15000/ Multiservice Switch 20000	PQC12v1, MS3	PQC12v1, MS3	1 VROs/SL	1 VROs/SL
PHP = Penultimate Hop Popping T-Label = Transport Label				

Monitoring remote service labels usage and associated hardware resources

Use the operational commands described in this section to monitor the usage of remote service labels and their associated hardware resources (VROs).

Displaying remote service labels and VRO usage information

To display the remote service labels and VRO usage information for all the VRFs on all access cards, issue the following command:

```
d rtr/<router_name> VRF/* slu/*
```

To display the remote service labels and VRO usage information for VRF 1 on all access cards, issue the following command:

```
d rtr/<router_name> VRF/1 slu/*
```

To display the remote service labels and VRO usage information for VRF 1 on access card 1, issue the following command:

```
d rtr/<router_name> VRF/1 slu/1
```

Variable definitions

Variable	Value
<router_name>	is the router name.

Setting the datapath forwarding mode based on ServiceLabelUsage

Use the operational commands described in this section to set the datapath forwarding mode as software or hardware.

Setting the datapath forwarding mode as software

To set the datapath forwarding mode as software for all the VRFs on all access cards, issue the following command:

```
set rtr/<router_name> VRF/* slu/* fwdMode sw
```

To set the datapath forwarding mode as software for VRF 1 on all access cards, issue the following command:

```
set rtr/<router_name> VRF/1 slu/* fwdMode sw
```

To set the datapath forwarding mode as software for VRF 1 on access card 1, issue the following command.

```
set rtr/<router_name> VRF/1 slu/1 fwdMode sw
```

Setting the datapath forwarding mode as hardware

To set the datapath forwarding mode as hardware for all the VRFs on all access cards, issue the following command:

```
set rtr/<router_name> VRF/* slu/* fwdMode hw
```

To set the datapath forwarding mode as hardware for VRF 1 on all access cards, issue the following command:

```
set rtr/<router_name> VRF/1 slu/* fwdMode hw
```

To set the datapath forwarding mode as hardware for VRF 1 on access card 1, issue the following command:

```
set rtr/<router_name> VRF/1 slu/1 fwdMode hw
```

Resulting forwarding mode after the set fwdMode operation (page 79) shows the resulting forwarding mode after the setting the forwarding mode as hardware or software.

Resulting forwarding mode after the set fwdMode operation

Set fwdMode	fwdMode (Current)	fwdMode (Resulting)
hardware	hardware	hardware
	software	hardware*
software	hardware	software
	software	software
* If there are hardware resources to perform the action, then the datapath will be reconfigured to run in hardware; otherwise, it remains in software		

Variable definitions

Variable	Value
<router_name>	is the router name.

BGP/MPLS VPN over Carrier's Carrier MPLS networking overview

The Nortel Multiservice Switch Border Gateway Protocol/Multiprotocol Label Switching (BGP/MPLS) Virtual Private Network (VPN) over Carrier's Carrier MPLS networking solution allows the Nortel Multiservice Switch RFC2547 service provider (SP) to leverage another SP RFC2547 service for WAN interconnectivity. It does this by allowing a BGP/MPLS VPN service provider to transit across another service provider delivering hierarchical BGP/MPLS VPN (Carrier's Carrier) services. This hierarchical VPN approach is introduced in RFC 2547bis.

Carrier's Carrier resides exclusively on Multiservice Switch 15000 and Multiservice Switch 20000 nodes performing the role of the CE' node. For information on how to provision the Carrier's Carrier network, see NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.

The Carrier's Carrier network is an extension of the BGP/MPLS VPN network. For more information, see BGP/MPLS VPN overview (page 47).

Navigation

Main Carrier's Carrier networking components

Carrier's Carrier networking components include the BGP/MPLS networking components as well as two new components unique to Carrier's Carrier. For more information on the BGP/MPLS networking components, see Main BGP/MPLS VPN components (page 47). Here are the descriptions of the two new components

Carrier's Carrier customer edge (CE') router

This router interfaces with the Carrier's Carrier PE router (PE'), and it performs label distribution functionality between the customer carrier and carrier's carrier to utilize the MPLS VPN transit service provided by the carrier's carrier. It also acts as a PE router to the end CE router of the customer carrier. Note that the functionality of the PE and CE' is engineered on the same Nortel Multiservice Switch node.

Carrier's Carrier provider edge (PE') router

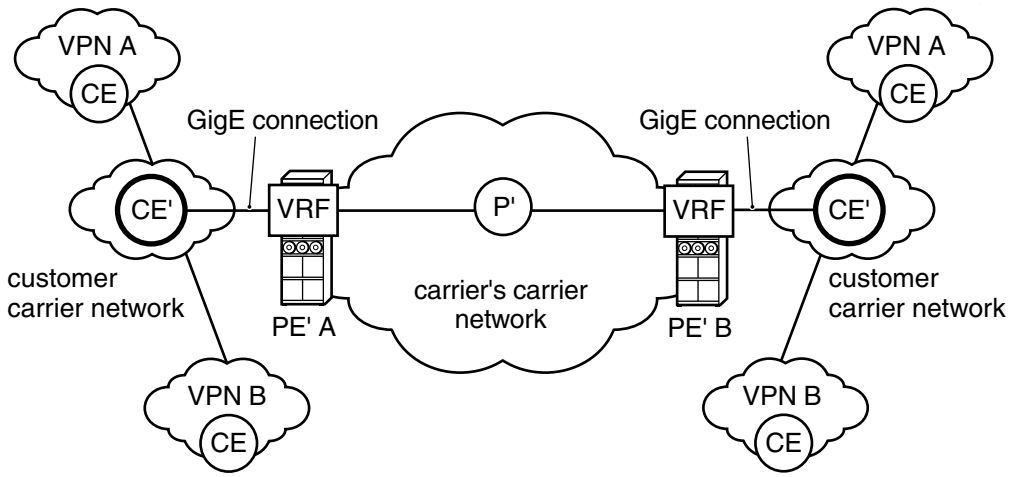
This router provides traditional MPLS VPN service to the CE router, and it performs label distribution functionality between the carrier's carrier and the customer carrier to provide MPLS VPN transit service to the customer carrier.

Carrier's Carrier network topology

Carrier's Carrier implements the requirements of the CE' node to interwork with a Carrier's Carrier PE' node, on top of acting as a PE node in its customer carrier space.

Figure Carrier's Carrier networking implementation topology (page 83) shows the implementation of the Carrier's Carrier network topology.

Carrier's Carrier networking implementation topology



CE	Customer Edge
CE'	Carrier's Carrier Customer Edge Router
PE'	Provider Edge Router
P'	Carrier's Carrier Provider Node
VPN	Virtual Private Network
VRF	VPN Routing and Forwarding table

MSS 3402 001 AA

Why use Carrier's Carrier networking solution?

The main benefits of implementing the Carrier's Carrier solution are:

- You can use Nortel Multiservice Switch 15000 and Multiservice Switch 20000 nodes to offer the RFC 2547 enterprise VPN service while taking advantage of an IP/MPLS base core network.
- Support of various topology options to ensure reliability of the service.
- Allows the SPs to use multiple backbone solutions.
- Clear administrative border can be defined over the MPLS interface between the CE' and PE'. While the customer carrier is able to independently manage the CE routers for its own enterprise customers, configuration, maintenance and operation in the transit service provider network is the responsibility of the carrier's carrier.
- As a benefit in a regular BGP/MPLS VPN, the customer address space and routing information are independent of the address space and routing information of other customer carriers. The same kind of independency also applies between the customer carrier and the carrier's carrier network, as they are operating in hierarchical BGP/MPLS VPNs.

- BGP, as the preferred routing protocol in general for connecting two SPs, can be used to interwork between a customer carrier and a carrier's carrier network.

Architecture

The architecture for Carrier's Carrier is based on a hierarchical MPLS model. The link between the CE' and the PE' must support MPLS. Multi-protocol BGP is used to advertise labeled CE' loopback addresses between CE' and PE' nodes. The ability to distribute this label information is enabled by RFC 2858. The negotiation between peers to distribute this label information is detailed in RFC 3392.

RFC 3107 allows a BGP peer to advertise more than one route to a given destination, as long as each route has its own unique label(s). This implies that only one route to the loopback address of the remote CE' will be advertised.

CE' access interfaces

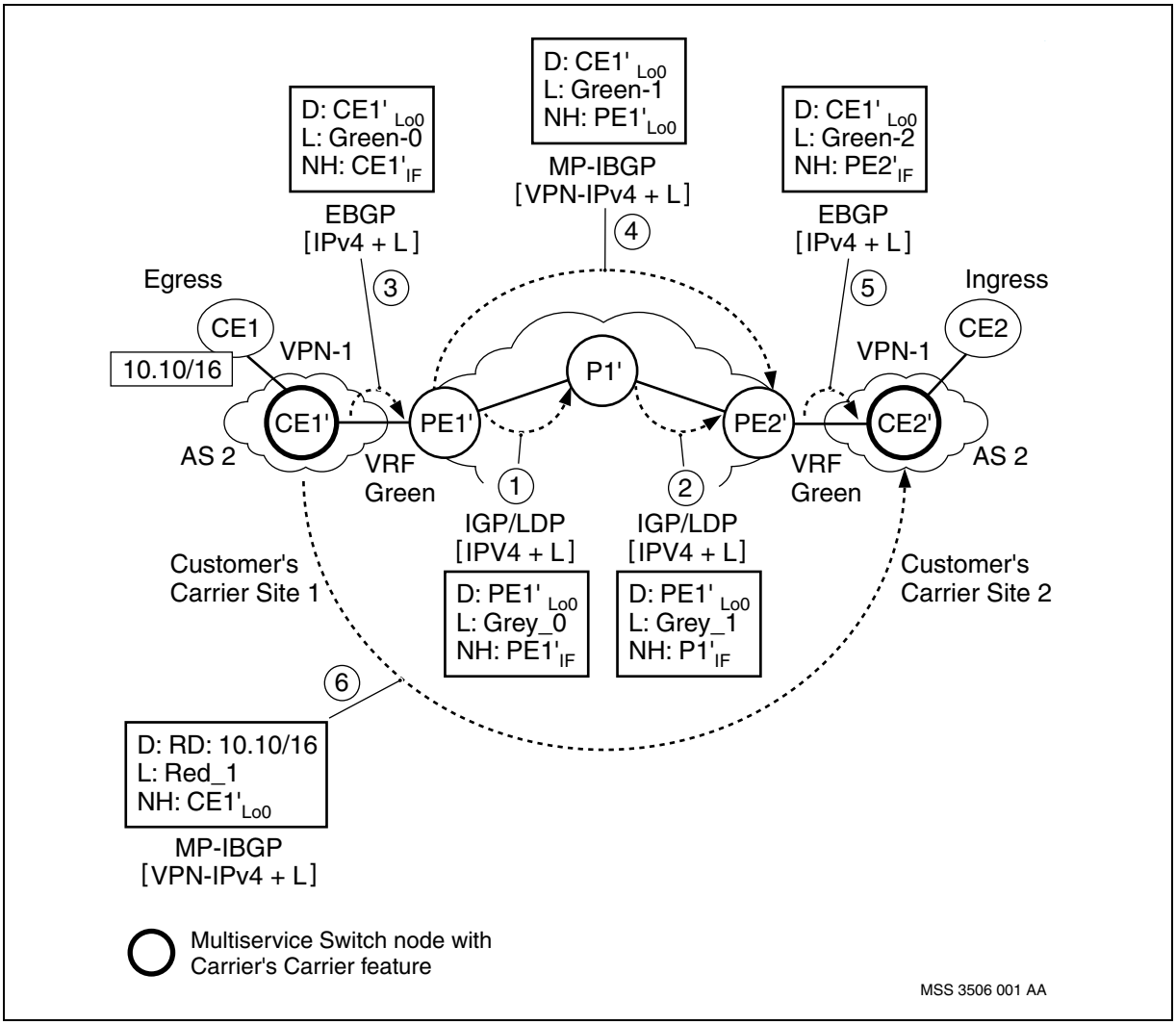
The access interface on the CE' node can be either an IP-based VRF interface or an IP-based non-VRF interface.

IP-based VRF interface

Control plane

Figure Routing and label binding protocols (page 85) shows the protocols needed to provide VPN services between CE nodes in a Carrier's Carrier configuration. The traffic is assumed to flow from CE2 (ingress) to CE1 (egress). Therefore, only the flow of the routing information from CE1 to CE2 is explained.

Routing and label binding protocols



1 PE1' uses LDP to advertise the implicit NULL label (LGrey-0), which it assigns to its loopback address (PE1'Lo0) to P1'. P1' assigns a local label (LGrey-1) to PE1'Lo0 before advertising it to PE2'. The following entry will be added to P1's MPLS forwarding table:

Label In: Lgrey-1
Label Out: -
Action: pop label, forward to PE1'IF
(A1)

Attention: PE1' can assign any label to its loopback, not just the implicit NULL.

2 PE2' learns the label assigned to PE1'Lo0 by P1', (LGrey-1), and adds the following entry to its forwarding table:

Destination: PE1'Lo0, pushLGrey-1, forward to P1'IF, (A2)

To send a packet to PE1'Lo0, PE2' needs to attach LGrey-1 label to the packet and forward it to P1'.

3 Independent of steps 1 and 2, CE1' uses EBGP (address family 1/4) to advertise the implicit NULL label (LGreen-0) that it assigns to its loopback address (CE1'Lo0) to PE1'.

4 Using the standard RFC2547 procedure, PE1' assigns a local label (LGreen-1) to CE1'Lo0 and advertises it to PE2'. PE1' uses its loopback address (PE1'Lo0) as next hop when advertising this route. PE1' installs the following entry in its MPLS forwarding table:

Label In: LGreen-1
Label Out: -
Action: pop label, forward to CE1'IF
(A3)

5 PE2' assigns a local label (LGreen-2) to CE1'Lo0 and advertises it to CE2' using EBGP (address family 1/4). PE2' installs the following entry in its MPLS forwarding table:

Label In: LGreen-2
Label Out: LGreen-1
Action: swap, forward to PE1'Lo0
(A4 - a)

However, according to (A2), to forward the packet to PE1', PE2' needs to push the LGrey-1 label into label stack and forward the packet to P1'. Therefore the previous MPLS FIB entry can be rewritten as:

Label In: LGreen-2
Label Out: LGreen-1, LGrey-1
Action: swap LGreen-1, push LGrey-1, forward to P1'IF
(A4)

CE2' installs the following entry in its forwarding table:

Destination: CE1'Lo0, pushLGreen-2, forward to PE2'IF, (A5)

6 In the previous step, CE2' learned the loopback address of CE1' from PE2'. Similarly CE1' learns the loopback address of CE2'. The MP-IBGP (address family 1/128) is established between CE' nodes and the customer carrier external routes (routes learned from CE1 and CE2) are exchanged between CE' nodes along with the service labels assigned to them by these nodes (Red labels). As an example: CE1' will advertise

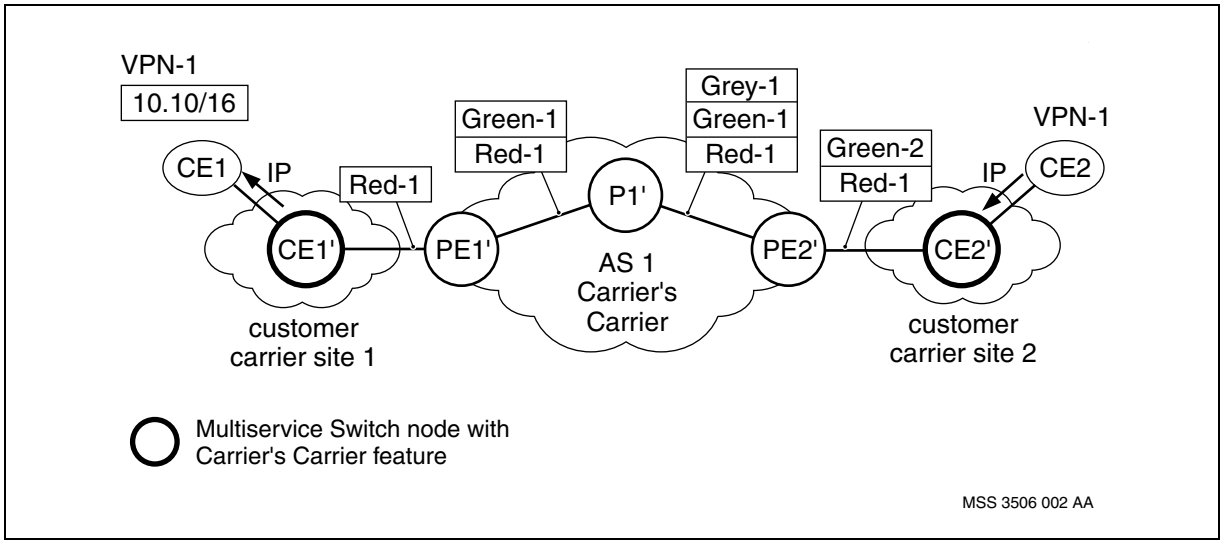
route RD:10.10/16 (learned from CE1 in VPN-1) to CE2' using its loopback address (CE1'Lo0) as next hop. CE2' will install the following route in its VPN forwarding table (VRF) for VPN-1:

Destination: 10.10/16, pushLRed-1, forward to CE1'Lo0, (A6)

Forwarding (data) plane

Figure Packet forwarding from CE2' to CE1' (page 87) shows how the packets are forwarded from CE2 to CE1. Assume that CE2 is sending traffic to CE1.

Packet forwarding from CE2' to CE1'



When IP traffic from CE2 with the destination prefix of 10.10/16 (located at CE1 site) arrives at CE2', CE2' makes an IP lookup in its VPN-1 VRF table and finds the (A6) routing entry. Based on this entry, CE2' pushes the LRed-1 label (service label) into the label stack and tries to forward the packet to CE1'Lo0. CE2' make another lookup for CE1'Lo0 in its main forwarding table and finds the (A5) routing entry:

Destination: CE1'Lo0 push LGreen-2 forward to PE2'IF

CE2' pushes LGreen-2 to the label stack. The next hop for this route (PE2'IF) is on the same subnet as CE2'IF, so no extra label is needed to send the packet to PE2'.

PE2' uses the (A4) MPLS FIB entry to forward the packet. It swaps LGreen-2 with LGreen-1 and uses the LGrey-1 to tunnel the packet across the Carrier's Carrier network (through node P1') to PE1'.

P1' uses the (A1) MPLS FIB entry to forward the packet to PE1'.

PE1' uses the (A3) MPLS FIB entry to forward the packet. PE1' removes the LGreen-1 label and forwards the packet to CE1'.

CE1' uses the red label (LRed-1) to identify the egress interface, and after removing the Red label, forwards the original IP packet to the egress customer site (CE1).

IP-based non-VRF interface

Figure IP-based non-VRF interface: control plane (page 89) shows a configuration in which customer carrier PEs (CE1'and CE2') have IP based non-VRF interfaces. This type of interface can be used to provide management access to CE' nodes. The only difference from the IP-based VRF interface is that the customer routes are distributed between CE1' and CE2' routers using normal IBGP instead of MP-IBGP with address family 1/128.

IP-based non-VRF interface: control plane

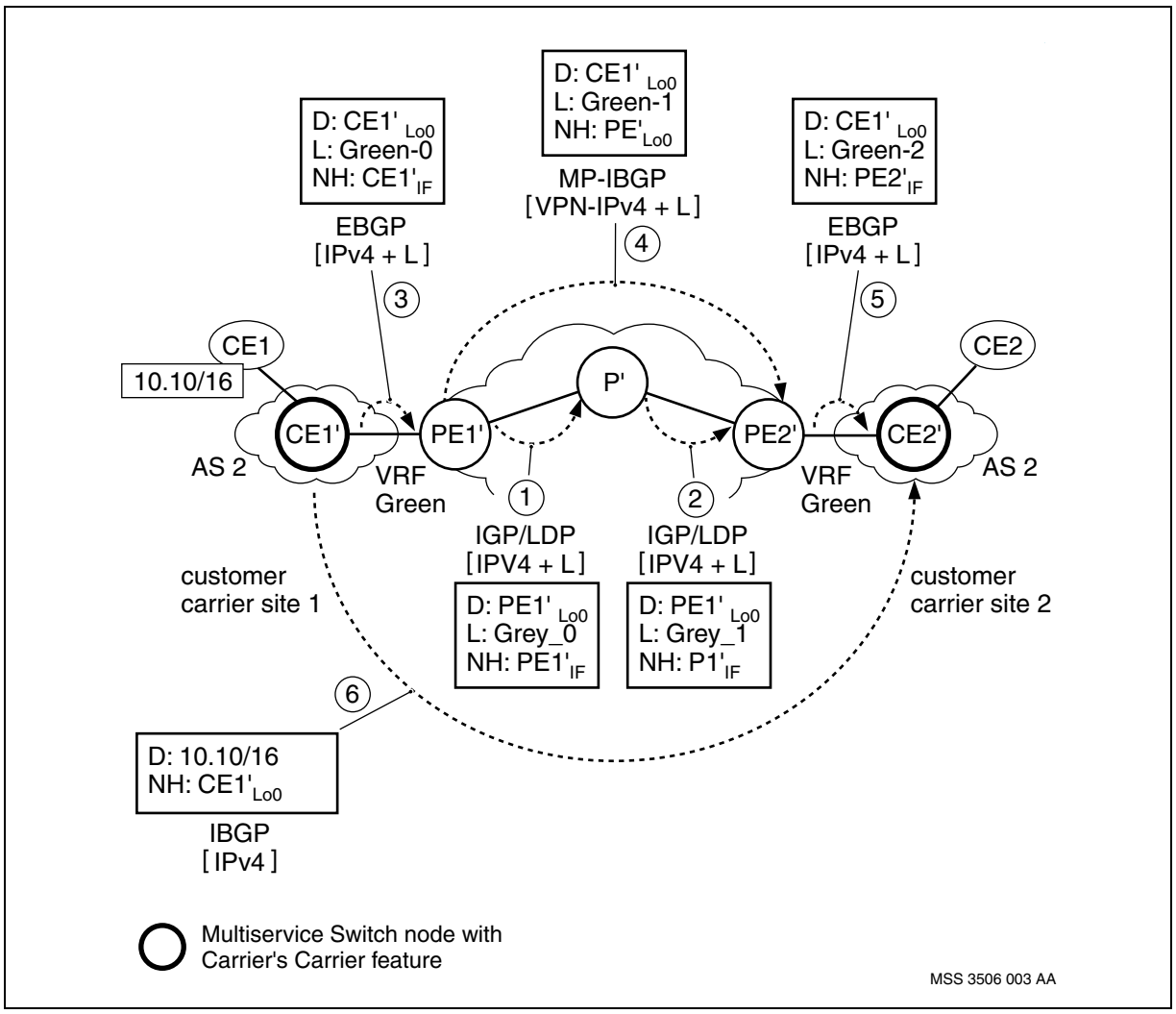
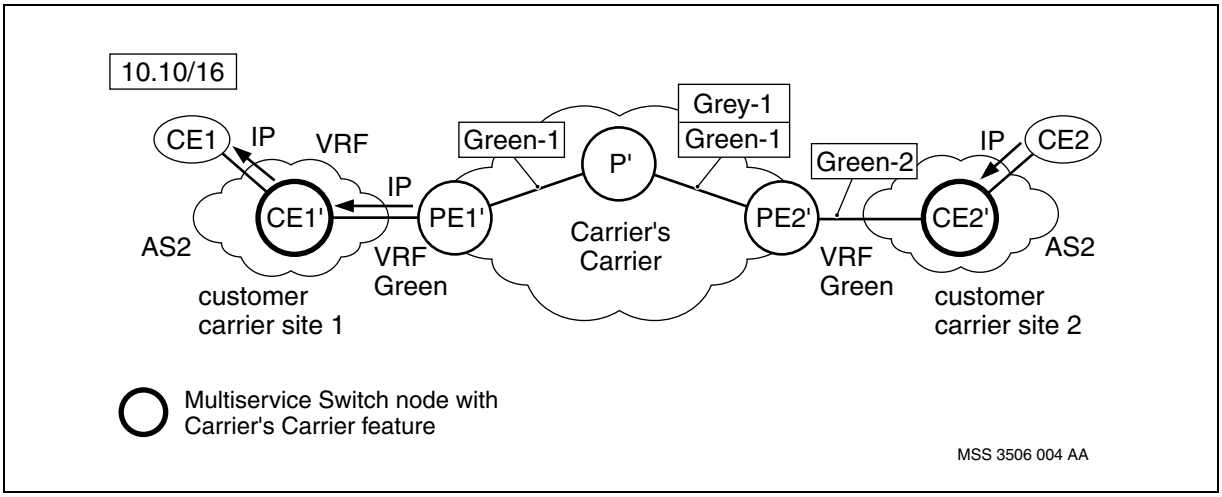


Figure IP-based non-VRF interface: forwarding plane (page 90) shows how the packets are forwarded from CE2 to CE1.

IP-based non-VRF interface: forwarding plane



Deployment of Carrier's Carrier

The strategy for deploying the Carrier's Carrier feature for a customer who already uses the BGP/MPLS VPN services is to setup a parallel connection to a Carrier's Carrier network and gradually transfer the customer traffic flows from the existing RFC2547 backbone to the Carrier's Carrier backbone. Traffic redirection can be achieved by making the routes learned from the Carrier's Carrier backbone more preferred over the routes learned from the old RFC2547 backbone.

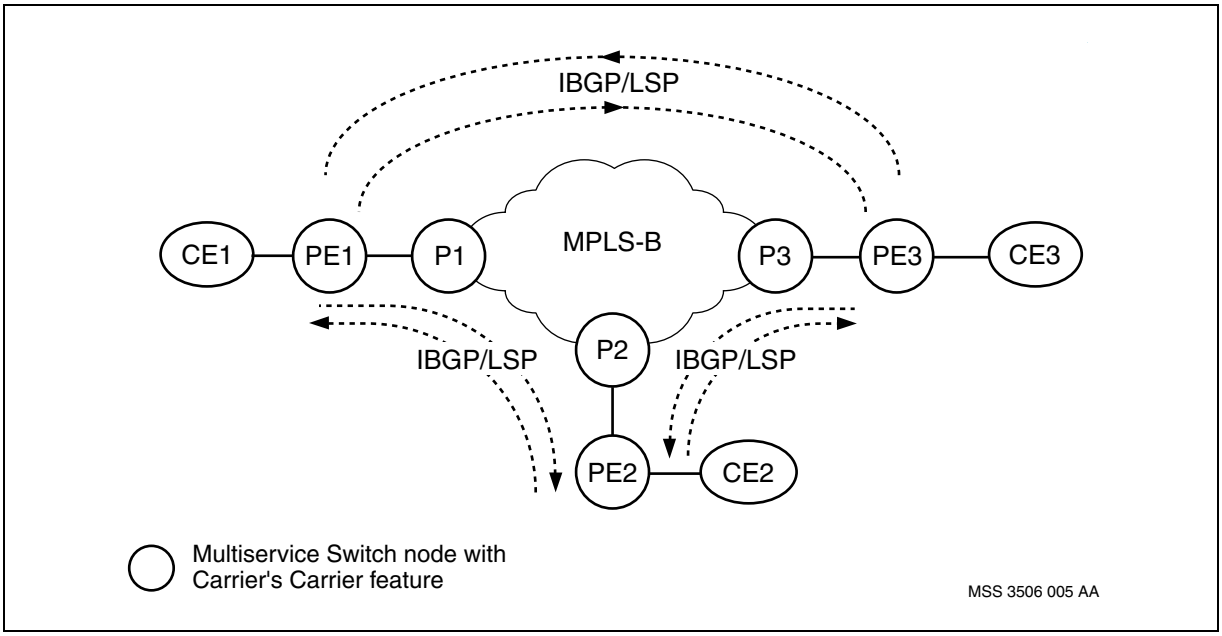
The two different methods for deploying the Carrier's Carrier feature are: a parallel connection to the Carrier's Carrier backbone is established by adding a new link between the existing customer PE and Carrier's Carrier backbone (method A), or the parallel link to the Carrier's Carrier backbone is provided through adding a new customer PE to the customer network (method B).

Method A: using the existing customer PE

During the transition, each customer carrier PE needs to maintain links to both standard RFC2547 and Carrier's Carrier backbones. The transition is accomplished by shifting the control and data traffic from one link to another:

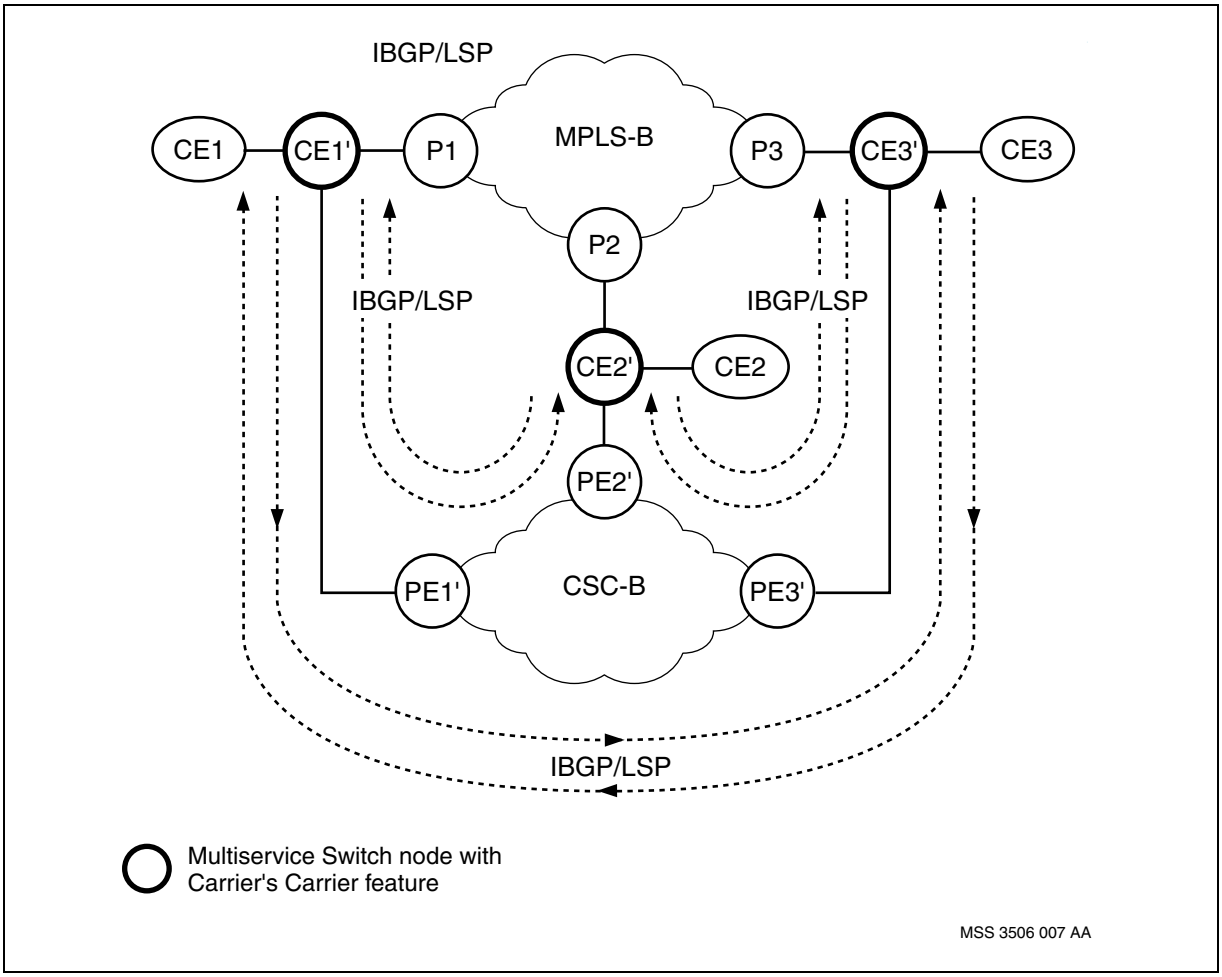
- Initially, Customer carrier PEs (PE1 to PE3) are only connected to the standard RFC2547 backbone (MPLS-B) as shown in Method A - initial phase (page 91). The MPLS-B nodes are acting as P nodes to provide IGP connectivity as well as LSPs between customer carrier PEs in both directions.

Method A - initial phase



- The software on each customer carrier PE node is upgraded to support the Carrier's Carrier feature. To preserve the CE-to-CE connectivity during the software upgrade, each CE can be dual-homed to two different PEs which are not upgraded at the same time.
- A new link is added between CE' and Carrier's Carrier PE (PE') and EBGP (address family 1/4) is provisioned on that link. Each CE' node learns the loopback address of the remote CE' nodes through both MPLS-B (IGP) and CSC-B (EBGP). By default, IGP routes are preferred over labeled EBGP routes and therefore, both MP-IBGP sessions and LSPs are still maintained over MPLS-B backbone.
- On CE1' node, the route preference of the IGP routes are modified to make the labeled EBGP routes preferred over the IGP routes. As a result, the IBGP packets from CE1' to remote CE' nodes will start flowing through the Carrier's Carrier backbone (CSC-B). The LSPs from CE1' to remote CE' nodes are also switched to CSC-B backbone.
- Finally, The route preference of the IGP routes are modified on the remaining CE' nodes (one node at a time) to make the labeled EBGP routes preferred over the IGP routes. As a result, all IBGP sessions and LSPs are switched to the CSC-B network as shown in Method A - final phase (page 92).

Method A - final phase



At this point the transition from standard RFC2547 to Carrier's Carrier configuration is complete and the connection between CE' and standard RFC2547 backbone (MPLS-B) can be disconnected.

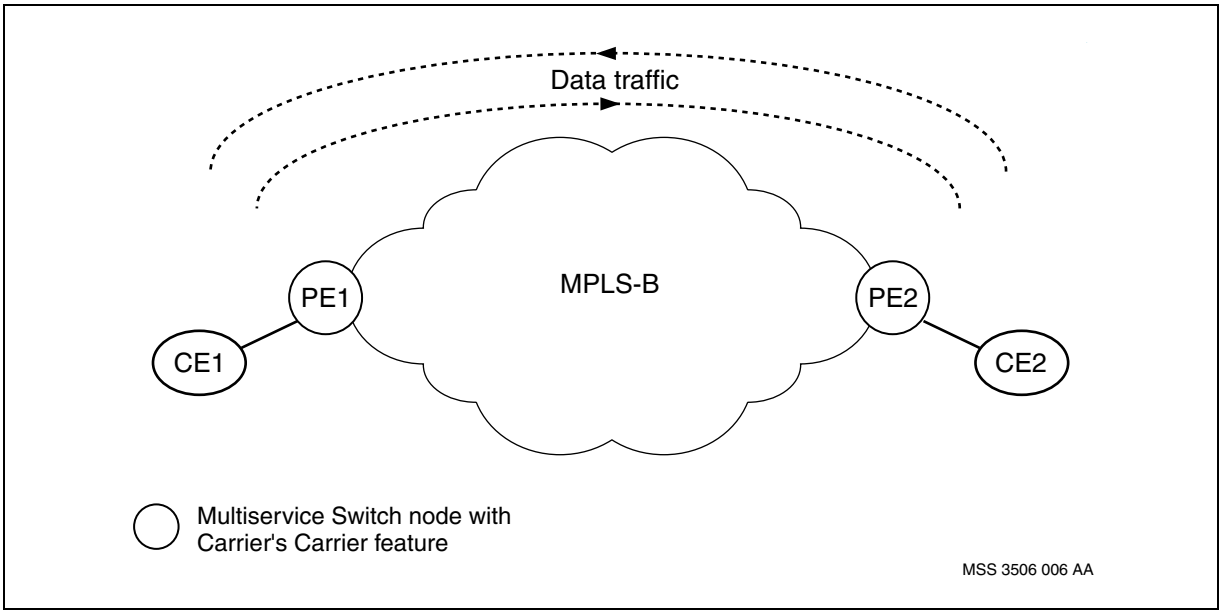
Method B: adding a new customer PE

During the transition, each CE needs to maintain links to two separate customer provider PE nodes. One of these PE nodes is connected to a standard RFC2547 backbone while the other one is connected to a carrier's carrier backbone and acts as a CE for that network (CE'). The transition from the standard RFC2547 to Carrier's Carrier service is accomplished by shifting the customer traffic from CE-PE access link to CE-CE' access link as explained in the following steps:

- Initially, each CE node is connected to only one customer provider PE node, which provides the standard RFC2547 service using the MPLS-B backbone Method B - initial phase (page 93). The access protocol

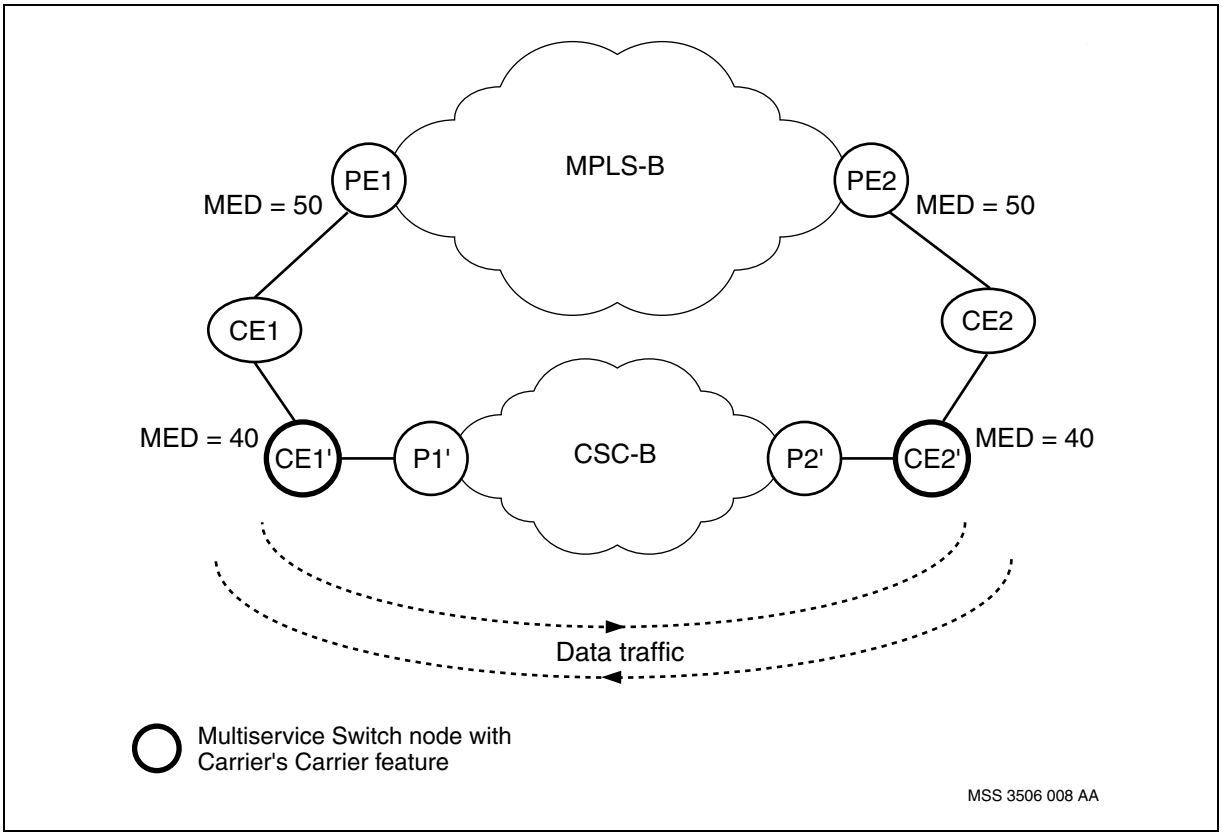
between CE and PE node is EBGP. Both IBGP sessions and LSPs are established across the MPLS-B backbone.

Method B - initial phase



- A new customer provider PE node (CE') is configured to provide the connectivity between the CE nodes across the Carrier's Carrier backbone. Access protocol between CE and CE' node is also EBGP. New MP-IBGP session and LSPs will be established between CE' nodes across the CSC-B network in addition to those which are still maintained across the MPLS-B backbone. MED on CE' VRF is set to a higher value than the MED on PE VRFs to ensure the traffic is still forwarded through the MPLS-B backbone.
- The MED on CE2' VRF is lowered to 40. As a result, the traffic from CE2 to CE1 will flow through the Carrier's Carrier network.
- Finally, the MED on CE1' VRF is lowered to 40. Now the traffic between CE nodes use the Carrier's Carrier backbone in both directions Method B - final phase (page 94).

Method B - final phase



At this point the transition from standard RFC2547 to Carrier's Carrier configuration is complete and the connection between CE and PE nodes can be disconnected.

Inter-AS VPN networking overview

The Nortel Multiservice Switch inter-autonomous system (inter-AS) virtual private network (VPN) solution allows border gateway protocol/multiprotocol label switching (BGP/MPLS) virtual private networks (VPNs) to interconnect to multiple service providers in different autonomous systems (ASs). Techniques for interconnecting VPNs across AS boundaries are defined in the Internet Engineering Task Force (IETF) request for comments (RFC) 2547bis document and include multi-AS backbones types A and C. Before the introduction of the inter-AS feature, the Multiservice Switch only supported multi-AS type A. With the introduction of the inter-AS capability, the Multiservice Switch supports specific type C topologies as described in this section.

The inter-AS VPN network is an extension of the BGP/MPLS VPN network and the Carrier's Carrier network. For more information on BGP/MPLS VPN, see BGP/MPLS VPN overview (page 47). For more information on Carrier's Carrier, see BGP/MPLS VPN over Carrier's Carrier MPLS networking overview (page 81).

Navigation

Topologies

The Multiservice Switch inter-AS solution supports a number of topology options. These options are referred to as:

- Edge network topology with flat or hierarchical BGP/MPLS networking
- Backhaul network topology with flat or hierarchical BGP/MLPS networking

Terminology

Refer to the following list for term definitions:

- An edge network refers to a service provider BGP/MPLS VPN where PE nodes do not directly communicate with nodes from another AS. Specifically, the Multiservice Switch PE nodes are near the customer edge (CE) of the service provider network, and a Multiservice Switch P node is peering with other nodes in different ASs. All devices support IP and are connected using IP.
- A backhaul network refers to a service provider's use of a layer 2 network to extend the geographical reach of an IP VPN service without having to place a PE node closer to the edge of the service provider network. In a backhaul network, the PE node may directly communicate with nodes from another AS.
- A flat network refers to a service provider BGP/MPLS VPN that peers with another service provider BGP/MPLS VPN in a different AS to provide a VPN solution to an end customer. In this case, a two-label MPLS stack is used and labels are swapped across service provider boundaries.
- A hierarchical network refers to a service provider BGP/MPLS VPN that utilizes a Carrier's Carrier service to interconnect two PE nodes from different BGP/MPLS VPN service provider sites. The two PE nodes may or may not be within the same AS. In this case, a three-label MPLS stack is used and the Carrier's Carrier service provider pushes its label on top of the customer carrier service provider two-label stack, effectively tunneling the label switched path (LSP) through the Carrier's Carrier network.

Why use inter-AS VPN networking solution?

The main benefits of implementing the inter-AS VPN solution are to:

- Allow VPNs to span multiple service providers. Service providers managing different ASs can offer a BGP/MPLS VPN service to the same end customer.
- Allow VPN sites to span multiple geographical areas.

Inter-AS VPN in flat mode

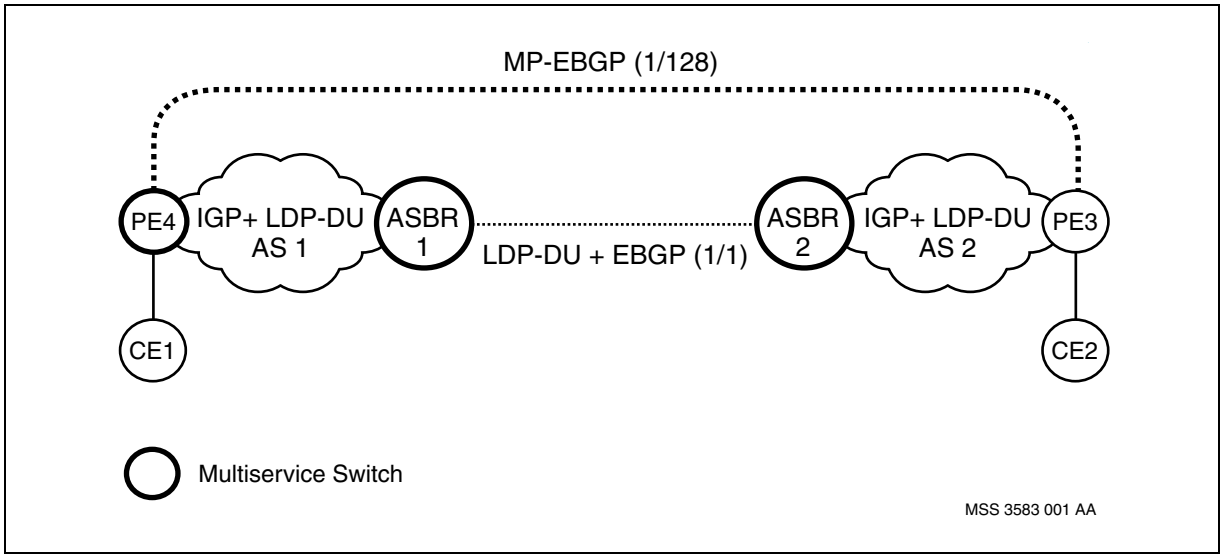
The service provider providing BGP/MPLS VPN service is peering with another BGP/MPLS VPN service provider in a different AS to reach other VPN customer sites in the same VPN. Typically, there exists a trusted relationship between the two ASs. For example, two organizations interwork together, but they need to control IPv4 address space leaks across network boundaries and thus do not want to interwork using the internal gateway protocol (IGP). The service provider can provide an edge or backhaul network deployment to the customer edge (CE) nodes of the end customer.

Edge network topology

The figure, Edge network deployment with inter-AS VPN in flat mode (page 97), illustrates a VPN which spans two different service providers and ASs.

External BGP (EBGP) is used between autonomous system border router 1 (ASBR1) and ASBR2. The ASBR1 learns the loopback address of PE3 using EBGP and redistributes that address to PE4 using IGP. Similarly, ASBR2 learns the loopback address of PE4 using EBGP and redistributes that address to PE3 using IGP. PE3 and PE4 can then use multi-hop EBGP to distribute VPN IPv4 routes and service labels. The MPLS label distribution protocol - downstream unsolicited (LDP-DU) is used to provide transport label distribution to set up an LSP between PE3 and PE4.

Edge network deployment with inter-AS VPN in flat mode



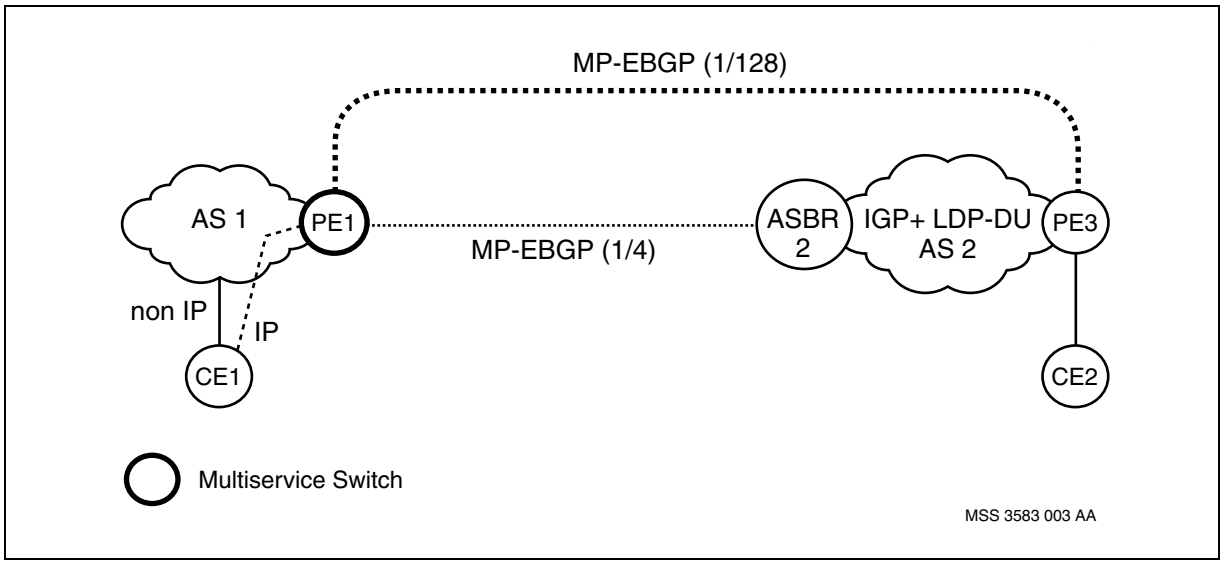
Backhaul network topology

The figure, Backhaul network deployment with directly connected ASs and inter-AS VPN in flat mode (page 98), illustrates a PE node with backhaul layer 2 access connecting directly to the other service provider AS. The figure, Backhaul network deployment with indirectly connected ASs and inter-AS VPN in flat mode (page 98), illustrates a PE node with a backhaul layer 2 access connecting to the other service provider AS using a transit network. In both scenarios, the end customer is connected to the PE1 node using a layer 2 network.

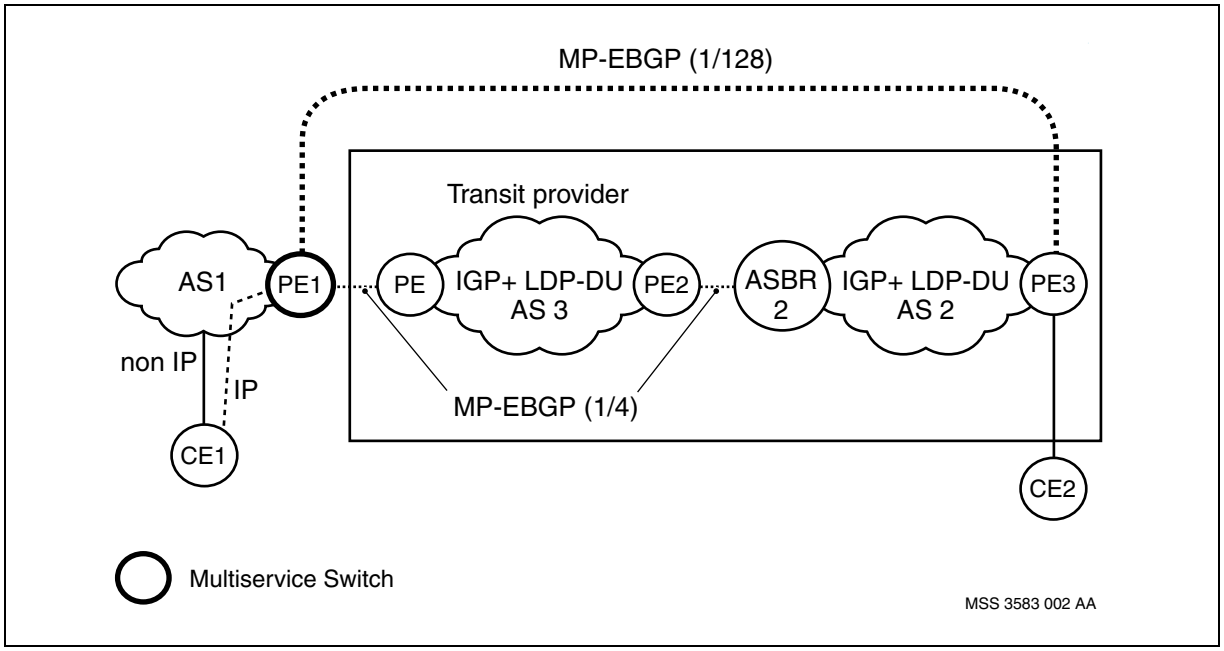
The multiprotocol - EBGP (MP-EBGP) is used to exchange VPN IPv4 routes and service labels. In this case, the PE1 learns the loopback address of PE3 using EBGP 1/4, which includes the label switching information. An LSP between PE1 and PE3 can then be established. PE1 and PE3 use multi-hop MP-EBGP to distribute VPN Ipv4 routes.

Attention: In the figure, Backhaul network deployment with indirectly connected ASs and inter-AS VPN in flat mode (page 98), if ABSR2 is a Multiservice Switch, then the combination of LDP-DU and EBGP IPv4 can be used between PE2 and ASBR2 instead of labelled MP-EBGP 1/4.

Backhaul network deployment with directly connected ASs and inter-AS VPN in flat mode



Backhaul network deployment with indirectly connected ASs and inter-AS VPN in flat mode



Inter-AS VPN in hierarchical mode

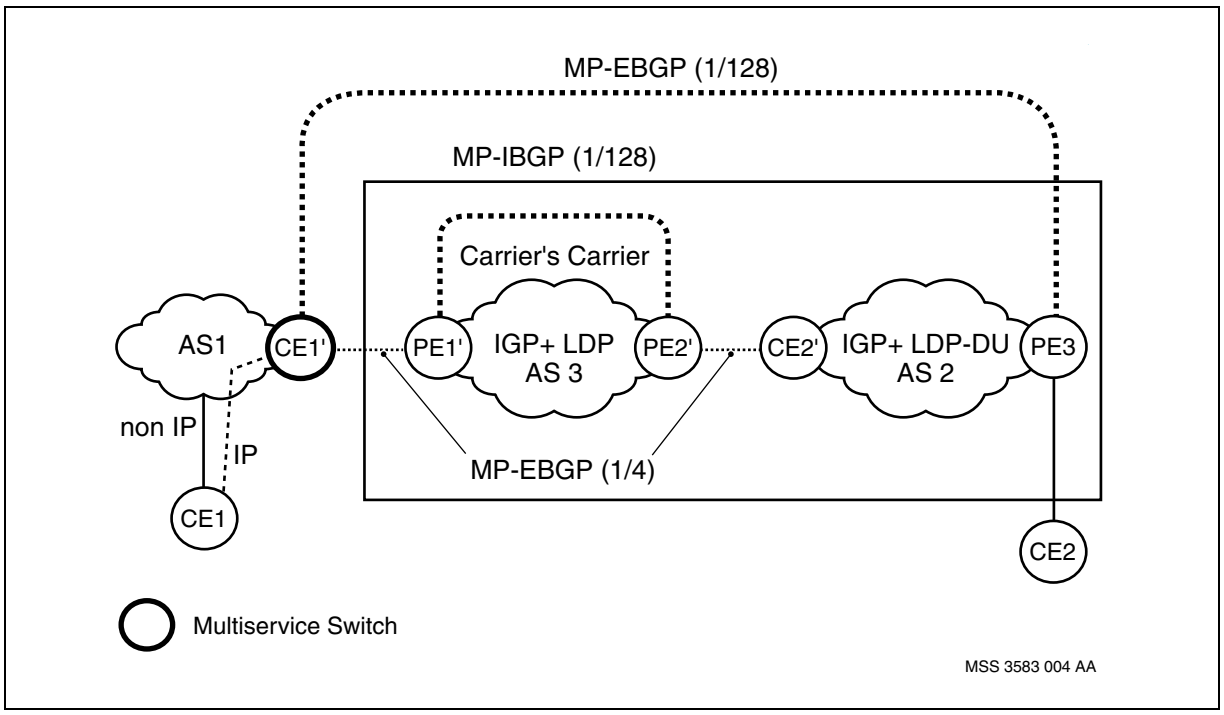
The service provider providing BGP/MPLS VPN service to the end customer transits across another service provider delivering hierarchical BGP/MPLS VPN (Carrier's Carrier) service. The service provider can use a backhaul network topology.

Backhaul network topology

The figure, Backhaul network deployment with inter-AS VPN in hierarchical mode (page 99), illustrates a service provider backhaul network offering BGP/MPLS VPN service, which spans multiple service providers and ASs.

MP-EBGP is used to exchange labelled IPv4 routes. In this case, the CE1' learns the loopback address of PE3 using EBGP including label switching information. The CE1' and PE3 use multi-hop EBGP to distribute VPN IPv4 addresses.

Backhaul network deployment with inter-AS VPN in hierarchical mode



Hub and Spoke networking overview

The Nortel Multiservice Switch BGP/MPLS VPN solution provides flexibility in constructing different kinds of VPNs by:

- controlling the distribution of routing information among various sets of sites
- setting up the Import Route Targets and Export Route Targets properly.

This chapter explains the Hub and Spoke VPN topology and the Star topology.

In a Hub & Spoke topology, all packet flows originating at spoke sites and destined to other spoke sites traverse through a central hub. A Star topology allows spoke to hub and hub to spoke packet flow only. Spoke to Spoke communication through a central hub is not allowed. Hub & Spoke topology and Star topology includes the possibility of configuring a number of spoke VRFs on the same PE (or on a different PE) where the hub VRFs are configured.

Attention: Hub & Spoke network is an extension of the BGP/MPLS VPN network. For more information, see BGP/MPLS VPN overview (page 47).

Navigation

Hub and Spoke networking components

The Hub & Spoke networking components include the BGP/MPLS networking components with special configuration for the Import Route Target and the Export Route Target of the VRF. The following terminology is defined:

Hub VRF

A VRF that is connected to a Hub site. The Hub site connects to the PE via two VRFs. One VRF imports routes from the Spoke sites and advertises them to the Hub site. The other VRF learns the routes from the Hub site and exports them to the Spoke sites.

Spoke VRF

A VRF that is connected to a Spoke site.

Import-Spoke VRF

A Hub VRF with its Import Route Target equal to the Export Route Target of the Spoke VRF. This is the VRF that learns the routes from the Spoke site and advertises them to the Hub site.

Export-Hub VRF

A Hub VRF with its Export Route Target equal to the Import Route Target of the Spoke VRF. This is the VRF that learns the routes from the Hub site and advertises them to the Spoke site.

Remote VPN-IPv4 route

A route that is learned via the MP-BGP protocol from other PE nodes in the network. The routing protocol type is bgpMplsInternal.

Local VPN-IPv4 route

A route that is learned via Inter-VRF functionality from other VRFs on the same PE node. The routing protocol type is bgpMplsInternal.

Locally learned route

An IPv4 route learned via a supported routing protocol (EBGP, OSPF, RIP) or static routes from the CE. The routing protocol type can be one of bgpExternal, ospfExternal, rip, remote, or local, depending on the means by which the route was learned.

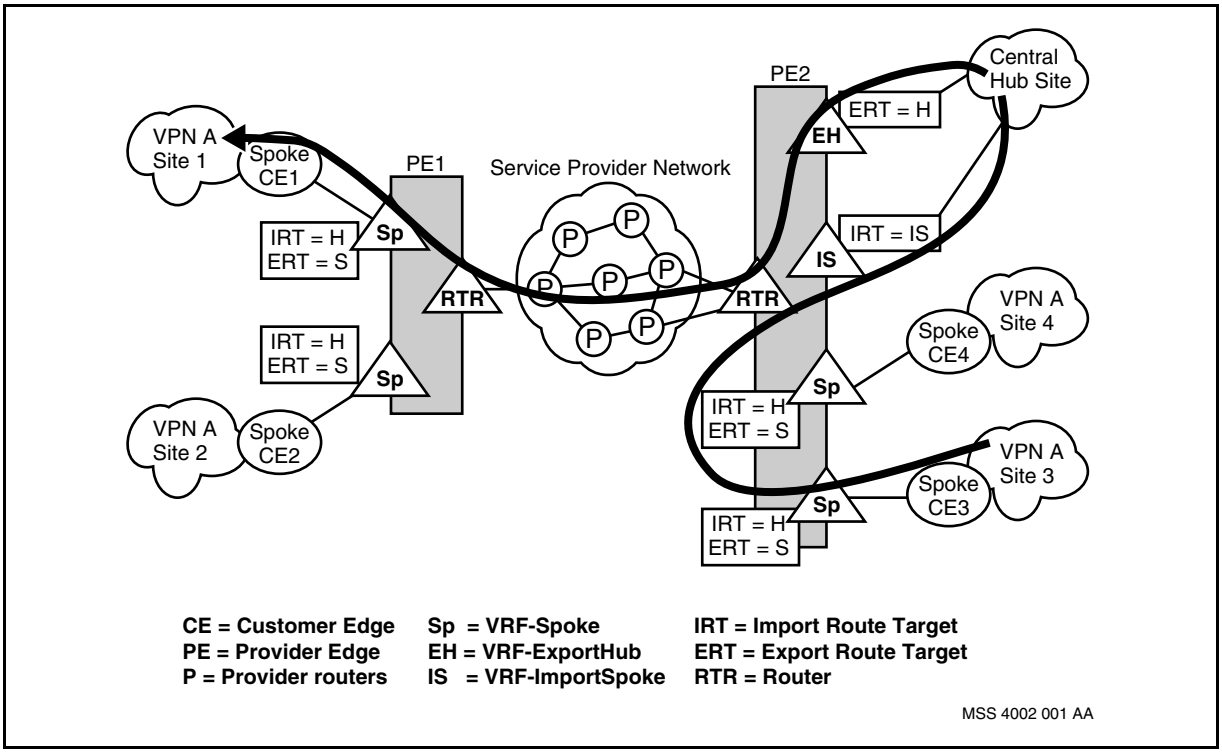
Hub and Spoke topology

Hub and Spoke is a VPN topology where packet flows between the spoke sites are allowed, but only through the hub site.

Inter-PE flow

The figure Inter-PE Hub and Spoke topology (page 102) depicts the Hub and Spoke network where spoke sites 3 and 4 and the hub site are connected to the same PE2. In this example, the two spoke sites 1 and 3 are connected to different PEs. The arrow in the figure indicates the control flow of routes advertisement. The data flow is in the opposite direction.

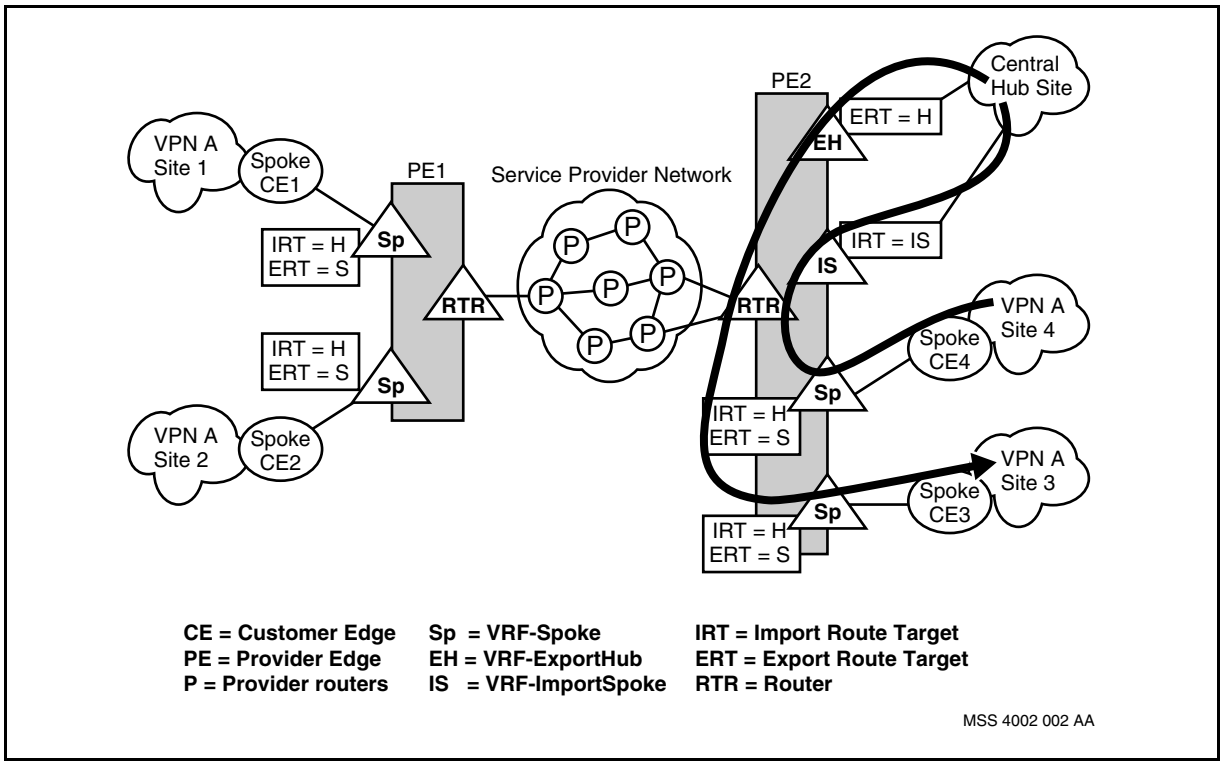
Inter-PE Hub and Spoke topology



Intra-PE flow

Figure Intra-PE Hub and Spoke topology (page 103) depicts the Hub and Spoke network where spoke sites 3 and 4 and the hub site are connected to the same PE2. In this example, the two spoke sites 3 and 4 are connected to the same PE2. The arrow in the figure indicates the control flow of routes advertisement. The data flow is in the opposite direction.

Intra-PE Hub and Spoke topology



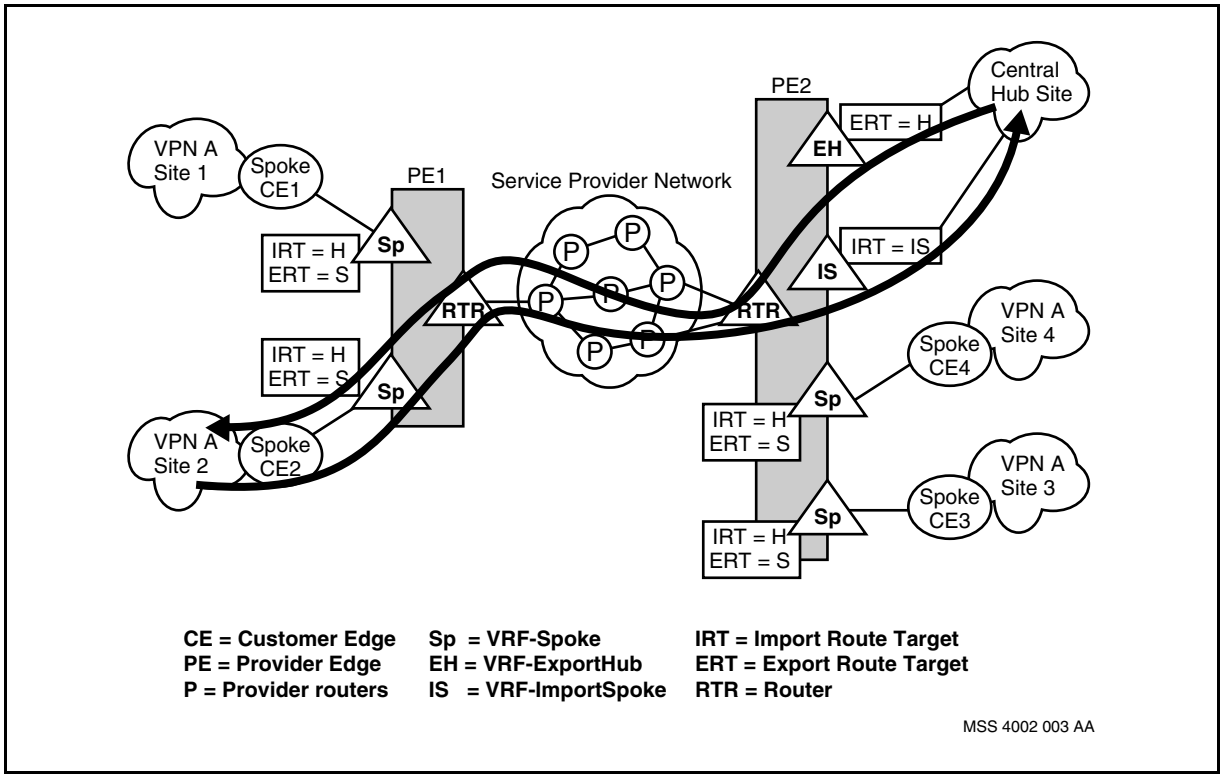
Star topology

Star is a VPN topology where packet flows between the Hub and Spoke sites are allowed, but no Spoke site to Spoke site communication is allowed.

Inter-PE flow

Figure Inter-PE Star topology (page 104) depicts a Star network where spoke site 2 and the hub site are connected to different PEs. The arrows in the figure indicate the control flow of routes advertisement. The data flow is in the opposite direction.

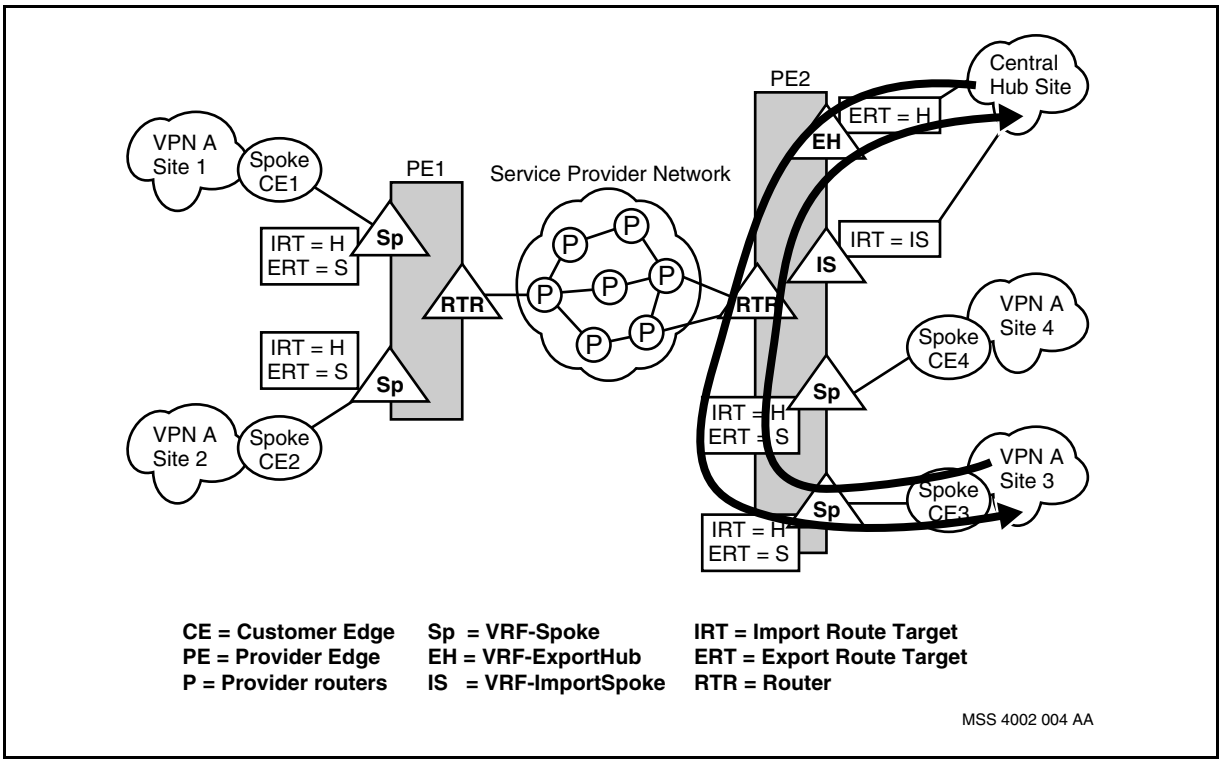
Inter-PE Star topology



Intra-PE topology

Figure Intra-PE Star topology (page 105) depicts the Star network where the spoke site 3 and the hub site are connected to the same PE2. In this example, spoke site 3 is communicating with the central hub. The arrows in the figure indicate the control flow of routes advertisement. The data flow is in the opposite direction.

Intra-PE Star topology



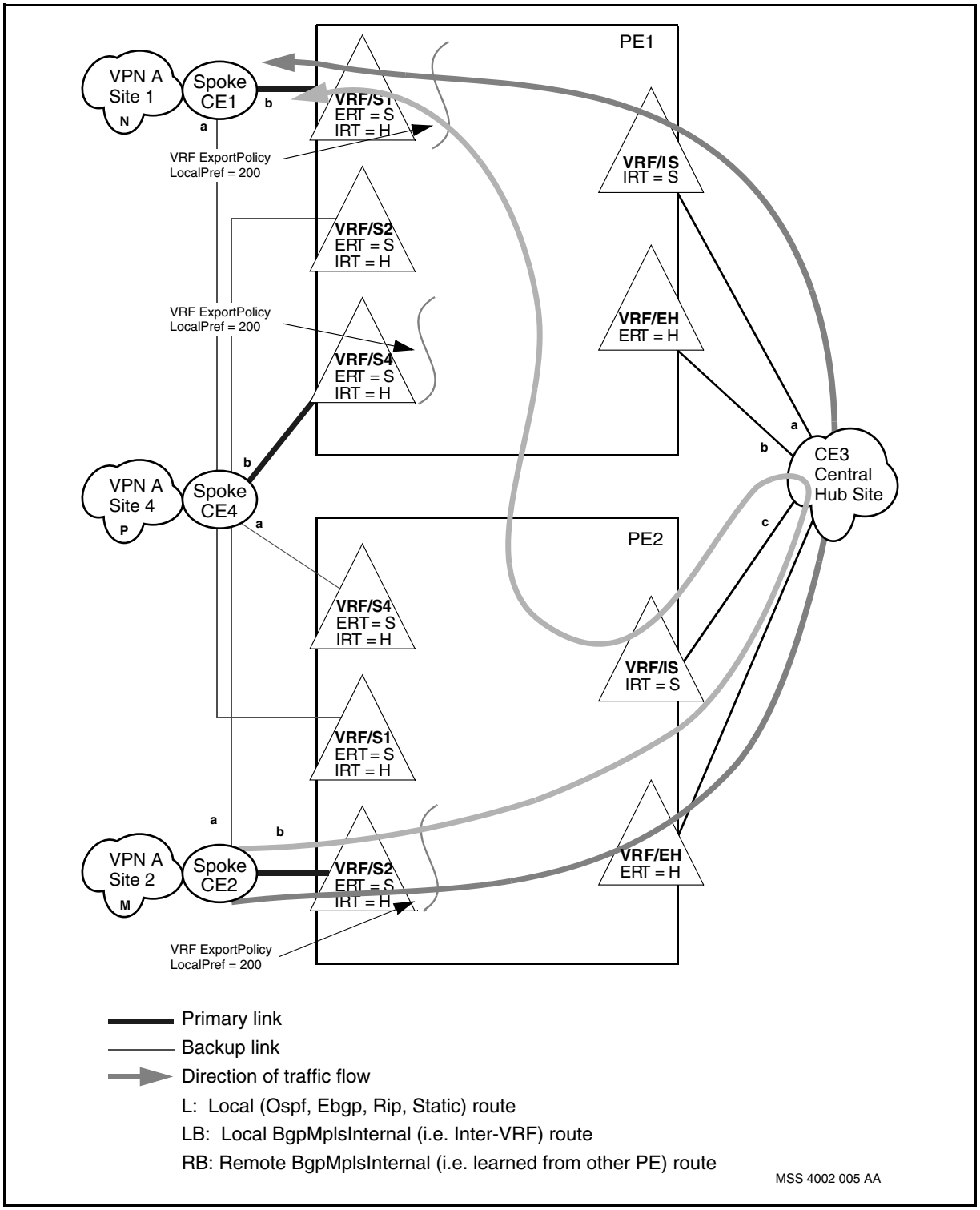
Hub and Spoke topology with site redundancy

Figure Hub and Spoke topology with site redundancy (page 107) depicts an example of a Hub & Spoke topology with multiple site redundancy for a single VPN. It shows an example of multiple sites connected to two different PE nodes where each site has a primary connection to one PE node and the secondary (or backup) connection to the other PE node. The Hub site is also connected to two PE nodes. The Hub site may be connected in a redundancy configuration as well whereby it will always prefer one path for specific sites. For example, the link to PE1 would always be preferred for traffic destined to destinations N and P, and the link to PE2 would always be preferred for traffic destined to destination M. A more common use of this topology connection would be for the HUB site to load balance the traffic over the two link connections to the PE nodes.

In either case, it is desirable that traffic destined to the Site always use the primary connection unless the primary connection is not available, then the backup connection can be used. For example, traffic destined to Site 1 should go through PE1 and traffic destined to Site 2 should go through PE2. To achieve this, a VRF Export Policy needs to be configured for the VRFs with the primary connection to set the LocalPreference value to be higher than the default LocalPreference value of 144. This will allow the VRF/IS to prefer the remotely learned VPN-IPv4 routes over the locally learned ones from Site 2 (i.e. route learned via the primary connection) and to prefer the locally learned VPN-IPv4 routes over the remotely learned ones from Sites 1 and 4 (route learned via the primary connection).

The data flow direction is also depicted showing two possible directions traffic can flow originating from Site2 and destined to Site1. The direction chosen depends on the HUB site decision, if one link will always be chosen for traffic destined to Site1 or if a load balancing mechanism is used.

Hub and Spoke topology with site redundancy



Route distribution

The following are the events that occur to exchange routing information between VPN sites in a Hub & Spoke configuration.

1 An IPv4 route is learned by the Spoke VRF and installed in the VRF's routing database. These routes are learned through an access routing protocol such as OSPF, EBGP, or RIP, or by Static Routing.

2 If the same route is learned by different means, then route selection (in the case of EBGP and MP-BGP routes) and route preference is used to determine the best route to be installed in the VRF forwarding table.

3 A locally learned IPv4 route that passes the Spoke VRF export policy is installed in the RTR's MvRib as a VPN-IPv4 route with the RD value of the Spoke VRF and a route target value of the Spoke VRF Export Route Target.

4 If the Import-Spoke Hub VRF is on the same PE as the Spoke VRF, then the VPN-IPv4 route is installed in the Import-Spoke Hub VRF's routing database. If the Import-Spoke Hub VRF is on a different PE from the Spoke VRF, then the VPN-IPv4 route is advertised to the Import-Spoke Hub VRF via the RTR Bgp Peer using MP-BGP. An Import-Spoke Hub VRF would have its Import Route Target equal to a route target value in the VPN-IPv4 route.

5 At the Import-Spoke Hub VRF, route selection and route preference is used to determine the best route to the destination.

6 The best route whether it was a locally learned VPN-IPv4 route or a remotely learned VPN-IPv4 route is further advertised to the Hub Site.

7 The Hub site re-advertises the route back as an IPv4 route to the PE via the Export-Hub VRF. Note that in a Star topology network, the Hub site would not re-advertise the route learned by the Spoke site back. It would only advertise routes in the Hub site.

8 The IPv4 route that passes the Export-Hub VRF export policy is installed in the RTR's MvRib as a VPN-IPv4 route with the RD value of the Export-Hub VRF and a route target value of the Export-Hub VRF Export Route Target.

9 If the other Spoke VRF is on the same PE as the Export-Hub VRF, then the VPN-IPv4 route is installed in the Spoke VRF's routing database. If the other Spoke VRF is on a different PE from the Export-Hub VRF, then the VPN-IPv4 route is advertised to the Spoke VRF via the RTR Bgp Peer using MP-BGP. The Spoke VRF would have its Import Route Target equal to a route target value in the VPN-IPv4 route.

10 At the Spoke VRF, route selection and route preference is used to determine the best route to the destination.

11 The best route whether it was a locally learned VPN-IPv4 route or a remotely learned VPN-IPv4 route is further advertised to the Spoke site.

Data flow

Data flows for Hub and Spoke VPN topology and Star topology are illustrated as follows:

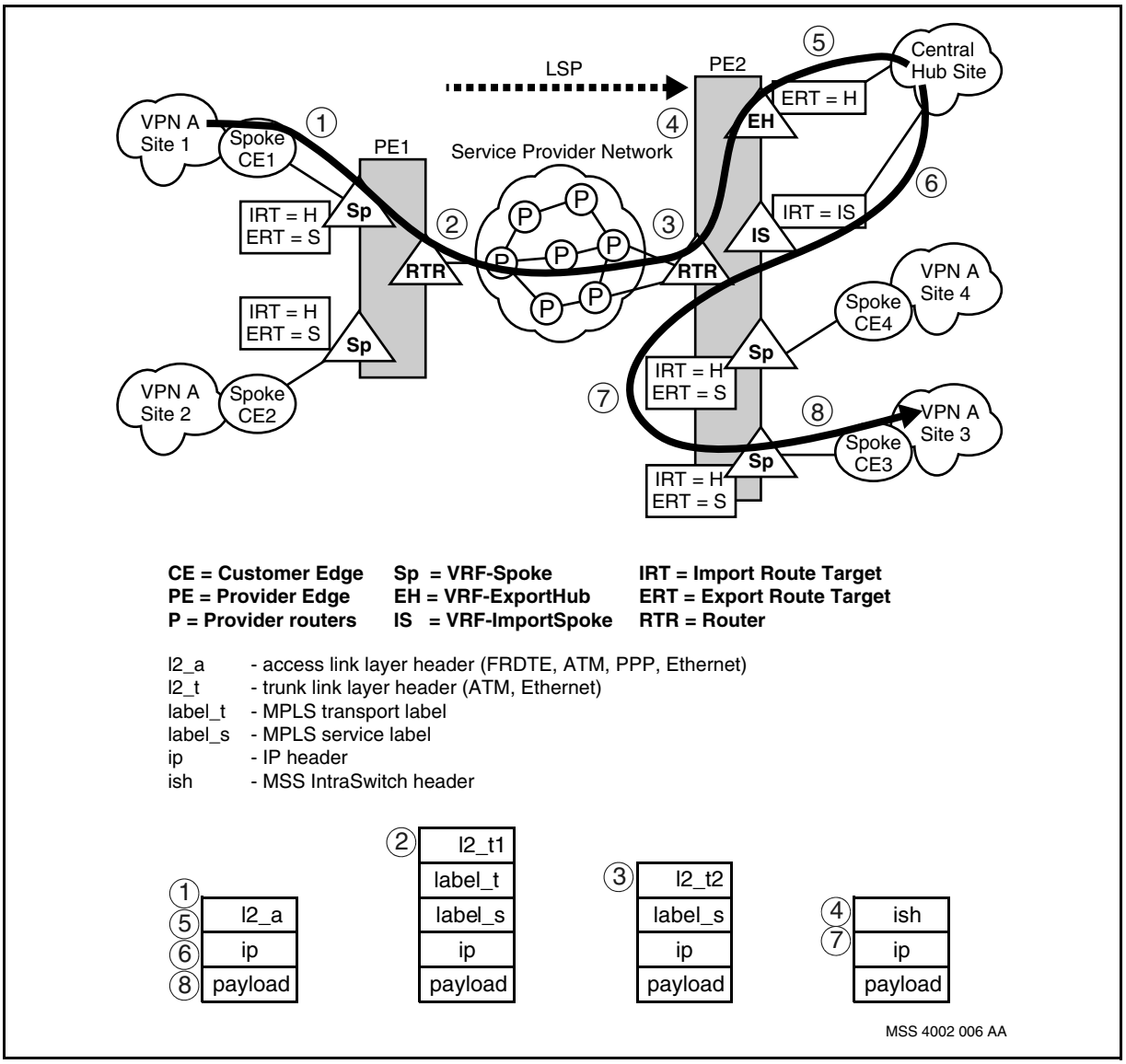
Hub and Spoke topology

There are two types of data flow in the Hub and Spoke topology:

Inter-PE flow

Figure Inter-PE Hub and Spoke data flow (page 110) depicts the corresponding data flow between the two spoke sites 1 and 3. The arrow in the figure indicates the direction of the data flow.

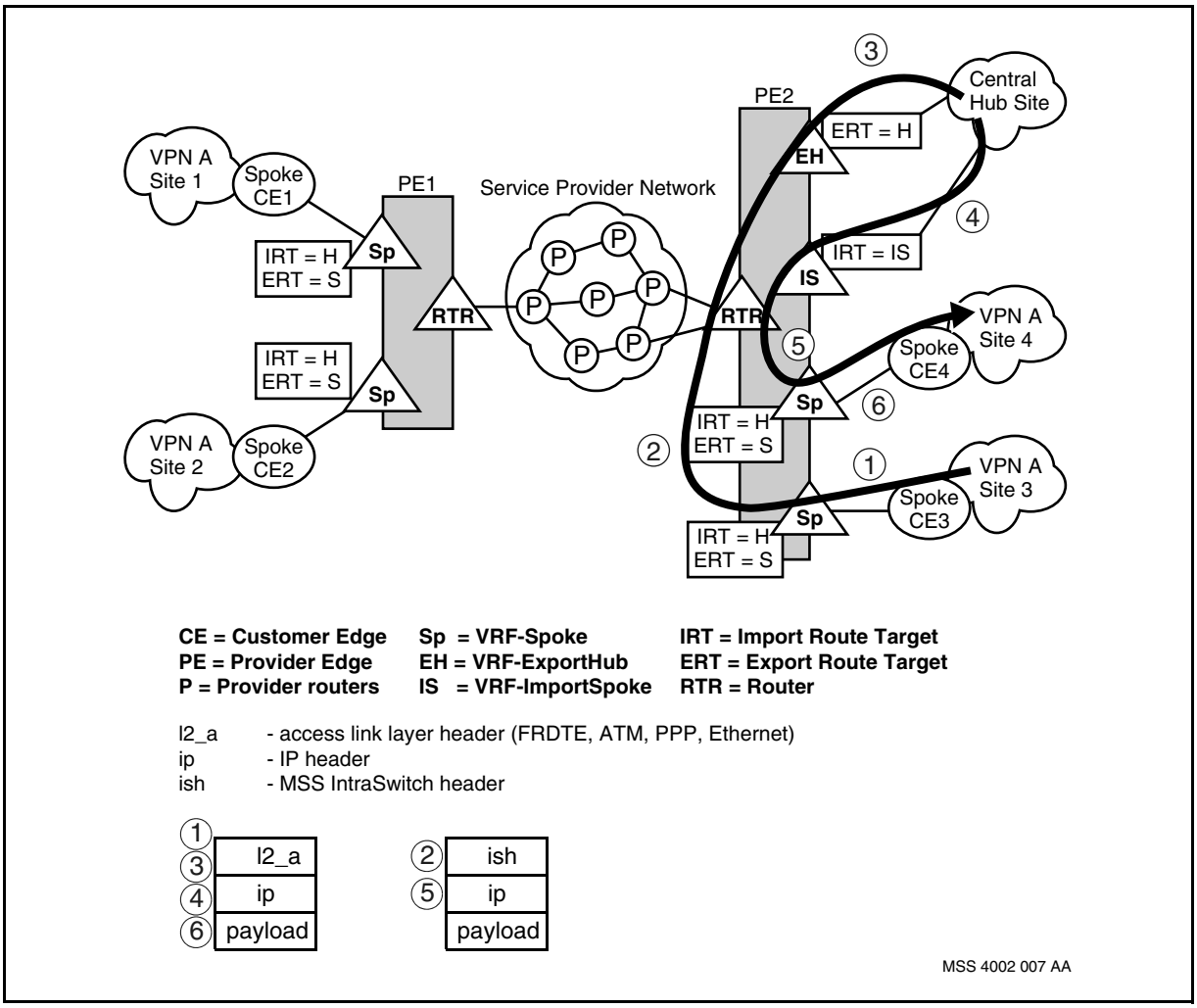
Inter-PE Hub and Spoke data flow



Intra-PE flow

Figure Intra-PE Hub and Spoke data flow (page 111) depicts the corresponding data flow between the two spoke sites 3 and 4. The arrow in the figure indicates the direction of the data flow.

Intra-PE Hub and Spoke data flow



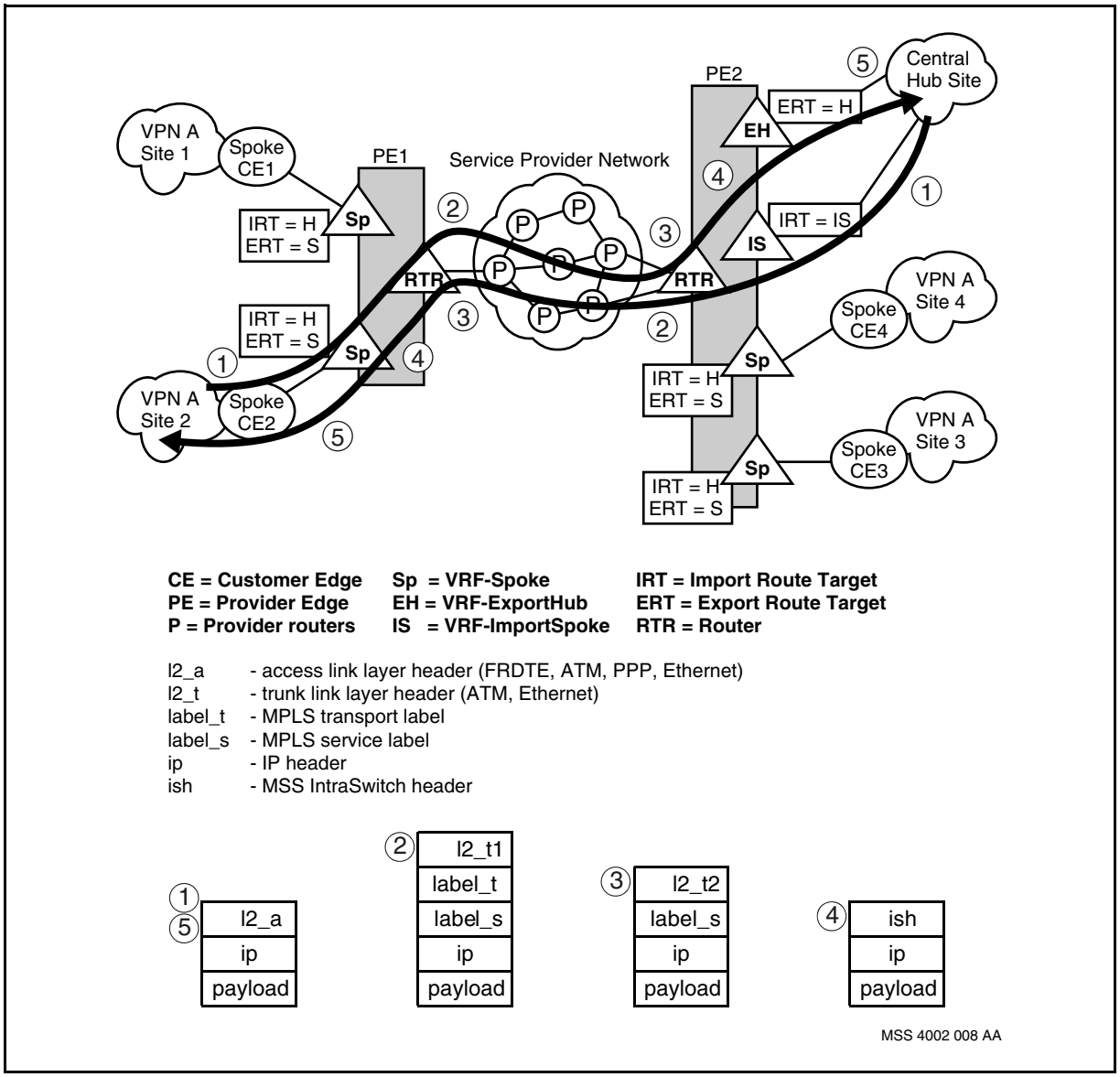
Star topology

There are two types of data flow in the Star topology:

Inter-PE flow

Figure Inter-PE Star data flow (page 112) depicts the data flow between the hub site and spoke site 2.

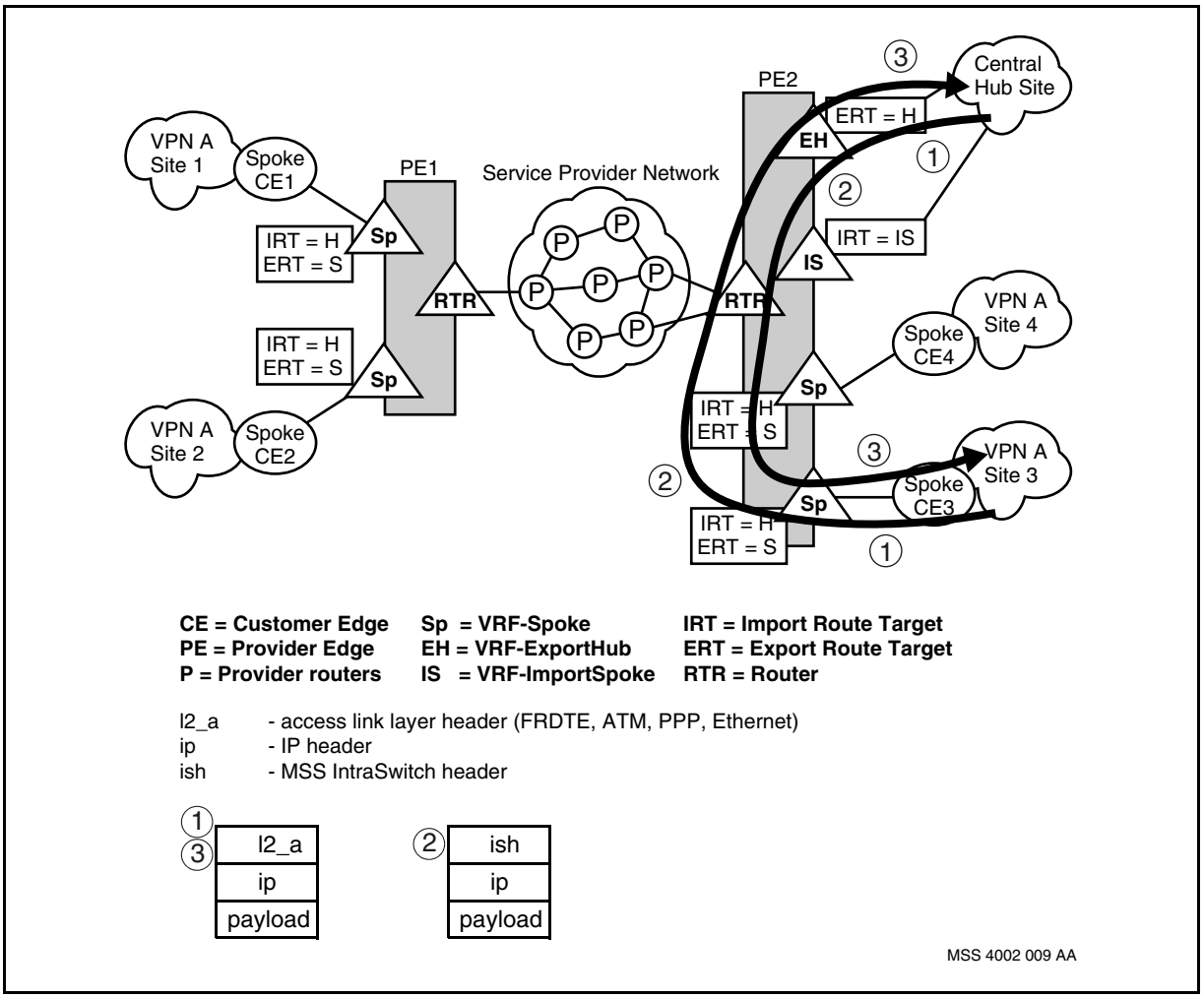
Inter-PE Star data flow



Intra-PE flow

Figure Intra-PE Star data flow (page 113) depicts the data flow between the hub site and spoke site 3.

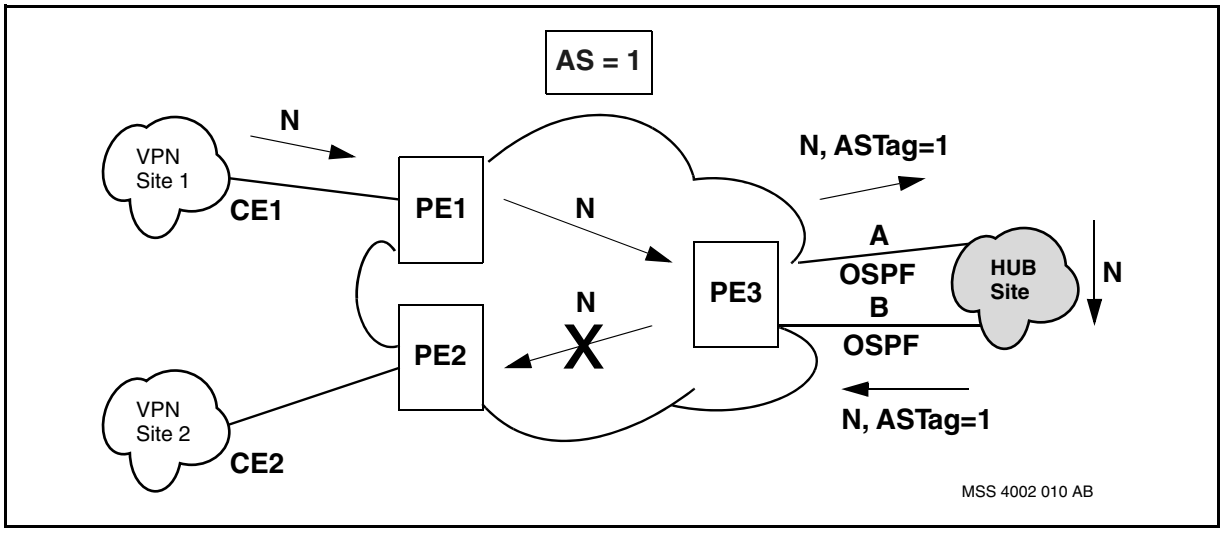
Intra-PE Star data flow



Using OSPF at the access

When OSPF is used as the access protocol on the VRF access, the AS tag is used when redistributing the MP-BGP routes into OSPF in order to prevent routing loops. The figure Using OSPF at the access (page 114) shows the default behavior when OSPF is used at the access. In this case, PE3 will encode its local AS number in the ASTag field of the LSA when it sends the route to the Hub Site via link A. On receiving the OSPF route from the Hub Site via Link B, and if a VRF export policy is configured, PE3 will check if the ASTag value is equal to its local AS number. If it finds they are equal, a routing loop is detected, and the route is not installed in the MvRib to be further advertised to other Spoke sites.

Using OSPF at the access



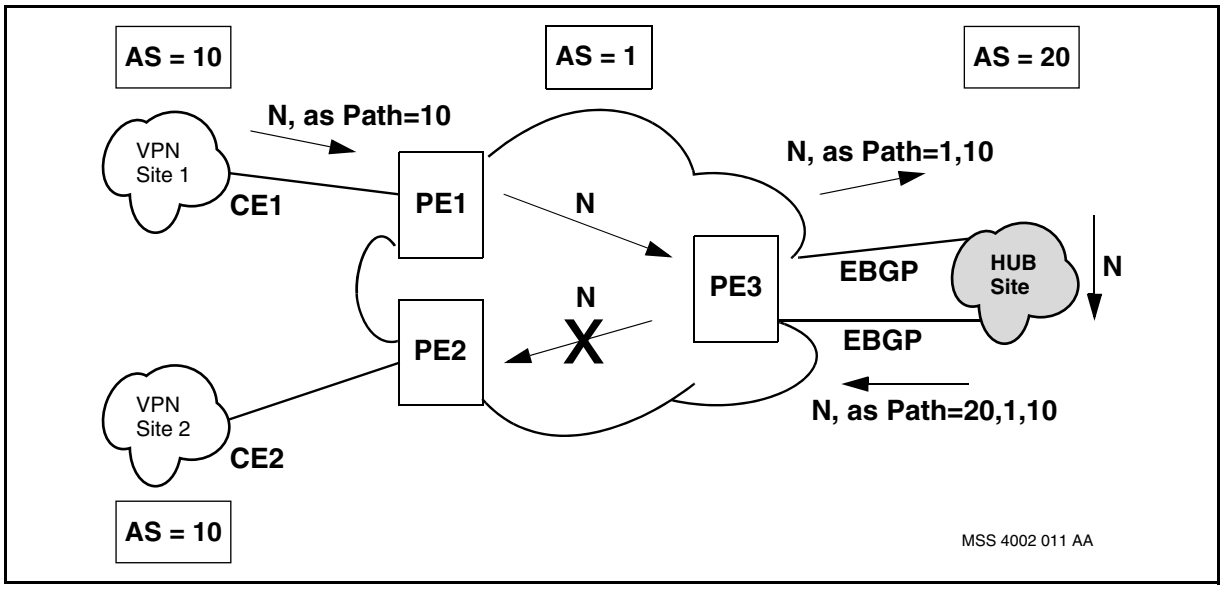
This behavior is not desirable in a Hub & Spoke topology as it prevents proper route distribution. For Hub & Spoke engineering, the attribute attachAsTag under the VRF OSPF should be disabled for the VRF which is sending routing messages to the Hub Site.

In the case of figure Using OSPF at the access (page 114), the attribute *attachAsTag* should be disabled on the Import Spoke VRF for PE3 (the OSPF interface attached to link A) so that the ASTag is not encoded in the LSA. As a result, PE3 does not receive routes from the Hub Site with AS numbers which match its local AS number, and routes are distributed correctly.

Using EBGP at the access

When EBGP is used as the access protocol on the VRF access, the AS PATH attribute is used in order to prevent routing loops. Figure Using EBGP at the access (page 115) shows the default behavior when EBGP is used at the access. In this case PE3 will encode its local AS number in the AS_PATH attribute of the UPDATE message. On receiving the EBGP route, PE3 will check if its local AS number is found in the AS_PATH attribute and if so, a routing loop is detected and the route is not installed in the MvRib to be further advertised to other Spoke sites.

Using EBGP at the access



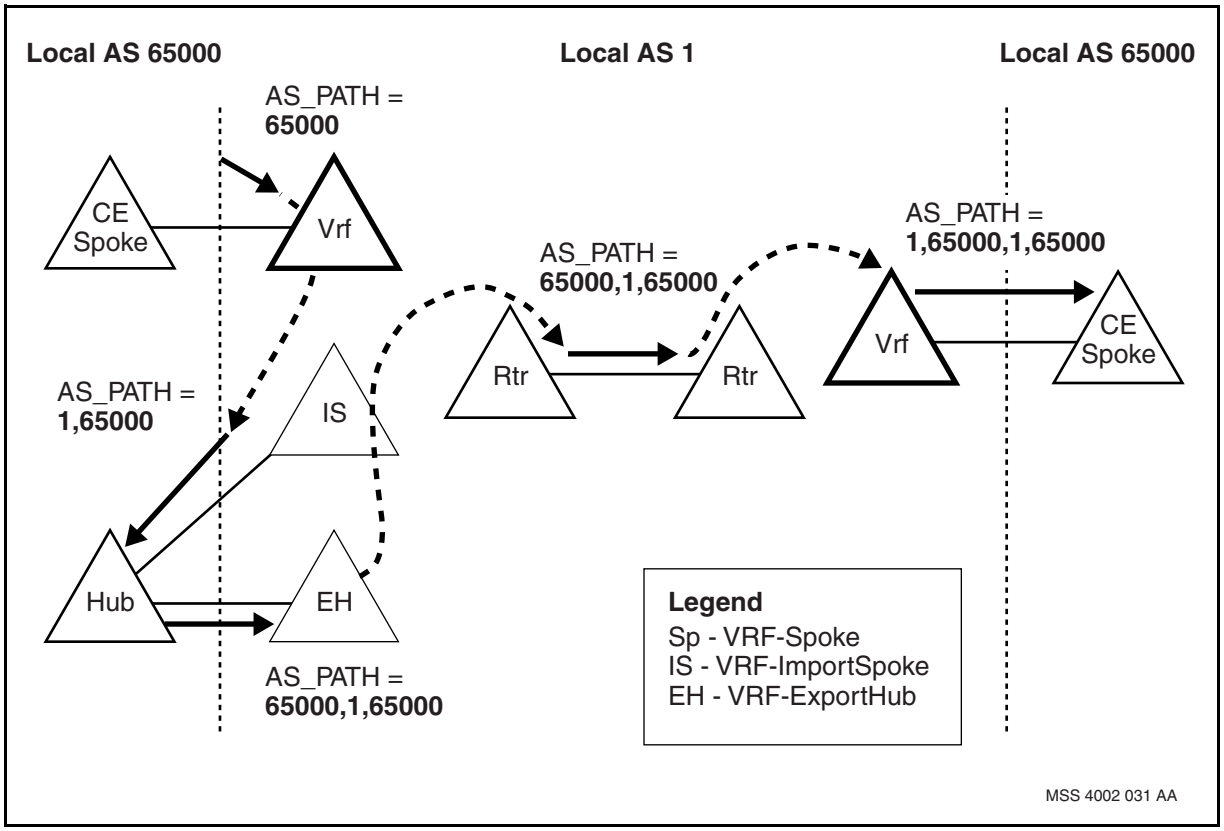
This default behavior is not desirable in a Hub and Spoke topology as it will prevent proper route distribution. For Hub and Spoke engineering, the attribute *asPathCheck* under the VRF BGP PEER should be disabled on the EH VRF. From the figure, this attribute should be configured on PE3.

Disabling the AS path check capability can result in routing loops occurring. In order to provide routing loop detection when asPathCheck is disabled, Site-Of-Origin (SOO) is used. For more information on how to use SOO, see Using Site of Origin (SOO) to prevent routing loops (page 117).

Using EBGP at the access with CEs and Hub in the same AS

Consider Hub and Spoke with EBGP on hub and spoke access (page 116) where the Hub and both CEs are in AS 65000 and the rest of the topology is in AS 1.

Hub and Spoke with EBGP on hub and spoke access



This configuration does not work as there are three points where routing loop prevention discards routes due to its Local AS being found in the AS_PATH: at the Hub, the EH, and the remote router.

To overcome this problem, the *asPathCheck* and *asOverride* attributes must be set appropriately for the RTR BGP PEERs and RTR VRF BGP PEERs.

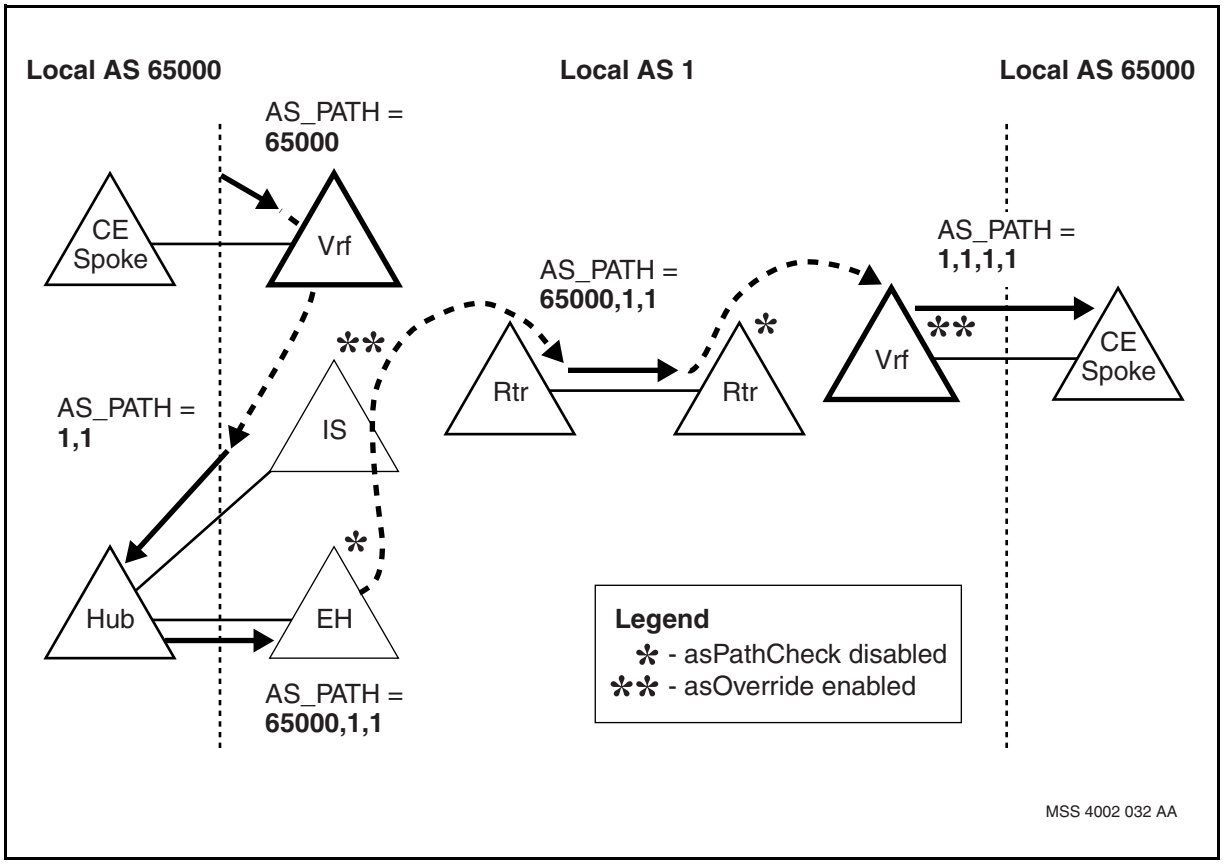
Disabling the *asPathCheck* attribute on the EH and peer RTR prevents the peer from checking a routes AS_PATH list for its local AS (routing loops), therefore preventing discarding from occurring on this basis.

With the *asOverride* attribute enabled, peers will alter the AS_PATH list such that all AS numbers in the path list found to match the EBGP peer's AS are changed to the PE's AS number.

For example, given PE AS=1 and Peer CE AS=65000, the AS_PATH would be 1,65000,1,65000. With *asOverride* enabled however, the resulting transmitted AS_PATH would be 1,1,1,1.

This configuration is illustrated in Hub and Spoke with EBGP on hub and spoke access (asPatchCheck disabled, asOverride enabled) (page 117).

Hub and Spoke with EBGP on hub and spoke access (asPatchCheck disabled, asOverride enabled)

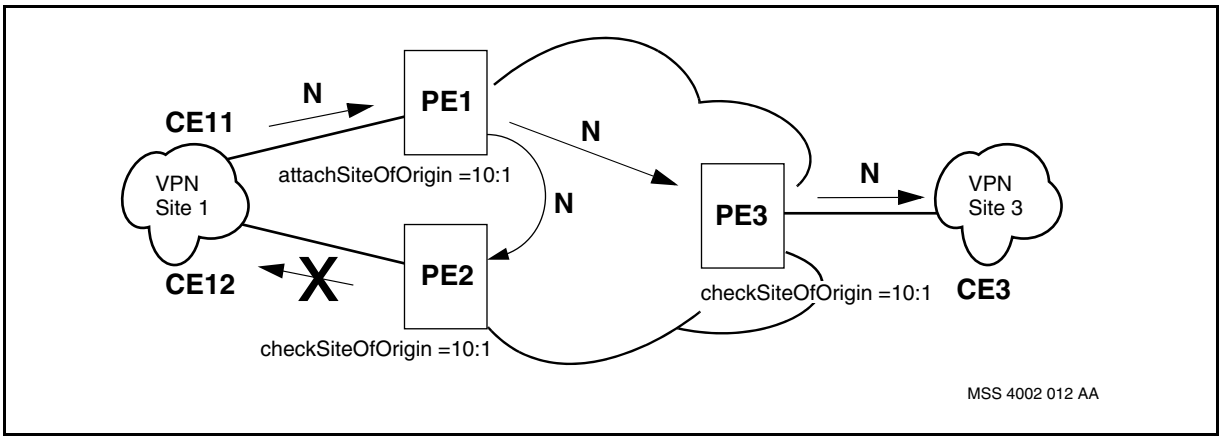


Using Site of Origin (SOO) to prevent routing loops

Site of origin (SOO) attribute can be used to detect routing loop wherever AS PATH loop detection is not available. The Site of Origin attribute, if used, uniquely identifies the set of routes learned from a particular site. This attribute is needed in some cases to ensure that a route learned from a particular site via a particular PE/CE connection is not distributed back to the same site through a different PE/CE connection. It is particularly useful if BGP is being used as the PE/CE protocol, and different sites have not been assigned distinct AS numbers.

Consider the scenario depicted in figure Use of SOO in a dual homed topology (page 118). Before a PE can redistribute a VPN-IPv4 route learned from a site, it may assign a Site of Origin attribute to the route. The attribute, attachSiteOfOrigin, under the VRF BGP PEER should be configured with an SOO value associated with Site 1. On the remote end, before the PE advertises the route to the site, it may check the Site of Origin attribute assigned to the route against a configured value, checkSiteOfOrigin attribute, to determine if the route should be advertised to the site or not.

Use of SOO in a dual homed topology



Using a default static route in a scaled environment

Depending on scalability requirements (for example a large number of spokes) it may be preferable to disable the dynamic routing protocol between the Export Hub and the Hub Site, and replace it with a default static route.

The Import Spoke VRF will continue to learn all the routes and send them to the Hub Site, but disabling dynamic routing between the Export Hub VRF (EH) and Hub Site prevents the EH from re-distributing all of the routes throughout the network.

This does not impact reachability, but prevents Forwarding Tables and MvRibs from growing to large, at hard to manage sizes.

Virtual router redundancy protocol

This section describes Nortel Multiservice Switch 7400/15000/20000 node implementation of the virtual router redundancy protocol (VRRP).

For information on configuring VRRP, see NN10600-803 *Nortel Multiservice Switch 7400/15000/20000 IP VPN Configuration Management*.

Navigation

Overview of VRRP

Nortel Multiservice Switch 7400/15000/20000 nodes use VRRP version 2, to provide router redundancy and availability to IP VPN routing. Router redundancy is achieved with VRRP virtual routers (VRs). RFC3768 describes VRRP in detail.

Nortel Multiservice Switch 7400/15000/20000 implementation of VRRP for the RFC2764 solution supports:

- IP over Ethernet with the 2-port 100BaseT Ethernet FP

Nortel Multiservice Switch 7400/15000/20000 implementation of VRRP for the RFC2547 solution supports:

- IP over Ethernet with the 4-port 10/100BaseT Ethernet FP
- IP over Ethernet with the 4-port gigabit Ethernet FP
- IP over Ethernet with the 8-port 10/100BaseT Ethernet FP
- multiple instances of VRRP VRs on each node's VR

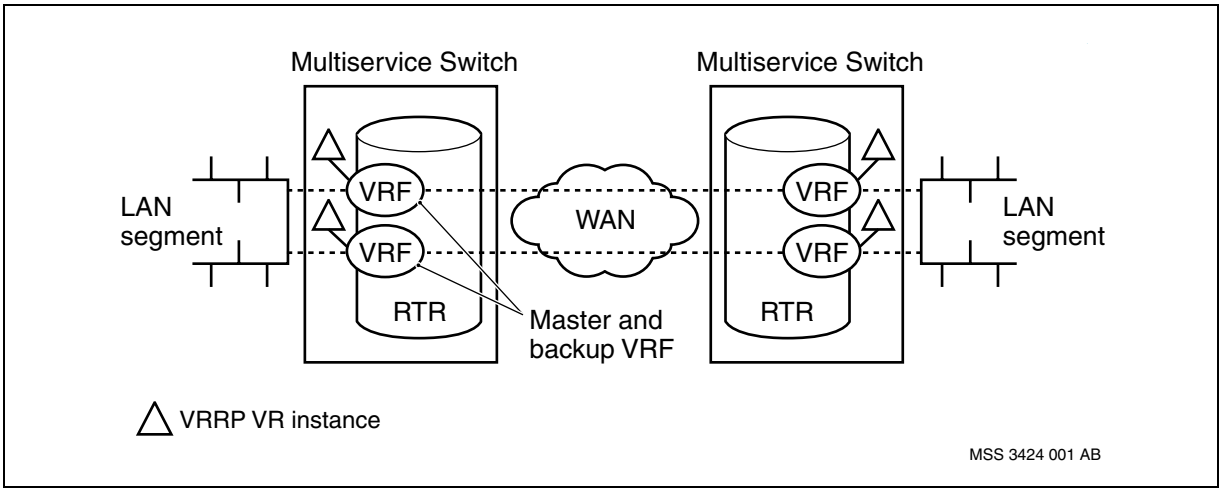
VRRP virtual routers

Implementing VRRP involves creating a VRRP virtual router (VR) made up of two or more routers in the same subnet sharing IP addresses and a virtual MAC address. Within the VRRP VR, one router (for example, a Nortel Multiservice Switch node's VRF) will act as the master and the others as backups. To an end-host, this VRRP VR appears as a single router. VRRP virtual router (page 120) depicts this arrangement. The VRRP routers communicate with each other using IP multicasts through the local Ethernet interfaces (Multiservice Switch node VR LAN protocol port).

VRRP VRs can communicate using the local LAN media. The protected VRFs configuration does not have an impact on the VRRP functionality. A VRRP VR consists of VRFs that are connected to the customer edge (CE) routers.

Each VRRP VR has a priority value that determines if it will act as a master or backup. The VRRP master router typically owns the IP addresses of the VRRP VR and has a priority of 255. If none of the VRRP VRs own an IP address, the VRRP VR with the higher priority is the master. In the case of equal priority, the higher interface IP address is the master.

VRRP virtual router



Router redundancy

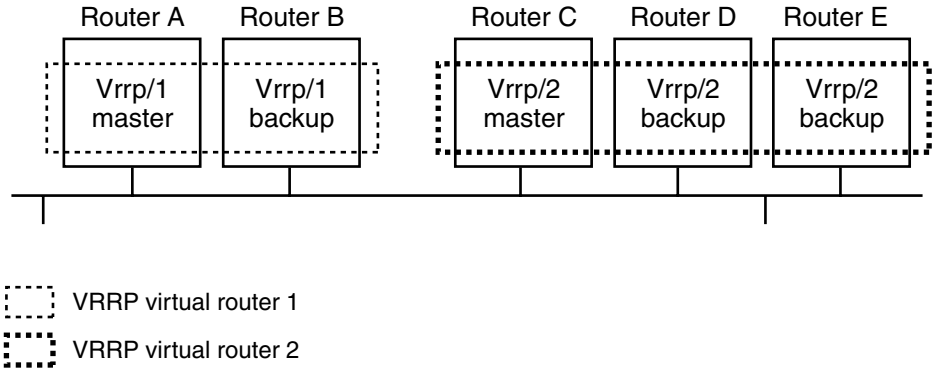
You can configure a single Nortel Multiservice Switch VRF to be protected by a single VRRP VR. How you configure VRRP redundancy depends on the unique characteristics of your network. The figure, Example VRF redundancy topologies (page 121), depicts the possible scenario where a LAN or VLAN segment uses multiple VRFs protected by a single VRRP VR for redundant access to the RFC2547 VPN.

VRRP provides redundancy on the Ethernet interface in both port-mode and VLAN-mode. When a protocol port with VRRP is associated with an Ethernet interface that is operating in port-mode, VRRP redundancy functionality behavior is unchanged. When a protocol port with VRRP is associated with an Ethernet interface that is operating in VLAN-mode, VRRP functionality behaves the same as port-mode, but only for that specific VLAN. A VLAN that is linked to a protocol port without VRRP configured is not redundant.

Attention: On the 4-port 10/100BaseT Ethernet, 4-port gigabit Ethernet, and 8-port 10/100BaseT Ethernet FPs, only one instance of VRRP VR per interface is supported. Interior gateway protocols, other than static protocols, on the same interface as VRRP VR must be in passive mode. Also, load balancing is not supported on these FPs.

For information about VRRP router redundancy with RFC2547, see VPN route forwarder redundancy with RFC2547 (page 121).

Example VRF redundancy topologies



Note: Routers A, B, C, and D are either Multiservice Switch VRFs or a combination of Multiservice Switch VRFs and external non-Multiservice Switch VRRP-compliant routers.

MSS 3424 003 AA

VPN route forwarder redundancy with RFC2547

Nortel Multiservice Switch 7400/15000/20000 nodes providing RFC2547 solution with VRRP on the same LAN/VLAN segment are designated as master VRRP VR and backup VRRP VR. The master VRRP VR enables its virtual MAC (VMAC) on the Ethernet interface and the backup VRRP VR disables the same VMAC on its Ethernet interface. The VRF with a master VRRP VR instance will then route traffic from the LAN/VLAN that is destined to the VMAC. After receiving an ARP message from the master VRRP VR, the hosts start sending traffic destined to the configured default route with an Ethernet header containing a MAC DA set to the VRRP VMAC. At intermittent

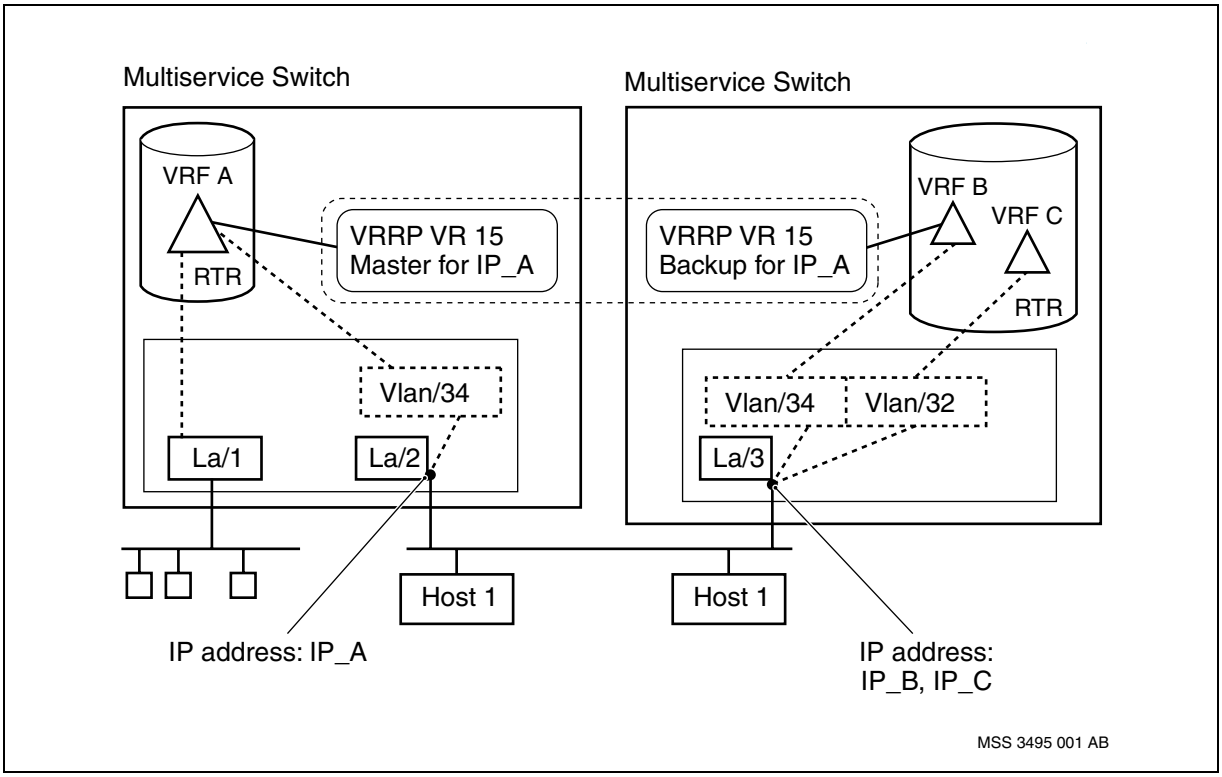
intervals, the master VRRP VR instance transmits a heartbeat control packet that is multicast onto the LAN/VLAN segment. The backup VRRP VR expects to receive that heartbeat message within a configured time interval, after which it assumes the master VRRP VR is no longer providing service on the LAN/VLAN segment.

If the master VRRP VR fails, the backup VRRP VR assumes the master role, and enables the VMAC on its Ethernet interface. The new master (originally the backup) VRRP VR would then start routing traffic from the hosts that are sending traffic to MAC DA equal to the VRRP VR VMAC.

If the original master VRRP VR recovers, it re-establishes its master role by sending out heartbeat messages at configured advertisement intervals. When the active master VRRP VR receives the heartbeat, it reverts to the backup role again. The active master VRRP VR relinquishes its master role because the heartbeat message from the original master has a higher priority.

The figure, VRRP configuration to provide Multiservice Switch RFC2547 VRF redundancy with VLANs (page 123), depicts an example of two Multiservice Switch nodes providing the VRRP redundancy for the RFC2547 solution. In this instance, VRF A and VRF B are on the same VLAN segment, identified by VID=34. Both VRFs are backed up by VRRP VR 15. The VRRP VR instance providing redundancy to VRF A is elected master, while the VRRP VR instance providing redundancy to VRF B assumes the backup role. The master enables its VMAC on the Ethernet interface and the backup disables the same VMAC on its Ethernet interface. RFC2547 VRF A routes traffic from La/2 Vlan/34. Hosts 1 and 2 receive an ARP from the master VRRP VR, associating IP address, IP_A, with its VMAC as the MAC DA. The hosts send traffic to the statically configured default route with an Ethernet header containing a MAC DA set to the VMAC of VRRP VR 15.

VRRP configuration to provide Multiservice Switch RFC2547 VRF redundancy with VLANs



The VRRP process

When operational, VRRP VRs are in one of three states: master, backup or initialize. VRFs with a master VRRP VR instance perform the routing duties for addresses associated with the VRRP VR. VRFs with a backup VRRP VR instance monitor the availability of the master VRRP VR instance. A VRF with a VRRP VR instance transitions to the initialize state when its *Vrrp* component is locked. The priority parameter of a VRRP VR determines if it acts as a master or backup. The table, Summary of the VRRP virtual router states in relation to network conditions (page 124), summarizes the states of the VRRP VRs under different conditions.

Summary of the VRRP virtual router states in relation to network conditions

network condition	VRRP virtual router state and activities	
	VRRP virtual router A priority = 255	**VRRP virtual router B priority = 100**
start up	master	backup
normal	master As master, VRRP VR A multicasts messages at a configured advertisement interval, advertising to the backup VRRP VR or VRRP VRs that it is operational.	backup As backup, VRRP VR B listens for the multicast advertisement. When it receives the advertisement message with a higher priority, an advertisement wait timer resets.
VRRP VR A failure	na	master VRRP VR B transitions to master when the advertisement wait timer expires.

Intermediate system to intermediate system Protocol

Intermediate system to intermediate system (ISIS) is a link-state routing protocol suitable for use as an Interior Gateway Protocol (IGP) within an Autonomous System (AS).

ISIS was originally specified in ISO 10589 as a protocol for exchanging CLNP routing information. The protocol was adapted, allowing it to be used for IP routing. The changes to the protocol are specified in RFC 1195. In Nortel Multiservice Switch systems the ISIS protocol is used for IP routing only.

Navigation

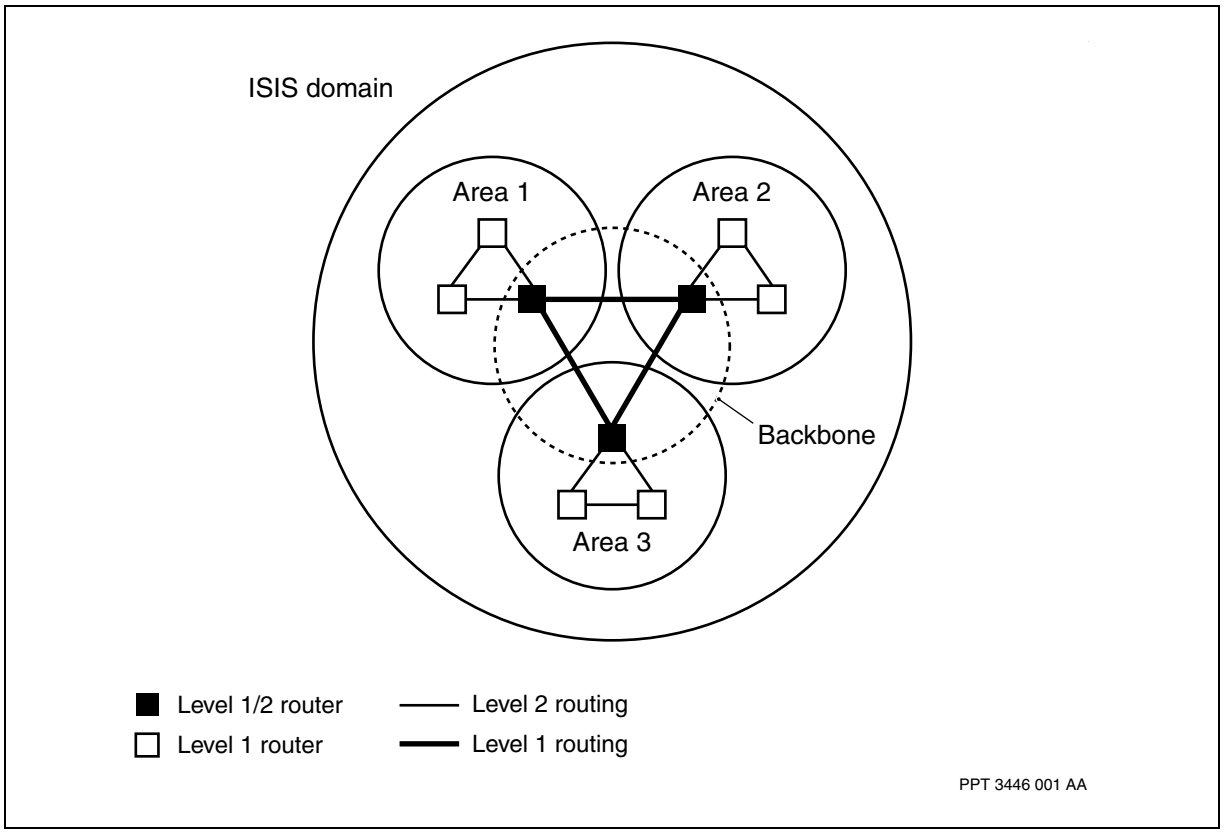
ISIS terminology

An ISIS routing domain is a network in which all the routers run ISIS to support intra-domain exchange of routing information. Routers within such a domain are called Intermediate Systems (ISs). An ISIS domain can be partitioned into smaller segments known as areas. Routers belonging to a common area engage in Level 1 routing, which involves the exchange of intra-area IP prefix information. Border routers in different areas may exchange inter-area routing information, this process is known as Level 2 routing. For information, see Level 1/Level 2 routing (page 126).

A router engaged in Level 1 routing generates a Level 1 link state packet (LSP). A Level 1 LSP contains intra-area routing information. A router engaged in Level 2 routing generates a Level 2 LSP, which contains inter-area routing information.

For Nortel Multiservice Switch systems, only Level 1 routing is supported. Multiservice Switch nodes will not form an adjacency with a router that does not belong to the same area. A Multiservice Switch node will not generate or accept Level 2 LSPs.

Level 1/Level 2 routing



ISO-based node identification

Even when used for IP routing only, ISIS is based on ISO concepts. For example, ISIS uses an ISO addressing scheme for identifying an ISIS node.

ISO network layer addresses are called Network Service Access Points (NSAPs). An NSAP address consists of an Area Address, System ID, and Network Selector (NSEL). The Area Address uniquely identifies an area within an ISIS domain; the System ID uniquely identifies a node within an area. The NSEL identifies a network layer service on the node. On an ISIS node used exclusively for IP routing, there is only one network layer service, the ISIS routing engine itself. When the routing engine is specified as the network layer service, the NSEL is set to zero and the NSAP is called a Network Entity Title (NET).

Multiple NETs are permitted per node. These NETs must have the same System ID and are differentiated only by the Area Address. This does not mean that the router is connected to multiple separate areas, rather the router belongs to one area, which is known by multiple synonymous Area Addresses. Normally, a router would be configured with only a single Area Address and would therefore have only a single NET. However, the ability to configure multiple Area Addresses is useful for migration purposes. For example, renumbering, merging, or splitting areas. This allows operators to perform a migration of their ISIS area topologies without suffering service interruptions during the reconfiguration period. Nortel Multiservice Switch ISIS implementation allows for provisioning of up to 3 NETs per ISIS instance to support this functionality. These NETs must have identical System ID components and only differ in Area Address.

Multiservice Switch ISIS implementation uses a fixed-size, non-configurable System ID length of 6 bytes.

As an example, the following address illustrates the ISO format:

```
49.000001.12ca.0065.90ab.00
```

The first portion (49.0001) is the Area Address. The area Address can be from 1 to 13 bytes in length. The first byte of the Area Address (49) is referred to as the Authority and Format Identifier (AFI). The next 6 bytes (12ca.0065.90ab) are the System ID. The System ID can be any 6 bytes that allow the ISIS node to be uniquely identified within the domain. Common methods for choosing a System ID include using one of the MAC Addresses on the node or some derivation of an IP address belonging to the node (e.g. the IP address 47.202.187.168 could be transformed into the System ID 0472.0218.7168). The last byte (00) is the NSEL.

Default route

ISIS Level 1 Routers exchange only intra-area prefix information and, therefore, do not know about any routes outside their areas. Backbone routers exchange Level 2 LSPs with routers in other areas, but also exchange Level 1 LSPs with routers in their own area. A backbone router will set the attached bit in its Level 1 LSP, to indicate that it is attached to the backbone. A Level 1 router will install a default route to the nearest Level 2 router that has set the attached bit. All traffic for destinations outside the local area will be sent to that backbone router.

Nortel Multiservice Switch ISIS implementation functions as a Level 1 router only. If a Level 2 router with its attached bit set exists in the ISIS area, a default route to that Level 2 router will be installed into the routing table by ISIS.

Media types

The ISIS protocol distinguishes 2 main types of link layer media, general topology (point-to-point) and broadcast.

In Nortel Multiservice Switch systems, ISIS supports the following types of media:

- Broadcast: Ethernet (4 port Gig E card only)
- General Topology: ATM (ATM PQC cards only)

Direct VR-to-VR connectivity

An IP virtual private network (VPN) consists of multiple customer virtual routers (VR), each representing a private customer VPN site. In addition, a single customer VR can support multiple VPN customer sites.

In a direct VR-to-VR IP VPN, carriers connect customer VRs in the same IP VPN directly, using dedicated virtual circuits (VC) between different Nortel Multiservice Switch nodes. Full VC-meshing between customer VRs is recommended in this VPN configuration, but not required. Customer VRs in the same IP VPN use internal gateway protocols (IGPs) to exchange routing information.

Navigation

- Dedicated layer 2 connections (page 129)
- IP VPN accounting (page 130)
- Routing information between VRs (page 131)

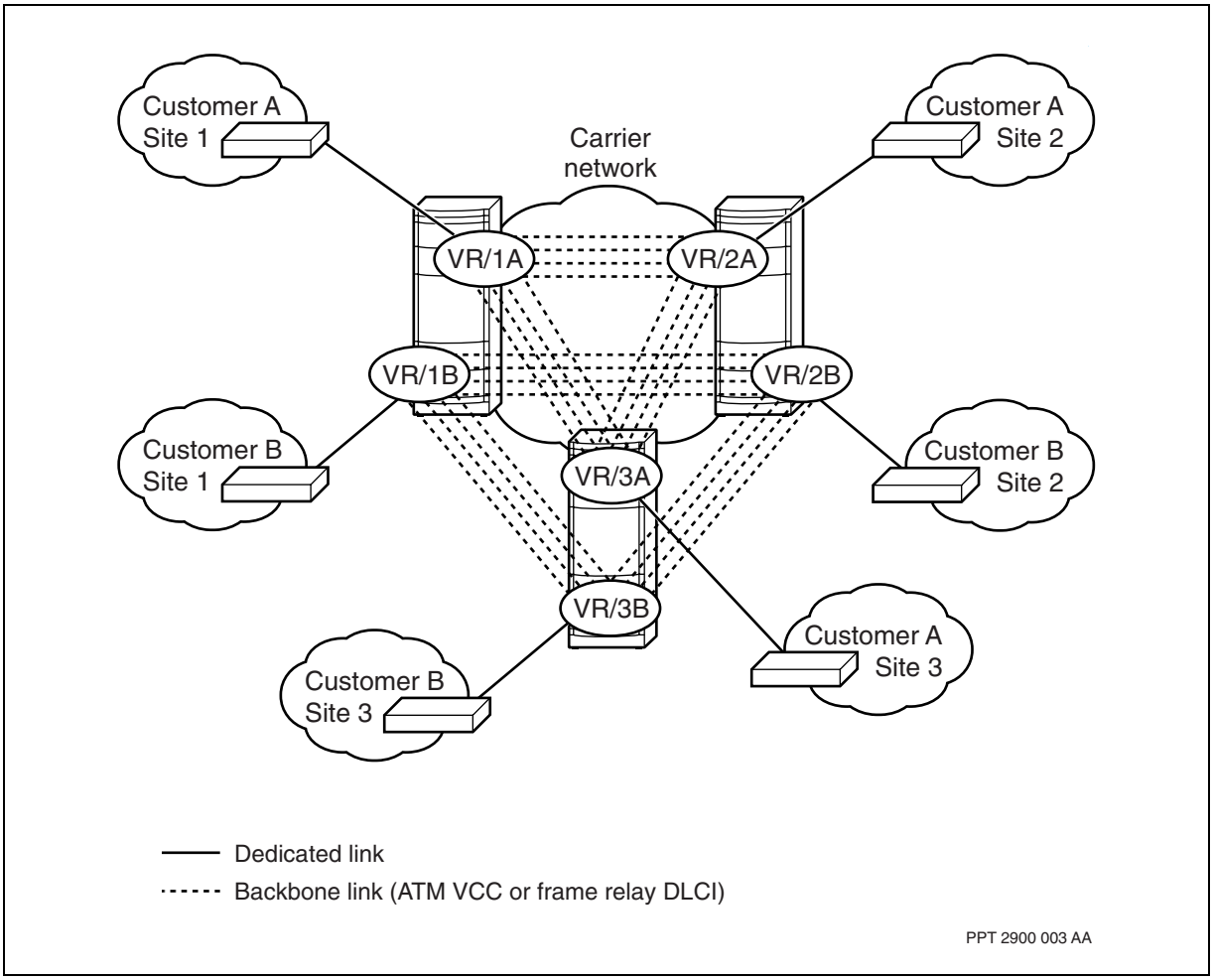
Dedicated layer 2 connections

For secure site-to-site intranet connectivity within an IP VPN, carriers can connect customer VRs directly through dedicated layer 2 connections.

Carriers can use dedicated VC connections to connect customer VRs over the public network. Each customer VR connects to a remote customer VR on another Nortel Multiservice Switch node through a backbone logical connection (either ATM VCC or frame relay DLCI).

In this configuration, the carrier must configure one or more separate VCs on the customer VR for each remote customer VR to which it connects.This deployment configuration requires full VC meshing to achieve connectivity between customer VRs within an IP VPN. See the figure Dedicated backbone VCs between customer VRs (page 130).

Dedicated backbone VCs between customer VRs



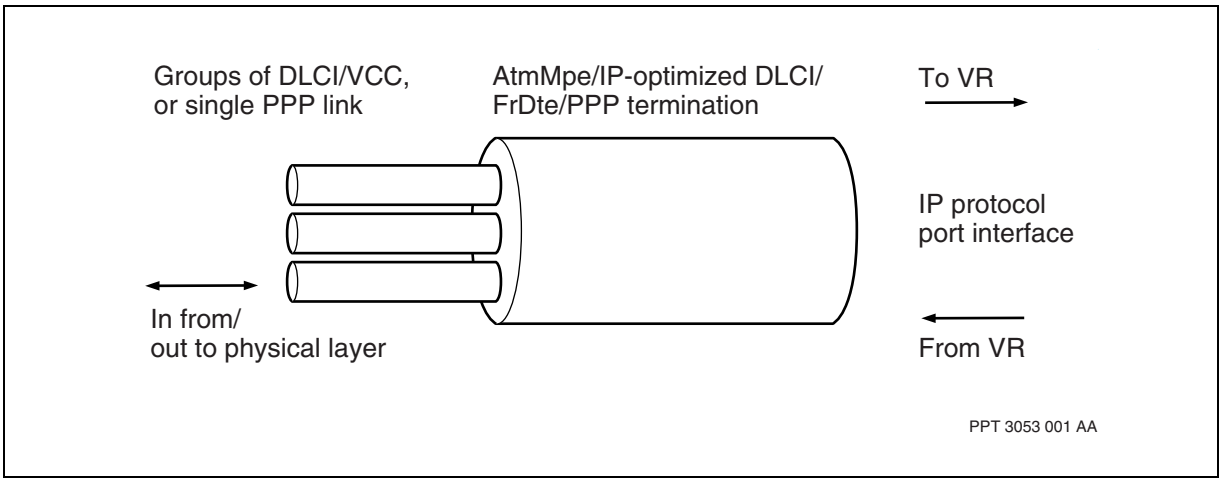
IP VPN accounting

IP accounting usage statistics are collected for each VR that is part of an IP VPN. Network interface statistics are generated for layer 2 connections. The statistics generated provide the total number of connections that terminate at the AtmMpe, IP-optimized DLCI, FrDte and PPP traffic ports. Since the VPN address is unknown, the accounting records gathered at the outgoing traffic ports are aggregate statistics of the number of packets received and sent by the VR to the network. These statistics are measured by class of service CoS. For further information on CoS, see NN10600-808 *Nortel Multiservice Switch 7400/15000/20000 Layer 3 Traffic Management Fundamentals*. The figure Mapping of VCs to a protocol port (page 131) illustrates the mapping process for AtmMpe, IP-optimized DLCI, FrDte and PPP VCs to a protocol port.

Only single-ended accounting is possible for VR-to-VR connections. One accounting record is generated for each protocol port.

For more information on IP VPN accounting, see NN10600-800 *Nortel Multiservice Switch 7400/15000/20000 IP Technology Fundamentals* and NN10600-560 *Nortel Multiservice Switch 7400/15000/20000 Accounting*.

Mapping of VCs to a protocol port



Routing information between VRs

In a direct VR-to-VR configuration, IGPs such as the routing information protocol (RIP), open shortest path first (OSPF), and internal border gateway protocol (IBGP) configured between VRs ensure the exchange of routing information within the IP VPN.

On the local VR, the carrier configures a WAN access protocol port running IP to connect to the backbone (for example, ATM MPE). On the remote VR, the carrier configures a corresponding WAN access protocol port that belongs to the same subnet. The carrier configures one or more IGPs (that is, IBGP, OSPF, or RIP) to run between the WAN access protocol ports. The layer 2 media distribution handles IP address resolution protocol (ARP) queries and packet forwarding to the local and remote protocol ports that belong to the same subnet.

Procedure conventions

This document uses the following procedure conventions:

- You can enter commands using full component and attribute names, or you can abbreviate them. The commands used in the procedures contain the full component and attribute names in the first instance. In the second instance, the component and attribute names are abbreviated. For more information on abbreviating component and attribute names, see NN10600-060 *Nortel Multiservice Switch 7400/15000/20000 Component Reference*. All component and attribute names are formatted in italics.
- The introduction of every procedure states whether you must perform the procedure in operational mode or provisioning mode. For more information on these modes, see Operational mode (page 132) or Provisioning mode (page 133).
- When you complete a procedure, you can verify your changes and then activate them as the new node configuration. For more information on completing configuration changes and exiting provisioning mode, see Activating configuration changes (page 133).

Operational mode

Procedures contained within this document can either be performed in operational mode or provisioning mode. When you initially log into a node, you are in operational mode. Nortel Multiservice Switch systems use the following command prompt when you are in operational mode:

```
#>
```

where:
is the current command number

In operational mode, you work with operational components and attributes. In operational mode, you can

- list operational components and display operational attributes to determine the current operating parameters for the node
- control the state of parts of the node by locking and unlocking components

- set certain operational attributes and enter commands to perform diagnostic tests

Provisioning mode

To change from operational mode to provisioning mode, type the following command at the operator prompt:

```
start Prov
```

Only one user can be in provisioning mode at a time. Nortel Multiservice Switch systems use the following command prompt whenever you are in provisioning mode:

```
PROV #>
```

where:

is the current command number

In provisioning mode, you work with the provisionable components and attributes that contain the current and future configurations of the node. You can add and delete components, and display and set provisionable attributes. For information on completing the configuration changes, exiting provisioning mode, and returning to operational mode see Activating configuration changes (page 133).

For information on operational and provisionable attributes, see NN10600-060 *Nortel Multiservice Switch 7400/15000/20000 Component Reference*.

Activating configuration changes

Several procedures in this document ask that you complete the configuration changes. When you complete the configuration changes, you are activating the configuration changes, confirming that you want to activate them, and saving the changes. You are instructed to complete the configuration changes only at the end of procedures that you perform in provisioning mode.

CAUTION

Activating a provisioning view can affect service

Activating a provisioning view can result in a CP reload or restart, causing all services on the node to fail. See NN10600-050 *Nortel Multiservice Switch 7400/15000/20000 Command Reference*, for more information.

CAUTION
Risk of service failure
When you activate the provisioning changes (see step 3), you have 20 minutes to confirm these changes. If you do not confirm these changes within 20 minutes, the shelf resets and all services on the node fail.

1 Verify that the provisioning changes you have made are acceptable.

```
check Prov
```

Correct any errors and then verify the provisioning changes again.

2 If you want to store the provisioning changes in a file, save the provisioning view.

```
save -f(<filename>) Prov
```

3 If you want these changes as well as other changes made in the edit view to take effect immediately, activate, confirm, and commit the provisioning changes.

```
activate Prov
confirm Prov
commit Prov
```

4 End the provisioning session.

```
end Prov
```

Nortel Multiservice Switch 7400/15000/20000

IP VPN Technology Fundamentals

Publication: NN10600-802
Document status: Standard
Document issue: 7.2S1
Document date: March 2006
Product release: PCR7.2 and up
Job function: Product Fundamentals
Type: NTP
Language type: U.S. English