CISCO SYSTEMS

Definitive MPLS Network Designs

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

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

ciscopress.com

Definitive MPLS Network Designs

Jim Guichard
François Le Faucheur
Jean-Philippe Vasseur

Cisco Press

800 East 96th Street
Indianapolis, IN 46240 USA

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

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

Static Routing Design Considerations

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

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

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

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

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

PE-CE BGP Routing Design Considerations

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

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

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

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

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

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

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

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

PE-CE IGP Routing Design Considerations

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

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

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

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

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

Specifics of the OSPF Service Deployment

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

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

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

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

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

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

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

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

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

Specifics of the EIGRP Service Deployment

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

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

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

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

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

Example 3-6 *EIGRP SoO Attribute Configuration Template*

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

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

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

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

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

IP Address Allocation for PE-CE Links

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

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

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

Controlling Route Distribution with Filtering

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

Example 3-7 *Maximum Routes Configuration Template*

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

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