CHAPTER 4

Layer 3 Implementation

This chapter contains the following major topics:

End-to-End Routing

In the multiple tenants' Data Center (DC) environment, the tenants must be separated. In order for the clients to access the resources in the DC, the clients must have route reachability to the DC. This solution uses Virtual Routing and Forwarding (VRF)-Lite technology to separate tenants and Border Gateway Protocol (BGP) and static routes as the routing protocols.

This section presents the following topics:

VRF-Lite in the Data Center

VRF is a key element in the DC that allows multiple instances of a routing table to coexist within the same router at the same time. Routing instances are independent, providing a separated environment for each customer. The same or overlapping IP addresses can be used without conflicting with each other.

Each VRF instance has its own:

- IP routing table
- Derived forwarding table
- Set of interfaces
- Set of routing protocols and routing peers that inject information into the VRF

VRF-Lite is a feature that equals to VRF without the need to run Multiprotocol Label Switching (MPLS). VRF-Lite uses input interfaces to distinguish routes for different customers and forms virtual packet forwarding tables by associating one or more Layer 3 (L3) interfaces with each VRF instance.

The VRF interface can be physical, such as Ethernet ports, or logical, such as a subinterface or VLAN Switched Virtual Interface (SVI). An end-to-end VRF-Lite instance supports network virtualization and provides total separation between customer networks. Communication between customers is not possible within the cloud and backbone network.

In the ASA and ACE, a similar concept "context" can be created. In the ASA, a single security appliance can be partitioned into multiple virtual devices, known as "security contexts." Each context is an independent device, with its own security policy, interfaces, and administrators. Multiple contexts are similar to having multiple stand-alone devices. In the ACE, a virtual environment, called a "virtual context," can be created using ACE virtualization. A single ACE appears as multiple virtual devices, and each is configured and managed independently. A virtual context allows for closely and efficiently managing system resources, ACE users, and the services that are provided to customers.

In this solution, every tenant has its own VRF instance in the ASR 1000, a VRF-Lite instance in the Nexus 7000, and its own contexts in the ASA and ACE.

Tenant Load Distribution

Redundancy and load balancing are a must in the DC design. From the network topology, we see there are redundant devices and links in every layer. To best use redundancy, traffic is divided into "left" and "right" to achieve load balance for both southbound and northbound traffic (Figure 4-1 and Figure 4-2). In a normal situation, the "left" traffic will use the left half of the topology (ASR1K-1, Nexus7k-Agg1), and the "right" traffic will use the right half of the topology (ASR1K-2, Nexus7K-Agg2). If the link or node fails, the traffic will converge to other nodes and links. If the "left" traffic reaches the right half of the topology by some reason, the first choice is going through the cross link and returning to the left half of the topology. As BGP is the routing protocol for every tenant, community is a natural choice to transmit the load-balancing information. After receiving the routes with community information, the ASR 1000 and Nexus 7000 use local preference to prefer the routes.

Figure 4-1 Topology Divided to Left and Right, Northbound Traffic

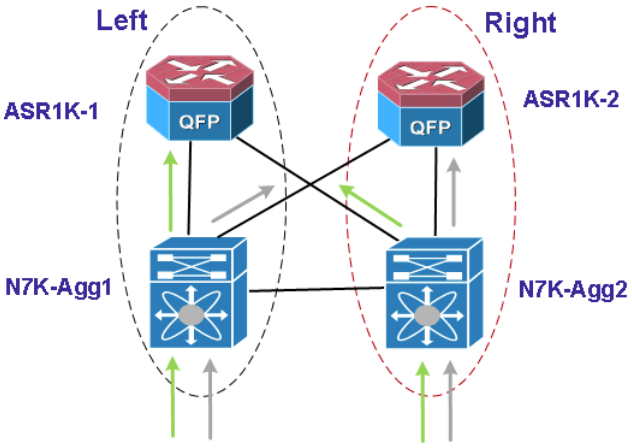
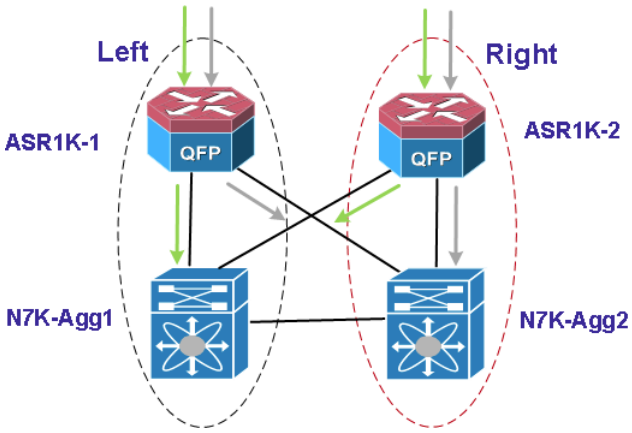


Figure 4-2 Topology Divided to Left and Right, Southbound Traffic



Implementation in the Nexus 7000

The Nexus 7000 runs BGP with the local PE and advertises the server subnets to BGP (redistribute to BGP from connect or static). To achieve load balance, a route-map is used to add community during the redistribution. In the Nexus 7000 Agg1, community 31:31 is attached to the "left" traffic, and 31:32 is attached to the "right" traffic routes. In the Nexus 7000 Agg2, community 32:31 is attached to the "left" traffic, and 32:32 is attached to the "right" traffic routes.

Implementation in the ASR 1000 (Local PE)

The ASR 1000 injects a default route to BGP and advertises the route to the Nexus 7000 for every tenant. To achieve load balance, ASR1K-1 adds community 21:21, and ASR1K-2 adds community 22:22 when the route is redistributed to BGP.

Overview of the flow

For the north to south traffic, when the ASR 1000 routers receive the routes of the server via BGP, it will use the community to set the local preference. For example, if the PE1 receives "left" traffic routes with community 31:31 (from Agg1), it will set local preference 10000, the same route, but with community 32:31 (received from Agg2), and it will set preference 5000. The PE1 will choose Agg1 as the next hop.

For the south to north traffic, the Nexus 7000 routers receive the default route via BGP from both the ASR 1000 PE1 and PE2. For the "left" traffic, the Agg1 sets the local preference higher for the routes learned from PE1 and sets PE1 as the next hop. For the "right" traffic, the Agg2 sets the local preference higher for the routes learned from PE2 and sets PE2 as the next hop.

Configuration Examples and Useful Commands

Below are configuration examples using tenant bronze_1.

Nexus 7000 a1

```
router bgp 65501

 template peer-policy PREFER->PE1
    send-community
    route-map PREFER-PE1 in
    next-hop-self

 template peer-policy ibgp-policy
    next-hop-self
```

```
 vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM
    neighbor 10.3.1.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.3.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.34.4
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community

route-map SERVER-NET-SET-COMM permit 10
  match ip address prefix-list SERVER-NET
  set community 31:31

route-map SERVER-NET-SET-COMM permit 20

route-map PREFER-PE1 permit 10
  match community COMM-1
  set local-preference 60000

route-map PREFER-PE1 permit 20

ip prefix-list SERVER-NET seq 5 permit 11.2.0.0/16

ip community-list standard COMM-1 permit 21:21
ip community-list standard COMM-2 permit 22:22
```

Nexus 7000 a2

```
router bgp 65501

 template peer-policy PREFER->PE1
    send-community
    route-map PREFER-PE1 in
    next-hop-self

 template peer-policy ibgp-policy
    next-hop-self

 vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM-PATH2
    neighbor 10.3.2.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.4.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.34.3
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community
```

```
route-map SERVER-NET-SET-COMM-PATH2 permit 10
  match ip address prefix-list SERVER-NET
  set community 32:31

route-map SERVER-NET-SET-COMM-PATH2 permit 20

route-map PREFER-PE1 permit 10
  match community COMM-1
  set local-preference 60000

route-map PREFER-PE1 permit 20

ip prefix-list SERVER-NET seq 5 permit 11.0.0.0/8 le 24

ip community-list standard COMM-1 permit 21:21
ip community-list standard COMM-2 permit 22:22
```

ASR 1000 PE1

```
router bgp 109

 template peer-policy DC2_PEER_POLICY
  route-map DC2_PATH_PREFERENCE in
  route-map default out
  default-originate route-map default-condition
  send-community both

 address-family ipv4 vrf customer_bronze1
  neighbor 10.3.1.2 remote-as 65501
  neighbor 10.3.1.2 activate
  neighbor 10.3.1.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.3.4.2 remote-as 65501
  neighbor 10.3.4.2 activate
  neighbor 10.3.4.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family
 !
route-map DC2_PATH_PREFERENCE permit 10
 match community PREFER-N7K1
 set local-preference 10000
!
route-map DC2_PATH_PREFERENCE permit 20
 match community PREFER-N7K2
 set local-preference 1000
!
route-map DC2_PATH_PREFERENCE permit 30
 match community BACKUP
 set local-preference 5000
!
route-map DC2_PATH_PREFERENCE permit 40

route-map default permit 10
 match ip address prefix-list default
 set community 21:21

route-map default-condition permit 10
 match ip address prefix-list default-condition
 set community 21:21

ip prefix-list default seq 5 permit 0.0.0.0/0
ip prefix-list default-condition seq 5 permit 169.0.0.0/8

ip community-list standard PREFER-N7K1 permit 31:31
ip community-list standard PREFER-N7K2 permit 32:32
```

```
ip community-list standard BACKUP permit 32:31
```

ASR 1000 PE2

```
router bgp 109

 template peer-policy DC2_PEER_POLICY
  route-map DC2_PATH_PREFERENCE in
  route-map default out
  default-originate route-map default-condition
  send-community both

 address-family ipv4 vrf customer_bronze1
  neighbor 10.3.2.2 remote-as 65501
  neighbor 10.3.2.2 activate
  neighbor 10.3.2.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.3.3.2 remote-as 65501
  neighbor 10.3.3.2 activate
  neighbor 10.3.3.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family

route-map DC2_PATH_PREFERENCE permit 10
 match community PREFER-N7K1
 set local-preference 1000
!
route-map DC2_PATH_PREFERENCE permit 20
 match community PREFER-N7K2
 set local-preference 10000
!
route-map DC2_PATH_PREFERENCE permit 30
 match community BACKUP
 set local-preference 5000
!
route-map DC2_PATH_PREFERENCE permit 40
!
route-map default permit 10
 match ip address prefix-list default
 set community 22:22

route-map default-condition permit 10
 match ip address prefix-list default-condition
 set community 22:22

ip prefix-list default seq 5 permit 0.0.0.0/0
ip prefix-list default-condition seq 5 permit 169.0.0.0/8

ip community-list standard PREFER-N7K1 permit 31:31
ip community-list standard PREFER-N7K2 permit 32:32
ip community-list standard BACKUP permit 31:32
```

Below are useful commands for the ASR 1000 and Nexus 7000. ASR 1000:

```
sh ip bgp vpnv4 vrf customer_bronze1 x.x.x.x
sh ip route vrf customer_bronze1
show route-map
sh ip community-list
```

Nexus 7000

```
sh ip bgp vrf customer_bronze1 0.0.0.0
sh ip bgp vrf customer_bronze1 x.x.x.x
sh ip route vrf customer_bronze1
show route-map
show ip community-list
```

ASR 1000 PE Implementation Overview

In order for the ASR 1000 to receive routes from its client networks across the Service Provider cloud, it must peer with the client PE router and receive VPNv4 prefixes from these peers. Also, in order to have Internet reachability, it must peer with the required IPv4 Internet routers. To achieve Service Provider client and Internet client reachability from the DC, the ASR 1000 conditionally injects a default route into the appropriate routing table, since it has all route prefixes in its routing table. Using this approach allows for efficient use of the routing table of devices/routers in the DC network.

To receive VPNv4 prefixes from the client PE routers, the ASR 1000 must run MP-iBGP with these routers. This requires running MPLS on the ASR 1000 and enabling Label Distribution Protocol (LDP) on relevant interfaces. In this solution, LDP is enabled on two ASR 1000 interfaces, one to the core and the other on the L3 interface that connects it to the other ASR 1000. The core interfaces on the ASR 1000 are 10G.

The ASR 1000 also has an Internet Protocol version 4 (IPv4) External Border Gateway Protocol (eBGP) neighborship with the Nexus 7004 aggregation routers. Using this peering, it advertises a default route into these routers and receives server specific network routes from them. Each tenant has a sub-interface in the VRF specific to the tenant and runs tenant specific eBGP sessions in these VRF instances between the ASR 1000 and Nexus 7004 aggregation routers.

ASR 1000 Core Routing Configuration

The core configuration on the ASR 1000 routers involves Open Shortest Path First (OSPF) and MPLS configuration, as well as the Multiprotocol Internal Border Gateway Protocol (MP-iBGP) configuration required to receive VPNv4 prefixes from client PE routers. Routing optimizations are also included for faster convergence. This includes Nonstop Forwarding (NSF) and Nonstop Routing (NSR) for OSPF, graceful-restart for MPLS, and BGP PIC Core and Edge and BGP graceful restart.

MPLS and OSPF Configuration

```
mpls ldp graceful-restart

router ospf 1
 nsr
 nsf
!
dc02-asr1k-pe1#sh run int te0/0/0
Building configuration...

dc02-asr1k-pe1#sh run int te0/0/0
Building configuration...

Current configuration : 298 bytes
!
interface TenGigabitEthernet0/0/0
 description uplink-to-core
 ip address 10.4.21.1 255.255.255.0
 ip ospf 1 area 0
 load-interval 30
 carrier-delay up 30
 plim qos input map mpls exp  5  queue strict-priority
 mpls ip
 cdp enable
 service-policy input wan-in
 service-policy output wan-out
end
```

```
dc02-asr1k-pe1#sh run int te0/1/0
Building configuration...

Current configuration : 313 bytes
!
interface TenGigabitEthernet0/1/0
 description connect to pe2
 ip address 10.21.22.1 255.255.255.0
 ip ospf network point-to-point
 ip ospf 1 area 0
 carrier-delay up 30
 plim qos input map mpls exp  5  queue strict-priority
 mpls ip
 cdp enable
 service-policy input wan-in
 service-policy output wan-out
end
```

Additional Routing Optimization Configuration

```
ip routing protocol purge interface
cef table output-chain build favor convergence-speed
cef table output-chain build indirection recursive-prefix non-recursive-prefix
cef table output-chain build inplace-modify load-sharing
```

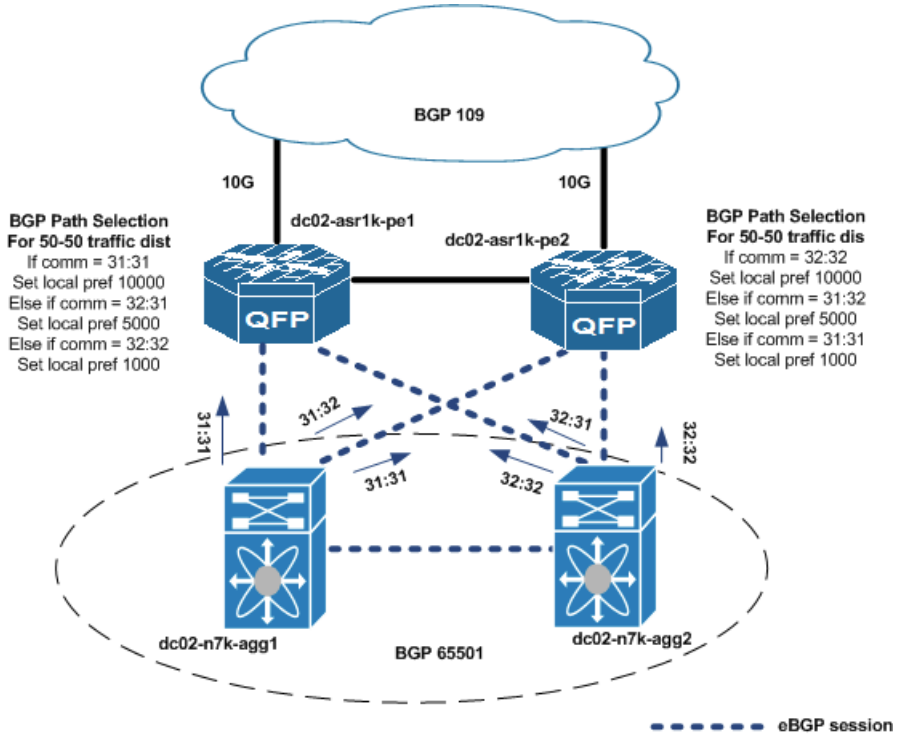
ASR 1000 DC Routing Configuration

BGP is used for DC routing between the ASR 1000 PE and the Nexus 7000 aggregation routers. The DC prefixes and server addresses are advertised by the Nexus 7000 aggregation routers, while a default route is conditionally advertised for each tenant configured on the aggregation routers. DC addresses advertised include server private and public prefixes, and the ASA public addresses used for VPN termination and the Copper tenants.

To ensure effective distribution of traffic to each of the Nexus 7000 routers, a BGP load-balancing scheme is configured. This scheme ensures 50-50 traffic distribution for all configured DC tenants by using the community value advertised by the aggregation routers to determine the preferred path. BGP path selection is based on the local preference set based on these received community values. See Tenant Load Distribution for a complete understanding of the BGP scheme used to forward traffic to the aggregation Nexus 7000 switches. BGP community values and BGP local preferences are used to determine a secondary path to be used if the primary path used by BGP fails. Using this scheme, both ASR 1000 PEs will forward traffic to the aggregation routers if the primary paths used to send tenant traffic fails. BGP PIC Edge optimization is configured to achieve faster convergence when the BGP paths fails. Both the primary and secondary BGP paths will be installed in the routing table with the secondary installed as a repair-path.

Figure 4-3 and Figure 4-4 show a diagrammatic overview of the BGP scheme used on the ASR 1000 for routing and overview of the secondary paths used by BGP to forward traffic if the primary paths fails.

Figure 4-3 ASR 1000 BGP DC Routing Overview

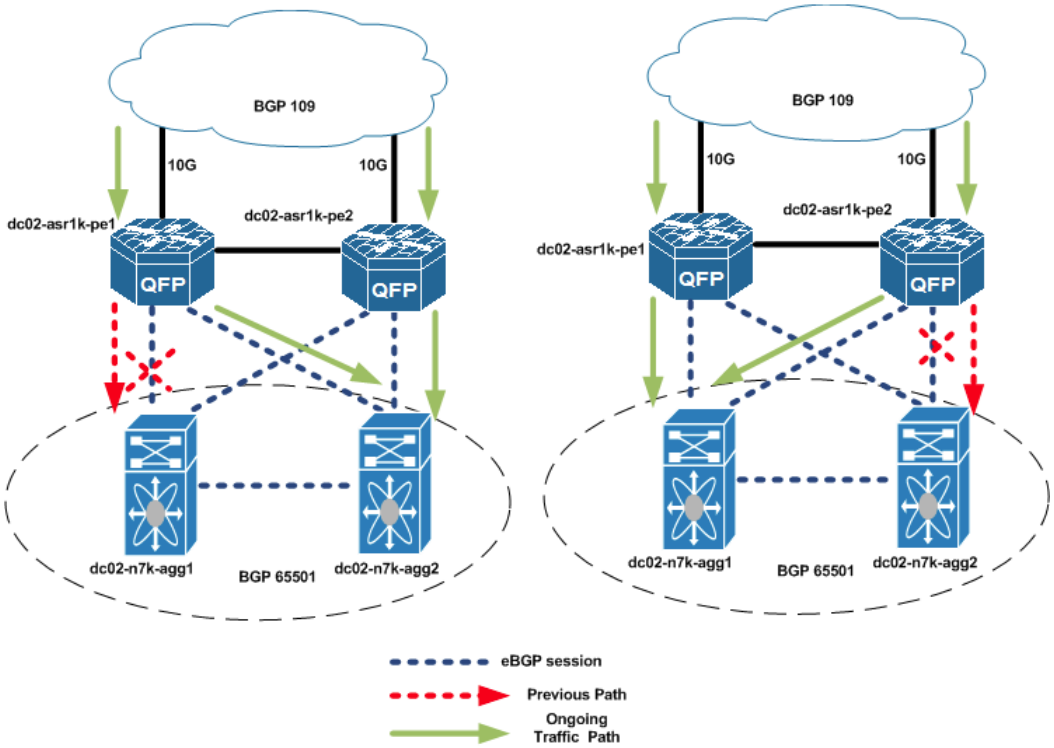


ASR 1000 BGP Routing Configuration For Sample Tenant

```
router bgp 109
 template peer-policy DC2_PEER_POLICY
  route-map DC2_PATH_PREFERENCE in
  route-map default out
  default-originate route-map default-condition
  send-community both
 exit-peer-policy
 !
address-family vpnv4
  bgp additional-paths install
  bgp recursion host
!
address-family ipv4 vrf customer_gold2
  neighbor 10.1.1.2 remote-as 65501
  neighbor 10.1.1.2 activate
  neighbor 10.1.1.2 inherit peer-policy DC2_PEER_POLICY
  neighbor 10.1.4.2 remote-as 65501
  neighbor 10.1.4.2 activate
  neighbor 10.1.4.2 inherit peer-policy DC2_PEER_POLICY
 exit-address-family
!

dc02-asr1k-pe1#sh ip bgp vpnv4 vrf customer_gold1  11.1.0.0
BGP routing table entry for 21:1:11.1.0.0/16, version 347789
Paths: (2 available, best #2, table customer_gold1)
  Additional-path-install
  Advertised to update-groups:
     998
  Refresh Epoch 1
  65501
```

```
    10.1.4.2 from 10.1.4.2 (10.1.5.3)
      Origin incomplete, metric 0, localpref 5000, valid, external, backup/repair
      Community: 32:31
      Extended Community: RT:21:1 , recursive-via-connected
      mpls labels in/out 2362/nolabel
      rx pathid: 0, tx pathid: 0
  Refresh Epoch 1
  65501
    10.1.1.2 from 10.1.1.2 (10.1.5.2)
      Origin incomplete, metric 0, localpref 10000, valid, external, best
      Community: 31:31
      Extended Community: RT:21:1 , recursive-via-connected
      mpls labels in/out 2362/nolabel
      rx pathid: 0, tx pathid: 0x0
dc02-asr1k-pe1#


dc02-asr1k-pe1#sh ip route vrf customer_gold1 repair-paths 11.1.0.0

Routing Table: customer_gold1
Routing entry for 11.1.0.0/16
  Known via "bgp 109", distance 20, metric 0
  Tag 65501, type external
  Last update from 10.1.1.2 1d17h ago
  Routing Descriptor Blocks:
  * 10.1.1.2, from 10.1.1.2, 1d17h ago, recursive-via-conn
      Route metric is 0, traffic share count is 1
      AS Hops 1
      Route tag 65501
      MPLS label: none
      MPLS Flags: NSF
    [RPR]10.1.4.2, from 10.1.4.2, 1d17h ago, recursive-via-conn
      Route metric is 0, traffic share count is 1
      AS Hops 1
      Route tag 65501
      MPLS label: none
      MPLS Flags: NSF
```

Figure 4-4 ASR 1000 BGP Routing with Failure of Primary Path



With failure of the primary BGP path for a tenant, traffic will be rerouted to the repair-path/secondary path associated with the tenant prefix to ensure 50-50 traffic distribution on both ASR 1000 PEs for all configured DC tenants. In Figure 4-4, if the primary path for a tenant, dc02-asr1k-pe1->dc02-n7k-agg1 or dc02-asr1k-pe2->dc02-n7k-agg2, fails, then based on the BGP routing configuration, traffic will be routed on the dc02-asr1k-pe1->dc02-n7k-agg2 or dc02-asr1k-pe2->dc02-n7k-agg1 paths respectively.

Layer 3 Implementation on the Nexus 7004 Aggregation

In this solution, a pair of Nexus 7004 switches are placed in the Aggregation layer. This section presents the following topics:

VRF-Lite

Cisco NX-OS supports VRF instances. Multiple VRF instances can be configured in a Nexus 7000 switch. Each VRF contains a separate address space with unicast and multicast route tables for IPv4 and IPv6 and makes routing decisions independent of any other VRF instance. Interfaces and route protocols can be assigned to a VRF to create virtual L3 networks. An interface exists in only one VRF instance.

Each Nexus 7K router has a default VRF instance and a management VRF instance. The management VRF instance is for management purposes only, and only the mgmt0 interface can be in the management VRF instance. All L3 interfaces exist in the default VRF instance until they are assigned to another VRF instance. The default VRF instance uses the default routing context and is similar to the global routing table concept in Cisco IOS.

Below is an example of creating a VRF instance and assigning a VRF membership to the interfaces for tenant customer_silver1.

```
vrf context customer_silver1

interface Vlan501
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.1.2/24
  no ipv6 redirects
  no ip arp gratuitous hsrp duplicate

interface Vlan601
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.2.2/24
  no ipv6 redirects

interface Vlan701
  vrf member customer_silver1
  no ip redirects
  ip address 11.2.3.2/24
  no ipv6 redirects

interface Vlan1821
  vrf member customer_silver1
  no ip redirects
  ip address 113.3.1.2/24
  no ipv6 redirects

interface port-channel343.501
  vrf member customer_silver1
  ip address 10.2.34.3/24

interface Ethernet3/9.501
  vrf member customer_silver1
  ip address 10.2.1.2/24
  no ip arp gratuitous hsrp duplicate

interface Ethernet4/9.501
  vrf member customer_silver1
  ip address 10.2.3.2/24
  no ip arp gratuitous hsrp duplicate
```

Routing protocols can be associated with one or more VRF instances. In this solution, BGP is used as the routing protocol. For example, below is the routing configuration for tenant customer_silver1 on the Nexus 7000 Agg1 device.

```
router bgp 65501
  vrf customer_silver1
    graceful-restart-helper
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM
      additional-paths send
      additional-paths receive
    neighbor 10.2.1.1
```

```
    remote-as 109
    address-family ipv4 unicast
      inherit peer-policy PREFER->PE1 1
  neighbor 10.2.3.1
    remote-as 109
    address-family ipv4 unicast
      send-community
  neighbor 10.2.34.4
    remote-as 65501
    address-family ipv4 unicast
      inherit peer-policy ibgp-policy 1
      no send-community
```

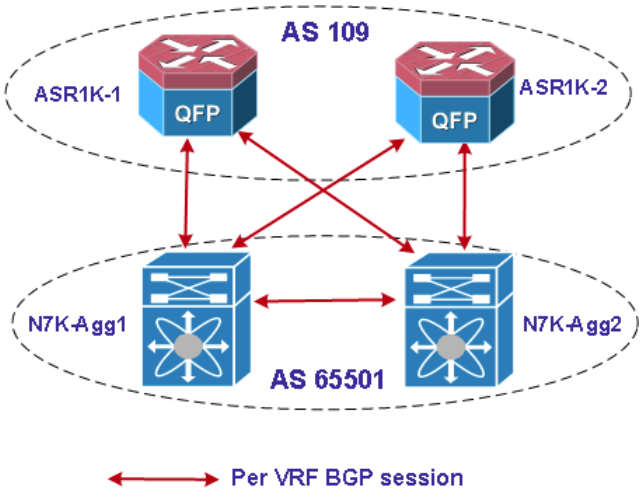
Below are useful commands.

```
show vrf XXX
show vrf XXX interface
show run vrf XXX
show ip route vrf XXX
show ip bgp vrf XXX
```

BGP

BGP is the routing protocol used in the tenants to convey routing information. Direct and/or static routes are distributed to BGP from the Nexus 7000 routers and are advertised out to the ASR 1000 routers. The Nexus 7000 routers also learn the default route from the ASR 1000 routers through BGP. Figure 4-5 shows the BGP sessions per tenant.

Figure 4-5 BGP Sessions Per Tenant



For example, for the tenant customer_bronze1, the Nexus 7000 Agg1 builds an eBGP session with ASR1k-1 using the e3/9.801 subinterface, and an eBGP session with ASR1k-2 using the e4/9.801 subinterface. To provide redundancy, there is also an iBGP session between the Nexus 7000 Agg1 and Agg2 using the po343.801 subinterface. In the iBGP session, we use the next-hop-self option. The Nexus 7000 Agg2 builds an eBGP session with ASR1k-2 using the e4/9.801 subinterface, and an eBGP session with ASR1k-1 using the e3/9.801 subinterface. As discussed in the Tenant Load Distribution section, the community is needed to convey the load-balancing information. Community sends the information to the eBGP peers (ASR 1000 routers).

Below are the related configurations of the Nexus 7000 routers for tenant customer_bronze1.

Nexus Agg

```
interface port-channel343.801
  vrf member customer_bronze1
  ip address 10.3.34.3/24

interface Ethernet3/9.801
  vrf member customer_bronze1
  ip address 10.3.1.2/24
  no ip arp gratuitous hsrp duplicate

interface Ethernet4/9.801
  vrf member customer_bronze1
  ip address 10.3.3.2/24
  no ip arp gratuitous hsrp duplicate

router bgp 65501

 template peer-policy PREFER->PE1
    send-community
    route-map PREFER-PE1 in
    next-hop-self

 template peer-policy ibgp-policy
    next-hop-self

 vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM
    neighbor 10.3.1.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.3.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.34.4
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community

route-map SERVER-NET-SET-COMM permit 10
  match ip address prefix-list SERVER-NET
  set community 31:31

route-map SERVER-NET-SET-COMM permit 20

route-map PREFER-PE1 permit 10
  match community COMM-1
  set local-preference 60000

route-map PREFER-PE1 permit 20

ip prefix-list SERVER-NET seq 5 permit 11.2.0.0/16

ip community-list standard COMM-1 permit 21:21
ip community-list standard COMM-2 permit 22:22
```

Nexus Agg2

```
interface port-channel434.801
```

```
  vrf member customer_bronze1
  ip address 10.3.34.4/24

interface Ethernet3/9.801
  vrf member customer_bronze1
  ip address 10.3.4.2/24
  no ip arp gratuitous hsrp duplicate

interface Ethernet4/9.801
  vrf member customer_bronze1
  ip address 10.3.2.2/24
  no ip arp gratuitous hsrp duplicate

router bgp 65501

 template peer-policy PREFER->PE1
    send-community
    route-map PREFER-PE1 in
    next-hop-self

 template peer-policy ibgp-policy
    next-hop-self

 vrf customer_bronze1
    log-neighbor-changes
    address-family ipv4 unicast
      redistribute direct route-map SERVER-NET-SET-COMM-PATH2
    neighbor 10.3.2.1
      remote-as 109
      address-family ipv4 unicast
        send-community
    neighbor 10.3.4.1
      remote-as 109
      address-family ipv4 unicast
        inherit peer-policy PREFER->PE1 1
    neighbor 10.3.34.3
      remote-as 65501
      address-family ipv4 unicast
        inherit peer-policy ibgp-policy 1
        no send-community

route-map SERVER-NET-SET-COMM-PATH2 permit 10
  match ip address prefix-list SERVER-NET
  set community 32:31

route-map SERVER-NET-SET-COMM-PATH2 permit 20

route-map PREFER-PE1 permit 10
  match community COMM-1
  set local-preference 60000

route-map PREFER-PE1 permit 20

ip prefix-list SERVER-NET seq 5 permit 11.0.0.0/8 le 24

ip community-list standard COMM-1 permit 21:21
ip community-list standard COMM-2 permit 22:22
```

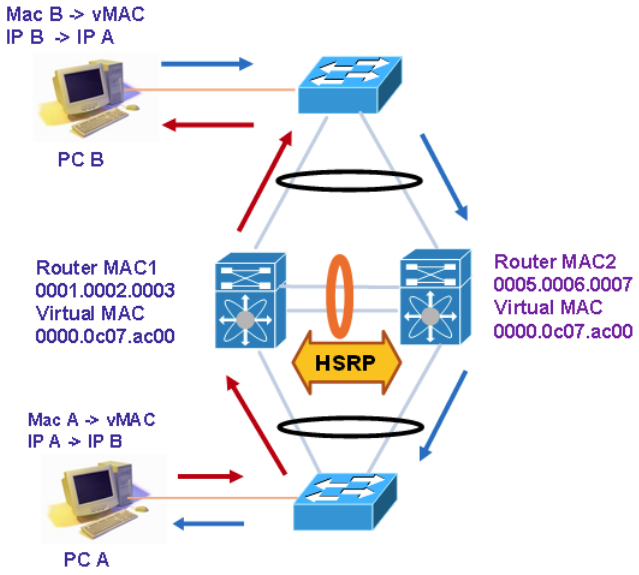
Below are useful **show** commands.

```
show ip bgp vrf XXX summ
show ip bgp vrf XXX neighbors
show ip bgp x.x.x.x vrf XXX
```

HSRP

Hot Standby Router Protocol (HSRP) is a First-Hop Redundancy Protocol (FHRP) that allows a transparent failover of the first-hop IP router. When HSRP is configured on a network segment, a virtual MAC address and a virtual IP address are provided for the HSRP group. HSRP will select one router in the group to be the active router. The active router receives and routes packets destined for the virtual MAC address of the group. In the Nexus 7000, HSRP interoperates with vPCs and behaves slightly different (Figure 4-6). Both active HSRP routers and the standby HSRP router will forward the traffic sent to it. In this solution, we put each Nexus 7000 as active (with high priority) for half of all the HSRP groups. It does not make a difference in the data plane, as the traffic is determined by the load-balance algorithm of the downstream devices (ACE, ASA, or Nexus 5000).

Figure 4-6 HSRP Behavior in a vPC Environment



vPC Peer Gateway and HSRP

Some third-party devices can ignore the HSRP virtual MAC address and instead use the source MAC address of an HSRP router. In a vPC environment, the packets using this source MAC address may be sent across the vPC peer link, causing a potential dropped packet. Configure the vPC peer gateway to enable the HSRP routers to directly handle packets sent to the local vPC peer MAC address and the remote vPC peer MAC address, as well as the HSRP virtual MAC address.

Below is the vPC configuration for the Nexus 7000 router.

```
vpc domain 998
  peer-switch
  role priority 30000
  peer-keepalive destination 192.168.50.21
  delay restore 120
  peer-gateway                    <====================
  auto-recovery
  delay restore interface-vlan 100
  ip arp synchronize
```

HSRP is used in the Nexus 7000 as the gateway of the server, ASA, and ACE. Below is the sample configuration of the server gateway for tenant customer_bronze1.

Agg1

```
interface Vlan801
  no shutdown
  vrf member customer_bronze1
  no ip redirects
  ip address 11.3.1.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 801
    preempt
    priority 150
    ip 11.3.1.1
```

Agg2

```
interface Vlan801
  no shutdown
  vrf member customer_bronze1
  no ip redirects
  ip address 11.3.1.3/24
  no ipv6 redirects
  hsrp version 2
  hsrp 801
    preempt
    priority 120
    ip 11.3.1.1
```

Below is the sample configuration of the ASA outside interface gateway for tenant customer_gold1.

Agg1

```
interface Vlan1301
  no shutdown
  vrf member customer_gold1_pub
  no ip redirects
  ip address 10.1.5.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 1301
    preempt
    priority 150
    ip 10.1.5.1
```

Agg2

```
interface Vlan1301
  no shutdown
  vrf member customer_gold1_pub
  no ip redirects
  ip address 10.1.5.3/24
  no ipv6 redirects
  hsrp version 2
  hsrp 1301
    preempt
    priority 120
    ip 10.1.5.1
```

Below is the sample configuration of the ASA inside interface gateway for tenant customer_gold1.

Agg1

```
interface Vlan1201
  no shutdown
  vrf member customer_gold1_priv
```

```
  no ip redirects
  ip address 10.1.6.2/24
  no ipv6 redirects
  hsrp version 2
  hsrp 1201
    preempt
    priority 150
    ip 10.1.6.1
```

Agg2

```
interface Vlan1201
  no shutdown
  vrf member customer_gold1_priv
  no ip redirects
  ip address 10.1.6.3/24
  no ipv6 redirects
  hsrp version 2
  hsrp 1201
    preempt
    priority 120
    ip 10.1.6.1
```

The ACE is running in one-arm mode, and the interface is in the same subnet as the server, so the ACE gateway is the same as the server gateway.

Below are useful commands.

```
show hsrp brief
show hsrp group
```

Services Layer Routing Configuration

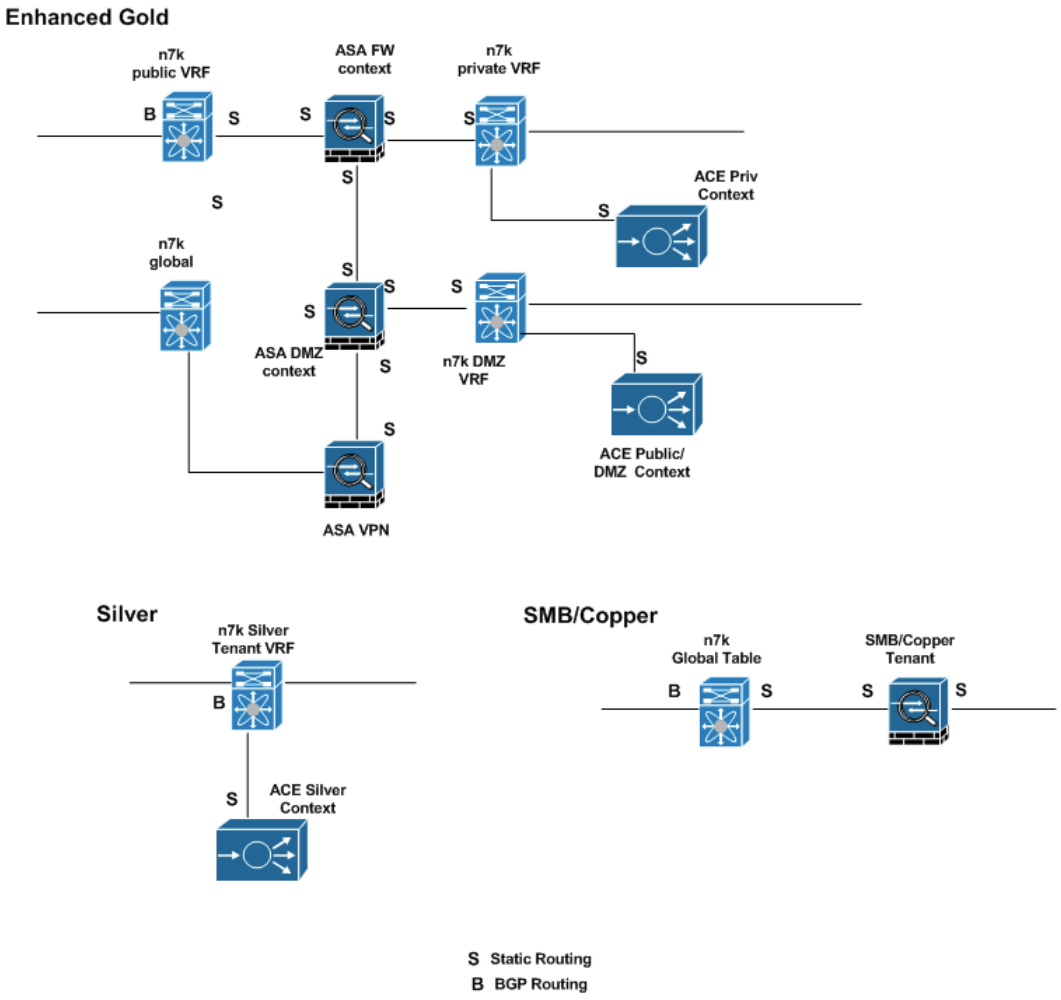
This section provides details on how end-to-end routing is achieved through the Services layer.

All Service Provider client-server Gold tenant traffic received on the Nexus 7000 aggregation routers is forwarded to the tenant ASA firewall context. Also, this traffic might be forwarded to the ACE appliances if the traffic is to be load balanced or is return traffic to load-balanced traffic. Internet client-server traffic will either be sent to the ASA VPN or the ASA firewall, depending on the type of security services associated with that traffic. All Service Provider Silver client-server traffic received on the Nexus 7000 aggregation routers will either be forwarded to the ACE appliance if application services (application load balancing, SSL overload, etc.) need to be provided, or will be forwarded to the Compute layer. Internet services are not provided to Silver tenants. All Service Provider Bronze client-server traffic received will be forwarded to the Compute layer. No Internet/security/application services are provided to these tenants. All Internet SMB/Copper tenants received on the Nexus 7000 routers will be forwarded to the ASA firewall context associated with this tenant.

This section also provides the routing configuration required for end-to-end traffic routing through the ASA firewall, ASA VPN, and ACE appliances. For end-to-end reachability through the Services layer, the appropriate static routes are configured on both the ASA and on the Nexus 7000 tenant VRF instance. Since the VIP and ACE client NAT pools are in the same subnet as the servers, the Nexus 7000 aggregation routes L2 forward packets destined to the ACE appliances.

An overview of Services layer routing is provided in Figure 4-7.

Figure 4-7 Services Layer Overview



This section presents the following topics:

- ASA Firewall Context Routing Configuration, page 4-19
- ASA VPN Routing Configuration, page 4-21
- ACE Routing Configuration, page 4-21

ASA Firewall Context Routing Configuration

For end-to-end routing, the ASA firewall should be able to provide routing for the following traffic:

1. **Tenant Service Provider client traffic that is destined to the Service Provider server private network.** This include Service Provider server and VIP private addresses. To achieve this, static routes are configured on the tenant firewall context to provide reachability to the Service Provider client and server network.
2. **Tenant Service Provider client traffic that is destined to the DMZ network.** This includes the DMZ server and VIP private addresses. To achieve this, Service Provider client addresses are NAT'd before traffic is sent to the DMZ firewall context. The firewall DMZ context has static routes that provide reachability to the Service Provider client NAT network and DMZ server network.

3. **Tenant Internet client traffic that is destined to the DMZ network.** To achieve this, static routes are configured on the DMZ firewall context that provides reachability to the Internet and DMZ server network.
4. **Tenant Internet VPN client (SSL or IPsec) traffic destined to the DMZ and private network.** Static routes are configured on the ASA VPN to provide reachability to the DMZ and private networks. Since VPN client addresses are in the same subnet with the VPN interface in the DMZ context associated with a tenant, static routes providing reachability to VPN client addresses are not required in the DMZ context.

Sample routing configurations for a Gold tenant's private and public ASA contexts are shown below:

Expanded Gold Tenant Private Context Routing Configuration

```
dc02-asa-fw1/customer-gold1-dmz# changeto c customer-gold1
dc02-asa-fw1/customer-gold1# sh route

!snip
dc02-asa-fw1/customer-gold1-dmz# changeto c customer-gold1
dc02-asa-fw1/customer-gold1# sh route

Codes: C - connected, S - static, I - IGRP, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
       i - IS-IS, L1 - IS-IS level-1, L2 - IS-IS level-2, ia - IS-IS inter area
       * - candidate default, U - per-user static route, o - ODR
       P - periodic downloaded static route

Gateway of last resort is 10.1.5.1 to network 0.0.0.0

S    111.0.0.0 255.0.0.0 [1/0] via 10.1.6.1, inside
C    10.1.8.0 255.255.255.0 is directly connected, dmz
C    10.1.6.0 255.255.255.0 is directly connected, inside
C    10.1.5.0 255.255.255.0 is directly connected, outside
S    11.0.0.0 255.0.0.0 [1/0] via 10.1.6.1, inside   # route to private server network
S    11.1.4.0 255.255.255.0 [1/0] via 10.1.8.11, dmz # route to DMZ server network
S    11.255.0.0 255.255.0.0 [1/0] via 10.1.8.11, dmz # route to VPN client networks
C    192.168.50.0 255.255.255.0 is directly connected, mgmt
S*   0.0.0.0 0.0.0.0 [1/0] via 10.1.5.1, outside # default route to private client
networks
S    192.168.0.0 255.255.0.0 [1/0] via 192.168.50.1, mgmt
dc02-asa-fw1/customer-gold1#
dc02-asa-fw1/customer-gold1# changeto c customer-gold1-dmz
dc02-asa-fw1/customer-gold1-dmz# sh route

Codes: C - connected, S - static, I - IGRP, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
       i - IS-IS, L1 - IS-IS level-1, L2 - IS-IS level-2, ia - IS-IS inter area
       * - candidate default, U - per-user static route, o - ODR
       P - periodic downloaded static route

Gateway of last resort is 100.200.1.1 to network 0.0.0.0

S    51.0.0.0 255.0.0.0 [1/0] via 10.1.8.21, inside  # route to private client
networks
C    100.200.1.0 255.255.255.0 is directly connected, internet
C    10.1.8.0 255.255.255.0 is directly connected, inside
C    10.1.7.0 255.255.255.0 is directly connected, dmz
```

```
S    11.1.1.0 255.255.255.0 [1/0] via 10.1.8.21, inside # route to private server
network
S    11.1.4.0 255.255.255.0 [1/0] via 10.1.7.1, dmz # route to dmz server network
C    11.255.1.0 255.255.255.0 is directly connected, vpn # interface to VPN network
C    192.168.50.0 255.255.255.0 is directly connected, mgmt
S*   0.0.0.0 0.0.0.0 [1/0] via 100.200.1.1, internet # default route to internet
S    192.168.0.0 255.255.0.0 [1/0] via 192.168.50.1, mgmt
dc02-asa-fw1/customer-gold1-dmz#
```

ASA VPN Routing Configuration

As mentioned in the previous section, the ASA VPN should have static routes that provide reachability to the DMZ and private network. The next hop for these routes must be associated with a VLAN interface, and this VLAN interface must be configured under the tunnel group used to establish the tunnel. See ASA IPsec VPN Configuration and ASA SSL VPN Configuration for details on how to associate a VLAN ID with a VPN tunnel group. For Internet reachability, the ASA has a default route that points to the Internet SVI interface on the Nexus 7000.

Sample ASA VPN Routing Configuration

```
route internet 0.0.0.0 0.0.0.0 100.200.1.1 1
route dmz1 11.1.0.0 255.255.0.0 11.255.1.251 1
route dmz2 11.1.0.0 255.255.0.0 11.255.2.251 2
route dmz3 11.1.0.0 255.255.0.0 11.255.3.251 3
route dmz4 11.1.0.0 255.255.0.0 11.255.4.251 4
route dmz5 11.1.0.0 255.255.0.0 11.255.5.251 5
route dmz6 11.1.0.0 255.255.0.0 11.255.6.251 6
route dmz7 11.1.0.0 255.255.0.0 11.255.7.251 7
route dmz8 11.1.0.0 255.255.0.0 11.255.8.251 8
route dmz9 11.1.0.0 255.255.0.0 11.255.9.251 9
route dmz10 11.1.0.0 255.255.0.0 11.255.10.251 10
route management 192.168.0.0 255.255.0.0 192.168.50.1 1
dc02-asa5555-1#
```

ACE Routing Configuration

Each ACE tenant context is configured with a default route that points to the HSRP VIP of the web VLAN interface on the Nexus 7000 switches. In this implementation, the web interface is used by the ACE to forward traffic to the Service Provider L3VPN client networks, and the ACE web VIP addresses are in the same subnet with the server and the VLAN interface on the Nexus 7000 aggregation switches. This eliminates the need to have static routes from the Nexus 7000 switches to the ACE, however, if required, separate subnets can be used for VIP addresses, and static routes would be needed on the Nexus 7004 VRF instances for the tenant pointing to the ACE interface address.

Layer 3 Best Practices and Caveats

Best Practices

1. To accelerate L3 convergence, spread the L3 ports on different SoCs on the F2 module. This is due to the fact that on the F2 module, each port is mapped to a VRF instance and then the FIB for that VRF is downloaded. If an SoC has all ports as L2, then during reload and possibly other conditions, when the ports come up, FIB download is delayed until the SVI to VRF mapping is done, and hence FIB download happens after the port comes up and L2 convergence and mapping of VLANs to that

port is complete. In VMDC 2.3 implementation, the L3 ports to the DC PEs and the VPC peer links were spread across five SoCs per module to get the benefit of FIB download immediately on reload. Refer to Cisco Nexus 7000 F2-Series 48-Port 1 and 10 Gigabit Ethernet Module Data Sheet for more information about F2 card and SoCs. Also, see CSCue67104 below.

2. To reduce traffic loss after system reload, delay the time that it takes for VLAN interface and vPCs to come online. By default, VLAN interfaces are brought online 10 seconds after the peer link is up, and vPCs are brought online 30 seconds after the VLAN interfaces are brought up. Based on scale characteristics of this validation, we delay VLAN interfaces and vPCs from coming online by 90 seconds each.
3. The ACE 4710 appliances do not support LACP, and hence their port-channels to the Nexus 7000 switches are static with mode on. We expect to see some traffic loss when the system comes online after a reload. To protect against this loss, carrier delays can be configured on the ACE GigabitEthernet interfaces to prevent this interface from coming online. Using this scheme will introduce a carrier-delay time during a vPC shut/no shut test or similar negative event.
4. Carrier delay can be configured on the ASR 1000 interfaces to the Nexus 7000 aggregation routers to delay the L3 interface from coming up. This ensures that these L3 interfaces are brought up at a time when the Nexus 7000 routers are ready to successfully set up and establish BGP sessions. In this validation, the carrier delay on the ASR 1000 PE was set to the maximum of 60 seconds.
5. By default, the ACE 4710 appliance will renew ARP entries for a configured host every 300 seconds. We increase the ARP rates to 1440 seconds to reduce the possibility of the ACE ARP request being lost as the system comes online after a reload.
6. To get better convergence performance, use BGP policy to divert traffic away from the Nexus 7004 aggregation switch under certain conditions such as VPC peer link fail or secondary shutdown. This is because the FIB programming on the F2 card is slower, leading to additional packet losses of up to 10 seconds in the scale validated, and this can be higher with a high-programmed prefix count. BGP configuration on the ASR 1000 and Nexus 7000 aggregation routers is set up so that the ASR 1000 reroutes traffic to an alternate path if the vPC peer link fails and shuts down the VPC secondary. This eliminates up to 10 seconds of traffic loss that occurs due to the F2 FIB programming delay. If the peer link fails, expect up to 13 seconds of traffic convergence, which is due to up to 8 seconds being required for the VLAN interface to go down, and due to up to 5 seconds being required for the BGP and RIB update on the Nexus 7000 aggregation routers. The causes of this convergence delay in FIB programming is under investigation. See CSCue59878 below. For overall vPC convergence, there are a few enhancements targeted for the next NX-OS software release 6.2.
7. BGP PIC, BGP graceful restart, and other routing optimization should be enabled on the ASR 1000 PE devices for faster convergence. BGP PIC and graceful restart are enabled by default on the Nexus 7000 aggregation routers.

Caveats

1. CSCud23607 was an HSRP programming issue seen if the MAC address table size limits are reached. This is fixed in NX-OS 6.1.3. Prior to NX-OS 6.1.3 , the workaround was to manually flap the affected HSRP interfaces.
2. CSCue59878 was filed to investigate the FIB programming delay after routing convergence during a vPC shut test or similar scenarios. This issue is under investigation. The reason for delay is due to the FIB programming mechanism used for the F2 module. The module has to program TCAM for all 12 SoCs, and as the number of prefixes gets higher, it takes additional time to calculate and program each of the SoCs. The workarounds are to reduce the number of SoCs used, i.e., less number of ports and to reduce the number of prefixes per SoC (by mapping specific VRF instances (ports) to SoCs so that the total prefix is less per SoC). If convergence times need to be quicker, and with a larger number of prefixes, consider using M2 or M1 series modules.

3. CSCue67104 was filled to investigate convergence delays due to packet losses after system reload of the nexus 7000 aggregation router. These losses are seen as FIB losses when the vPC port-channels are brought up and can last 10 or more seconds. This issue was closed as this is expected. On F2 modules, which have an SoC design, each SoC needs to map all of its ports into VRF instances, and then download the FIB. When all of the ports on an SoC are L2 only, the L2 ports need to come up and the SVIs need to be mapped to VRF instances before downloading the FIB for those VRF instances. This takes additional time after the port comes up (see CSCue59878 above, F2 FIB convergence is slow). To work around this issue, have a mix of both L2 and L3 ports on the same SoC. The L3 ports being on the SoC will cause all FIBs for the VRF instances on the L3 port to be downloaded as soon as the module comes up. In VMDC 2.3, all VRF instances used are allowed on the L3 port, so all FIBs will be downloaded to any SoC that has L3 ports. Since there are two L3 uplinks and four L3 peer links for iBGP per box, this provides one L3 port for uplink and two iBGP ports for peer per module. These ports should be spread on three different SoCs. Additionally, we can also spread the vPC peer link ports in different SoCs. Since there are four ports in the vPC peer link, two ports from each module, this covers two more SoCs. This helps with the reload case, as the vPC peer link will come online first and have SVIs mapped to it followed by FIB download, before the actual vPC port-channels come up, however, this will not help in the module restore case, as the vPC peer link port SoCs and FIB download will still be delayed. Additional L3 ports can help, if they are configured on any additional SoCs used. The goal with this workaround is to have all SoC FIBs programmed by the time the vPC port-channels come online.
4. CSCuc51879 is an issue seen during RP failover either due to RPSO or In-Service System Upgrade (ISSU). This is an issue related to traffic loss seen during RPSO or during ISSU on an ASR 1000 PE with a highly scaled up configuration.

The following performance fixes are expected in the 6.2 release of NX-OS. These fixes are expected to help with convergence.

- CSCtn37522: Delay in L2 port-channels going down
- CSCud82316: VPC Convergence optimization
- CSCuc50888: High convergence after F2 module OIR