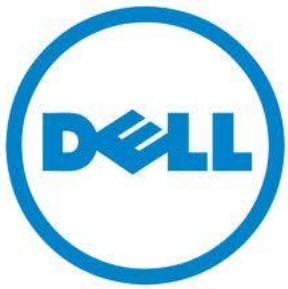
The Case of Routed VLT, Peer Routing, VLT Proxy Gateway and Their Relationship to VRRP

A Detailed Analysis of Layer 3 Forwarding with Dell Virtual Link Trunking

Victor Lama
Dell Network Solutions
Enterprise Campus and Data Center
Northeast Region



Date	Version	Description
10-11-2015	1.0	Initial Release
10-12-2015	1.1	Section 7.0 – changed "it was proven that Dell#**2**" to "it was proven that Dell#**1**" Section 8.0 – changed "Only the relevant configurations for VLT Proxy Gateway will be shown in section **7**.2" to Only the relevant configurations for VLT Proxy Gateway will be shown in section **8**.2"
1-6-2016	1.2	Added Section 9.0 and updated Summary

Table of Contents

Part I - Theory of Operation

Introduction

Dell Networking's Virtual Link Trunking (VLT or vLT) feature is a proprietary Layer 2 (L2) multipathing technology that delivers full bi-sectional bandwidth for today's demanding application workloads. One particular permutation of VLT can deliver L2 adjacencies across physically disparate sites for virtual workload mobility and other server virtualization-based high availability features. In addition, VLT can also be deployed with a dynamic routing protocol for flexibility in deployment scenarios and optimized traffic patterns, as will be discussed in more detail later in this paper.

As is the case with other vendors that introduce proprietary technologies, the terms used to describe certain features and functionality sometimes change as the technology develops. For example, "Routed VLT" and "Peer Routing" basically refer to the same type of functionality. Routed VLT refers to a more general feature-set to describe the fact that VLT and routing, in one form or another, are being deployed simultaneously. The VLT Proxy Gateway feature is an extension of peer routing and also falls under the rubric of Routed VLT. The terms "mVLT" and "eVLT" and even "VLT squared" have also made appearances in Dell Networking documentation over the years. All three refer to the exact same thing: connecting two VLT domains back-to-back, whether in a fully meshed (looped triangle) or partially meshed (square) topology. The term to be used moving forward is mVLT.

This paper is not meant to be a tutorial on the fundamentals of Virtual Link Trunking theory of operation and configuration, although some of that information will be provided. For more detail, the reader is advised to refer to other Dell Networking white papers on basic VLT functionality. A good place to start is the configuration and CLI guides for the particular switch platform that is being deployed. Also, see:

Dell Force10 VLT Technical Guide

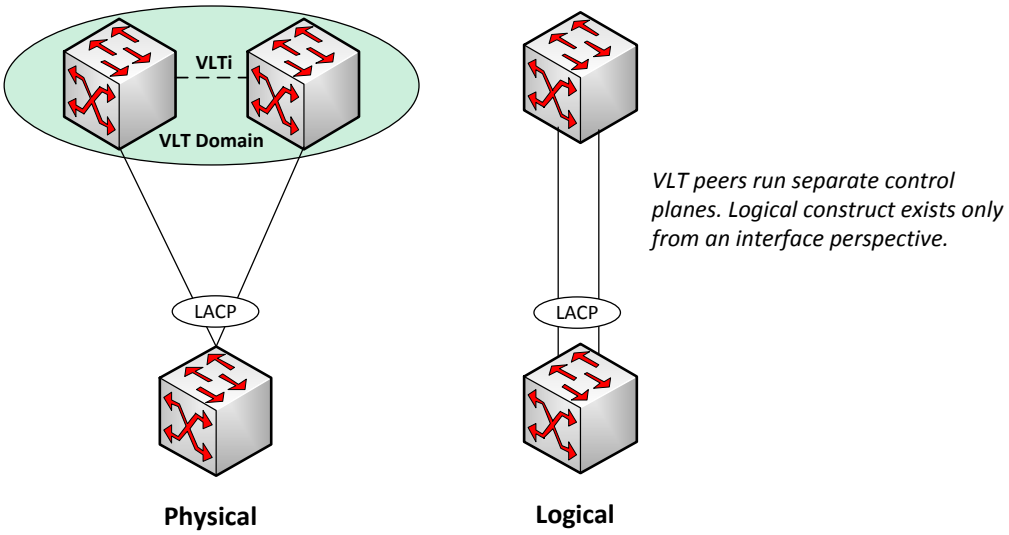
Virtual Link Trunking Reference Architecture

On a high level, a VLT domain is comprised of two Dell Networking S, Z or C-Series switches ("peers" in VLT parlance) that are connected to each other using a link that is known operationally as a VLT Interconnect (VLTi). The VLTi link is comprised of front-panel Ethernet switch ports. No special cables or interfaces are needed. The function of the VLTi link is to exchange protocol state information, as well as MAC address and ARP tables, between both VLT peers. All VLANs are also automatically allowed on the VLTi link without any manual intervention. What results is a pair of switches that present themselves as a single logical entity to a downstream node that is connected to each peer in the VLT domain.

Take note that, with VLT, each peer is still executing its own control plane protocols, unlike stacking. VLT allows the downstream node the ability to leverage a multi-chassis LACP group (LAG) whose members will span both switches in the VLT domain and evade Spanning Tree blocking semantics, sans the single point of failure that stacking presents. See figure 1 below.

Note: The current version of DNOS (Dell Networking Operating System) as of the writing of this paper is DNOS 9.9 (0.0).

Figure 1 – Basic VLT Connectivity



1.0 Peer Routing

Dell Networking's peer routing feature is deployed between two Layer 3 (L3) switches that are part of a single VLT domain. Peer routing is applicable **within** the VLT domain, as opposed to VLT Proxy gateway, which will be discussed later on. Moreover, that VLT domain will be acting as the network's L3/L2 boundary. That means the "northbound" connections to the existing infrastructure from the routed VLT domain will be L3 uplinks that are running dynamic routing protocols, such as OSPF, ISIS, BGP or RIP unicast routing, PIM multicast routing, or static routing.

Peer routing is enabled where one would normally configure a first hop router redundancy protocol, like VRRP (and speaking in general, HSRP or GLBP). VRRP and peer routing are usually deployed exclusively of the other, but that is not a requirement. They can coexist and can be found in "brownfield" environments that already have VRRP deployed to support existing VLANs. Network designers typically deploy a first hop router redundancy protocol on the distribution (aggregation) or core switching layer – wherever the L3/L2 boundary exists. Routed access layers are less common (leaving Network Virtualization Overlays aside for now),

especially in a virtualized data center that typically requires L2 adjacencies across racks of servers.

A routed VLAN is a Layer 2 construct that includes a Layer 3 component: the VLAN's Switched Virtual Interface (SVI). The SVI is the L2 termination point for the VLAN and its default gateway interface. In a legacy network that leverages VRRP, the SVI will be configured with an IP address and VRRP statements, including the router group's virtual IP address (VIP).

What exactly is peer routing? Peer routing can be described as an "agreement" between both boundary switches (VLT peers) **within** a single VLT domain to forward routed traffic on behalf of the other peer, regardless of the configured default gateway of the sending host. Basically, each peer creates a **LOCAL_DA** entry (see section 5.1.1) in its CAM table for the peer's MAC address, thereby taking ownership of all traffic destined for that address, regardless of whether the switch is the configured default gateway for the sending host. This way neither switch will have to forward traffic to the other peer across the VLTi crosslink. In short, it is a technique that offers network-based resiliency and packet forwarding efficiency.

Why is such a mechanism needed between the two VLT peers? In an environment that does not leverage an L2 multipathing mechanism, like VLT, the two boundary switches that are part of a VRRP group will act as master and backup when it comes to L3 forwarding. The master will take ownership of the VIP and the backup will remain in a standby state (the roles may be reversed for a particular VLAN in a per-VLAN Spanning Tree environment). The default gateway, which is considered a "virtual router," must receive the traffic and then forward it to the next L3 hop.

If a frame is received by the VRRP backup router – for example, if it happens to be the STP root for that VLAN – it will forward the frame to the master through the inter-switch link (crosslink). That is why, to minimize crosslink traffic and suboptimal switching paths, it has been a best practice to designate the same switch as the VRRP master and the STP root bridge for each configured VLAN.

VLT changes this virtual router paradigm. Because of the fact that VLT provides for full bi-sectional bandwidth, each VLT peer becomes a candidate to receive user traffic from the downstream source, regardless of an end-station's default gateway address. It's the downstream node's LACP hashing algorithm that will dictate which uplink will be used. Remember, both links will be actively forwarding traffic. In addition, the VLTi link, through which database and state information are exchanged between VLT peers, does not pass user traffic by default, except in the case of a link failure, orphaned ports (those that do not belong to a VLT), and the forwarding of multicast and broadcast traffic. This is part of VLT's loop prevention mechanism.

All this makes it necessary to engineer an efficient traffic forwarding solution to accommodate these realities. That is where peer routing comes into play. It will allow both of the switches in the VLT domain to be active routers to the next L3 hop without consideration for the default gateway's MAC address. No crosslink traffic, no extra hop and no suboptimal traffic patterns.

1.1 Peer Routing vs. VRRP

As explained earlier, peer routing can be deployed as a replacement for VRRP in "greenfield" environments or as a complementary mechanism in "brownfield" environments. Whereas a VRRP group of two switches has a master and a backup forwarder, and the master "owns" the VIP and forwards traffic accordingly – with peer routing, both VLT peer switches forward L3 traffic on behalf of the default gateway. There is no configured VIP in a peer routing design. Instead, each VLT peer's L3 VLAN SVI for any particular VLAN will be assigned an IP address, as usual, and the default gateway on the host can be set to either one of them. Again, it does not matter because both VLT peers will forward the packet.

Up until now, VRRP's behavior in a legacy (blocking) environment has been covered. The fact is that VRRP will behave differently in a VLT scenario. When VRRP is configured on switches that are part of a boundary VLT domain, an enhancement to VRRP will allow **both** members to become active forwarders, regardless of who the master is or if the end host is directing its packets to a VRRP peer's IP Address, instead of the VIP, as some well-known storage vendor equipment has been known to do.

So, what then is the benefit of deploying peer routing over VRRP? Perhaps the most important reason is scalability. VRRP can only scale to 255 VLANS, where peer routing can support as many VLANs as are allowed on the switch. Peer routing is enabled globally for all VLAN gateway interfaces with one command line, as opposed to VRRP, which requires multiple configuration lines for every routed VLAN interface on the switch. That brings us to the next point: VRRP is more CPU-intensive, since an instance has to be run for every participating VLAN interface. Lastly, peer routing is a building block for another Dell Networking innovation known as VLT Proxy Gateway, whereas VRRP is not an option.

2.0 Routed VLT

So, what then is Routed VLT? In short, it refers to the ability to run a dynamic routing protocol within a single VLT domain or between VLT domains in an mVLT topology. With regard to the former, it allows routing adjacencies to be formed across the VLTi link, which is a Dell Networking differentiator. In the case of the latter, routing adjacencies are formed across the port-channel that connects the two VLT domains. Routed VLT may be deployed in conjunction with peer routing, but it is not necessary; a design that incorporates VLT, a routing protocol and VRRP is very prevalent.

There is a particularly interesting Routed VLT deployment scenario that can be found in Section 3.5 of the VLT Reference Architecture Guide v2.0 published by P. Narayana Swamy of the Dell Networking CTO's office. It consists of an architecture in which two separate access layer (leaf) VLT domains (four switches altogether) are "dual-homed" to a pair of core (spine) switches that are also in a VLT domain. Moreover, a routing protocol, OSPF, is deployed on that broadcast network with an OSPF network type of Broadcast.

In figure 2 below, the spine switches may be the DR and BDR, and among those three VLT domains, L2 and L3 VLANs can be "interspersed," as the CTO document describes it. For example, in VLT domain 1, VLAN 10 may be an L3 VLAN (the gateway for VLAN 10) and VLANs 20 and 30 may be only L2. Whereas, in VLT domain 2, VLAN 20 may be an L3 VLAN (the gateway for VLAN 20), while VLANs 10 and 30 may be configured as only L2...and so on. The main reason for deploying such a design is to scale very large virtualized server environments and distribute the burden of maintaining a large number of ARP entries across different VLT domains.

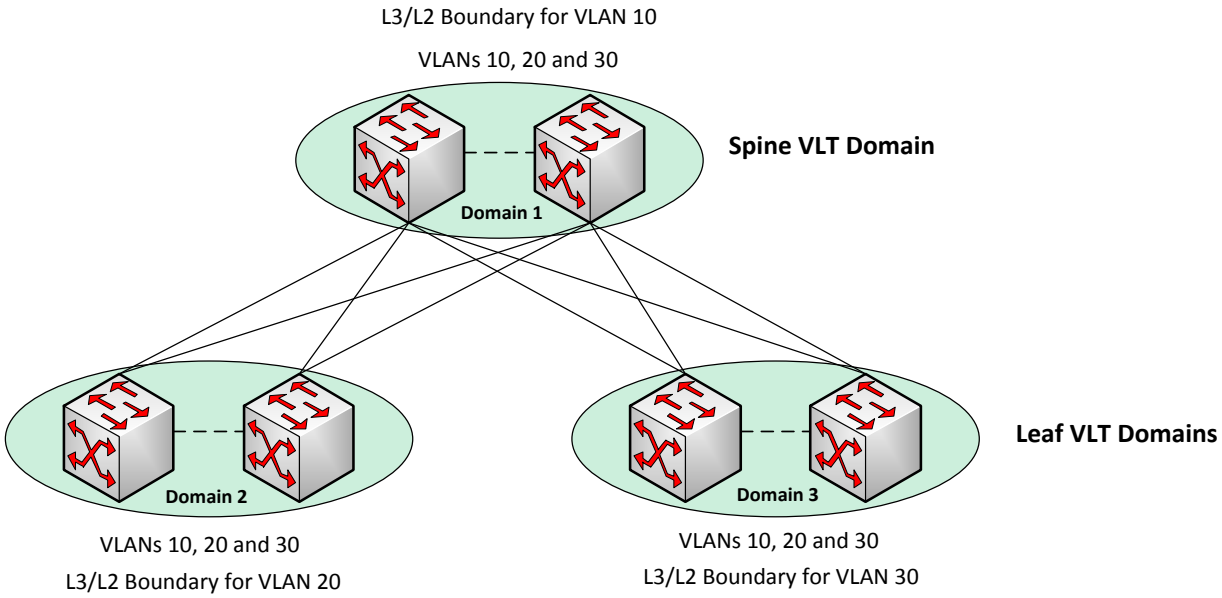


Figure 2 – Routed VLT Domains with Interspersed L3/L2 boundaries

Another example of when it may become necessary – or at least desirable – to deploy a routing protocol over a VLT-enabled port-channel is the case in which a router/L3 switch at a remote site has two "dark fiber" Metro Ethernet connections that are terminated on each of the VLT peers at a central location. This particular scenario is seen quite often in smaller environments, such as K-12 school districts and small-to-midsize university campuses where data center edge routers and a service provider WAN connectivity service are not in place.

In such cases, the remote L3 device and the L3-capable VLT domain at the central site will leverage a transit VLAN over a VLT port-channel to achieve routing protocol adjacencies and exchange routing information.

3.0 VLT Proxy Gateway

All this leads up to the VLT Proxy Gateway feature. Similar to the manner in which peer routing allows both peers **within** a single VLT domain to function as active forwarders on behalf of the other, VLT Proxy Gateway extends that functionality **across** two separate VLT domains in what is known as an mVLT group. An mVLT group consists of two separate VLT domains connected via a port channel, sometimes referred to here as crosslinks. Routing may or may not be deployed with mVLT. But when it is, the VLT Proxy Gateway feature allows one VLT group to proxy-route on behalf of the other.

Peer routing and VLT proxy Gateway together offer a combined solution as a collapsed core and an L2 Data Center Interconnect (DCI) between two sites. It provides a particularly elegant solution for virtual workloads that migrate between sites without the usual traffic flow pattern known as "tromboning," which refers to the situation in which inter-VLAN traffic flows back and forth between sites with "stretched" VLANs.

Whereas with peer routing, each peer within the VLT domain leverages the VLTi link to establish that peer routing agreement, with VLT Proxy Gateway, LLDP TLVs are exchanged across the port channel that connects the two VLT domains. This is the case if dynamic VLT Proxy Gateway is deployed. Otherwise, it can also be deployed in static fashion, thereby precluding the need to exchange LLDP TLVs.

In the scenario depicted in Figure 3 below, a routed mVLT domain is extended across two data centers with an L2 adjacency between them. VLT domain 1 resides in DC1 and VLT domain 2 in DC2. They are connected to each other via two Metro-Ethernet Data Center Interconnect (DCI) "crosslinks" that comprise a port channel in a square topology.

If switches A, B, C and D all have SVI interfaces for VLAN 10 configured on them, with peer routing configured **within** each VLT domain and VLT Proxy Gateway configured **between** VLT domains, all switches will be able to route packets (proxy) on behalf of the other, regardless of which switch is the default gateway for any particular workload on VLAN 10, **and** regardless of where that workload exists, whether in DC1 or DC2. Concomitantly, the default gateway for that workload on VLAN 10 will always be **local** to the location of the workload, hence mitigating traffic "tromboning" across data centers over the DCI links.

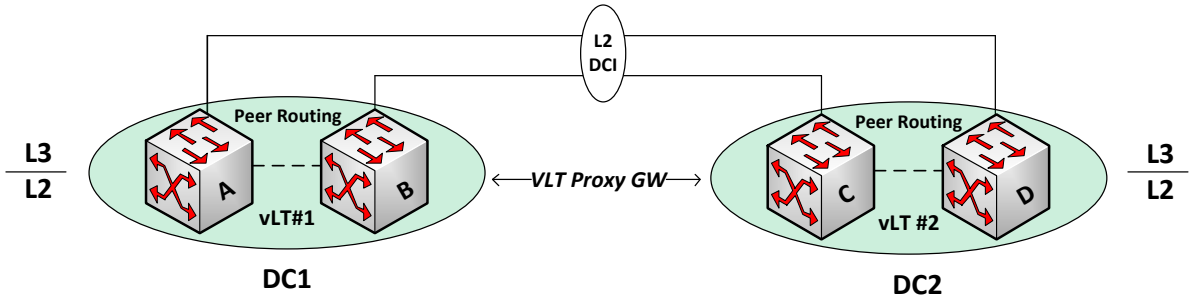


Figure 3 – Routed mVLT Connectivity – Square Topology

Remember, in this scenario, VRRP is not configured anywhere and a routing protocol, such as OSPF, will be configured on the VLT domains since they are the L3/L2 boundaries and will have to establish adjacencies with upstream routers, as well as between VLT domains to advertise local networks that are not spanned across the DCI link. This can be considered a Routed VLT scenario. Technically, it is a Routed **mVLT** scenario with VLT Proxy Gateway enabled between the two VLT domains that make up the mVLT domain.

Parenthetically, the question sometimes arises as to whether one can configure VRRP within each VLT domain (instead of peer routing) and leverage VLT Proxy Gateway between VLT domains to effectively extend the VRRP group across both VLT domains in the mVLT. The answer is that VRRP and VLT Proxy Gateway are mutually exclusive.

4.0 Spanning Tree in a Routed VLT Design

There is a common misconception that a network design that incorporates an L2 multipathing technology, like VLT, does not need to have any form of Spanning Tree Protocol (STP) deployed. **This is a fallacy and constitutes a dangerous practice.** While STP will not block any member of a port channel (conventional or multichassis), there is always the possibility that the VLT domain can experience a failure that results in the inability to synchronize state between members. Furthermore, STP prevents the formation of a bridging loop as a result of an errant connection to another switch outside the VLT domain.

STP should always be configured, and it should be done prior to configuring VLT to prevent any loop from forming as VLT converges. This best design practice is applicable in a switched or routed VLT scenario. Furthermore, if RSTP (802.1w) is deployed, then it is another best practice to ensure that the root bridge for the Common Spanning Tree (CST) is also the VLT primary. This can be achieved with Force10-PVST+ deployed as well, but that of course mitigates its load sharing capability since the root ridge for all VLANs will be the same switch.

As for deploying STP in the VLT Proxy Gateway scenario (or even in a non-routed mVLT scenarios across disparate data centers), the best practice is to apply BPDU filtering (or simply disable STP) on the DCI port channel to keep the spanning tree domains in each data center

separate from each other. This will reduce convergence and operational complexity and simplify troubleshooting by isolating the L2 control plane and its fault domain to each data center. Accordingly, a change in a particular data center will not cause transient connectivity problems or superfluous flooding in another data center. The segmentation will also make configuration of the topology of the various data centers easier, as it will be computed relative to a local root bridge.

As of DNOS version 9.7, Force10-PVST+ is compatible with VLT. For an in-depth understanding of Force10-PVST+ protocol semantics and its interactions with Cisco's rapid-PVST+, please navigate to the following link:

Dell Force10 PVST+ Protocol Semantics - A Detailed Analysis

PART II - Examples

5.0 Peer Routing Lab Example

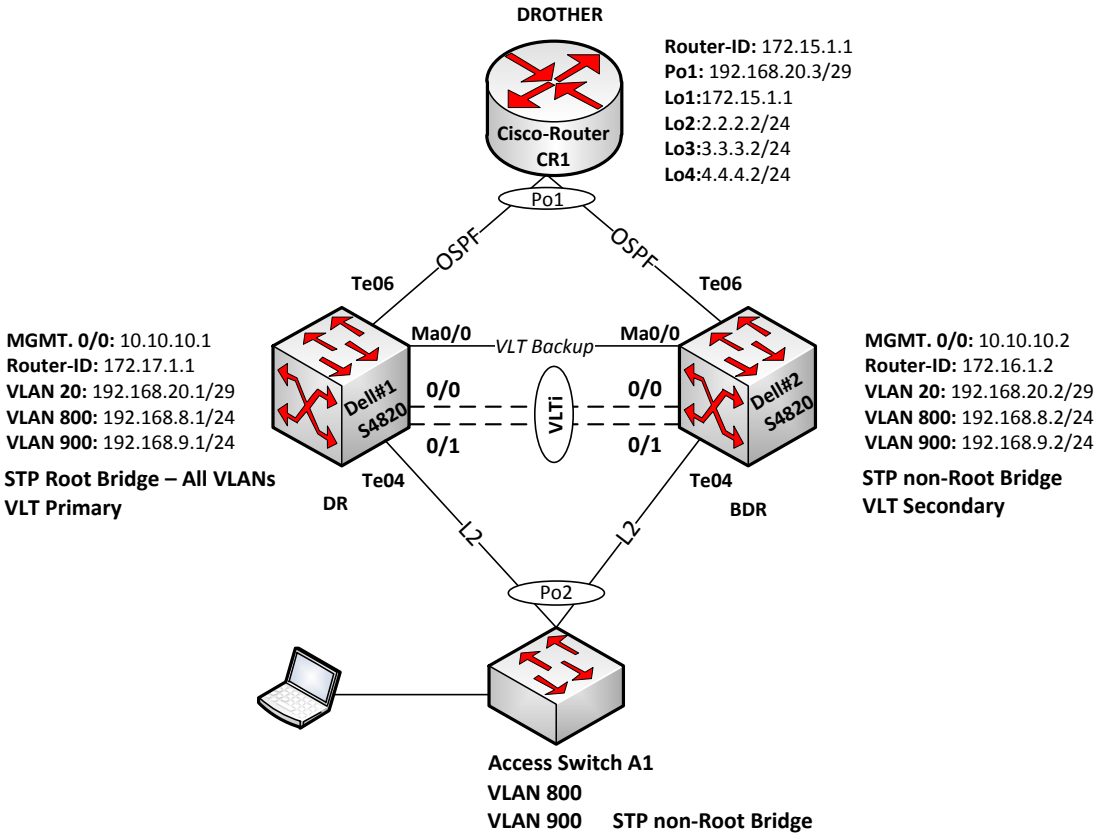


Figure 4 – Routed VLT Domain Baseline Topology

Peer routing is configured simply by entering the ***peer-routing*** command under the VLT domain configuration sub-section. Doing this precludes the need to deploy VRRP.

What follows below is a holistic look at the foundational topology, associated configurations and verification data regarding a baseline VLT deployment with peer routing added to it.

5.1 Dell#1 Switch Configurations and Verification

Dell#1#sh run | find protocol
protocol spanning-tree pvst
 no disable
 vlan 1,20,800,900 bridge-priority 0
(Take note that deploying VLT does NOT preclude the need to deploy the spanning tree protocol (STP). STP will be acting as a loop prevention mechanism in the event of a VLT failure or an errant connection that creates a physical bridging loop.
!
This bridge is the root for all VLANs. It's a best practice to make the root bridge the same as the VLT primary. While this practice mitigates the value of running PVST+ for uplink load sharing purposes, it does make failover more deterministic. Consider running RSTP, if feasible.)
!
!
Dell#1#sh vlan | find NUM

	NUM	Status	Description	Q Ports
*	1	Active		U Po10 (Te 0/0-1)
				U Te 0/4,47
	20	Active	OSPF PEERING VLAN	U Po1 (Te 0/6)
				V Po10 (Te 0/0-1)
	800	Active	Client-VLAN	V Po10 (Te 0/0-1)
	900	Active	Client-VLAN-2	V Po10 (Te 0/0-1)

!
!
Dell#1#sh run int ma0/0
interface ManagementEthernet 0/0
 description Used_for_VLT_Keepalive
 ip address 10.10.10.1/24
 no shutdown
 (The management interfaces are part of a default VRF and are isolated from the switch's data plane.)
!
Dell#1#sh run int te0/0
interface TenGigabitEthernet 0/0
 description VLTi LINK
 no ip address
 no shutdown
(VLTi Physical link)
!
Dell#1#sh run int te0/1
interface TenGigabitEthernet 0/1
 description VLTi LINK
 no ip address

```
 no shutdown
 (VLTi Physical link)
!
Dell#1#sh run int po10
interface Port-channel 10
 description VLTi Port-Channel
 no ip address
 channel-member TenGigabitEthernet 0/0-1
(The best practice is to configure the VLTi links for static port-channeling)
 no shutdown
!
Dell#1#sh run int te0/4
interface TenGigabitEthernet 0/4
 description To_Access_Switch_A1_fa0/13
 no ip address
 port-channel-protocol LACP
  port-channel 2 mode active
 no shutdown
!
Dell#1#sh run int te0/6
interface TenGigabitEthernet 0/6
 description To_CR1_fa0/13
 no ip address
 port-channel-protocol LACP
  port-channel 1 mode active
 no shutdown
!
Dell#1#sh run int po1
interface Port-channel 1
 description port-channel_to_CR1
 no ip address
 switchport
 vlt-peer-lag port-channel 1
 no shutdown
!
Dell#1#sh run int po2
interface Port-channel 2
 description port-channel_to_access_switch_A1
 no ip address
 portmode hybrid
 switchport
 vlt-peer-lag port-channel 2
 no shutdown
!
```

Dell#1#sh run int vlan 20
interface Vlan 20
 description OSPF PEERING VLAN
 ip address 192.168.20.1/29
untagged Port-channel 1
 no shutdown
!
Dell#1#sh run int vlan 800
interface Vlan 800
 description Client-VLAN
 ip address 192.168.8.1/24
 tagged Port-channel 2
 no shutdown
!
Dell#1#sh run | find vlt
vlt domain 1
(Each VLT domain is assigned a number)
 peer-link port-channel 10
(This is the VLTi link, which consists of physical ports Te0/0 and 0/1 [Po10] on each switch)
 back-up destination 10.10.10.2
(This is an out-of-band VLT integrity mechanism (heartbeat). The IP address is of the Management 0/0 interface on the VLT peer)
 primary-priority 4096
(One switch is the VLT primary and the other is the secondary. The switch with the quantitatively lower priority number becomes the VLT primary. Primary and secondary switch considerations are made when creating a design with deterministic failover. During a VLTi link failure, the secondary will shut down its VLT ports to prevent a loop.)
 system-mac MAC address 90:b1:1c:f4:01:01
(This is a common MAC address that is contrived and shared by both VLT peers as part of the switch virtualization mechanism)
 unit-id 0
(Each switch in the VLT domain will have a different unit-id)
peer routing
(Enables peer routing)
!
!
Dell#1#sh vlt brief
(Verification of VLT foundational configurations and operability)
 VLT Domain Brief

Domain ID:	1
Role:	Primary
Role Priority:	4096

ICL Link Status: Up

(This refers to the VLTi link, which consists of physical ports Te0/0 and 0/1 - Po10. The VLTi link will automatically allow any VLAN that is configured on the switch. They are known as Spanned VLANs)

HeartBeat Status: Up

(This refers to the keepalive that traverses the out-of-band management interfaces. If the VLTi link goes
down, the VLT process on each switch uses the heartbeat to determine whether the peer switch is still available to forward traffic. If so, what results is known as a split brain and VLT will take certain measures to ensure a loop-free topology. For details on failover, please refer to the Dell Networking configuration guide for the switch platform in use.)

VLT Peer Status: Up

(This is dependent on the integrity of the VLTi link, which consists of ports physical Te0/0 and 0/1 - Po10)

Local Unit Id: 0

Version: 6(3)

Local System MAC address: 90:b1:1c:f4:2c:bb

*(This is one of the system MAC addresses for the **local** VLT peer switch. It is different than the MAC address that is assigned to all the interfaces on the switch. The interface MAC address is the one that gets entered into the CAM table and leveraged by peer routing.)*

Remote System MAC address: 90:b1:1c:f4:29:f1

*(This is one of the system MAC addresses for the **remote** VLT peer switch. It is different than the MAC address that is assigned to all the interfaces on the switch. The interface MAC address is the one that gets entered into the CAM table and leveraged by peer routing.)*

Configured System MAC address: 90:b1:1c:f4:01:01

(A common MAC address is configured and shared by both VLT peers as part of the switch virtualization mechanism)

Remote system version: 6(3)

Delay-Restore timer: 90 seconds

(This is the time that the VLT process will wait before fully enabling itself after a VLTi link failure and its restoration. The delay is to allow the information tables and protocol state information on each switch to reconverge.)

Peer routing : Enabled

(self-explanatory)

Peer routing-Timeout timer: 0 seconds

(DNOS offers the option of allowing the peer routing function to operate for only a defined period of time after a VLT peer failure. A "0" means that there is no time limit; the surviving VLT peer will continue L3 forwarding indefinitely)

Multicast peer routing timeout: 150 seconds

!

Dell#1#sh vlt backup-link

(Verify that the heartbeat mechanism is operational)

VLT Backup Link

Destination: 10.10.10.2
Peer HeartBeat status: Up
Destination VRF: default

(The heartbeat mechanism uses the management interface and is part of the management VRF, which is enabled by default)

HeartBeat Timer Interval: 1
HeartBeat Timeout: 3
UDP Port: 34998
HeartBeat Messages Sent: 4
HeartBeat Messages Received: 5
!

Dell#1#sh vlt detail

(The fact that the foundational configurations for the VLTi and the heartbeat are correct does not necessarily mean that an active port-channel has been successfully configured. Use this command to verify that the multi-chassis LAG is up and that the correct VLANs are allowed)

Local LAG Id	Peer LAG Id	Local Status	Peer Status	Active VLANs
1	1	UP	UP	20
2	2	UP	UP	1, 800, 900

!

Dell#1#sh run | find router

router ospf 1

(Configuring a dynamic routing protocol is NOT necessary to enable peer routing itself, but one is typically used in conjunction with it since both are enabled on the L3/L2 boundary routers. For peer routing to function properly – that is, for each VLT peer to be able to proxy on behalf of the other when performing L3 forwarding – the IP routing tables on both peers must be fully converged.

!

Furthermore, take note that there are several ways to create routed connections between the VLT domain and CR1. Besides what is displayed in this example, one can take two other approaches: 1) Leveraging 2 x /30 point-to-point connections between CR1 and each VLT peer and 2) Leveraging a layer-2 port channel and an SVI for the OSPF peering VLAN on A1, with that VLAN allowed on the port channel.

!

The choice to take the approach shown in this example is to showcase a Routed VLT configuration that also mitigates some of the layer-2 complexities, such as spanning-tree convergence, by deploying a layer-3 port channel, instead of an SVI, on A1.

router-id 172.17.1.1

(The router-id is the same as the highest loopback interface IP address)

network 192.168.9.0/24 area 0
network 192.168.8.0/24 area 0
network 172.17.1.0/24 area 0
network 192.168.20.0/29 area 0
passive-interface default

(Prevents all interfaces from establishing an OSPF neighborship)

no passive-interface vlan 20

(Excludes the interface for VLAN 20, the OSPF peering VLAN, from the "passive interface" construct, thereby allowing it to establish adjacencies where appropriate. VLAN 20's interface will send LSAs for those interfaces covered by the "network" statements)

!

Dell#1#show ip ospf neighbor

(Dell#1 is the DR)

Neighbor ID	Pri	State	Dead Time	Address	Interface	Area
172.16.1.2	1	FULL/BDR	00:00:31	192.168.20.2	Vl 20	0
172.15.1.1	1	FULL/DROTHER	00:00:39	192.168.20.3	Vl 20	0

!

Dell#1#show ip route ospf

	Destination	Gateway	Dist/Metric	Last Change
	---	---	---	---
O	2.2.2.2/24	via 192.168.20.3, Vl 20	110/2	02:13:50
O	3.3.3.2/24	via 192.168.20.3, Vl 20	110/2	02:13:50
O	4.4.4.2/24	via 192.168.20.3, Vl 20	110/2	02:13:50

(Take note of the routes to the loopback addresses on the CR1 router that have been learned through OSPF)

O	172.15.1.1/32	via 192.168.20.3, Vl 20	110/2	02:13:50
O	172.16.1.2/32	via 192.168.20.2, Vl 20	110/1	02:13:50

!

Dell#1#show interfaces | grep Hardware

Hardware is DellEth, address is 90:b1:1c:f4:2c:bd
Hardware is DellEth, address is 90:b1:1c:f4:2c:bd
Hardware is DellEth, address is 90:b1:1c:f4:2c:bd
Hardware is DellEth, address is 90:b1:1c:f4:2c:bd

*(Output truncated. Every interface, physical and virtual, has the same system interface MAC address shown above. **It is this address that peer routing leverages as will be seen in the next section.**)*

5.1.1 Verifying the Contents of Dell#1's System CAM

Another way to verify that peer routing has populated the CAM table with the correct information is to use the following command from privileged exec mode:

show cam mac stack-unit 0 port-set 0

The output of this command offers a glimpse into the contents of the L2 CAM (Content Addressable Memory), which includes static, dynamic and LOCAL_DA MAC address information. This is a hardware-based lookup. This differs from the output of the "show MAC address table"

command, which is a data structure built in software and will only list a subset of the L2 CAM's contents, including static and dynamically learned MAC addresses.

The *stack-unit* keyword refers to the number designation of the switch with regard to its position in a stack. If it is a standalone unit (no stacking), it is designated as unit 0. If stacked, the switches will be numbered from 0 to 5.

The *port-set* keyword refers to all the ports that are part of the same Broadcom Trident chipset. For example, a Dell S4820's switch architecture is a single switch-on-a-chip, and therefore all sixty-four 10G ports are part of port-set 0, which is the first (and only) port set in this switch. And if the S4820 is one of a set of stacked switches, the stack-unit number will change (between 0 and 5), but the port-set will always be 0. Since each port set maps to one Broadcom Trident chip, in a switch that has multiple Trident chips deployed as part of its architecture, such as the Dell Z9500, multiple port sets would exist.

Dell#1#sh cam mac stack-unit 0 port-set 0
!

VlanId	Mac Address	Region	Interface
20	90:b1:1c:f4:29:f3	STATIC	Po 10
20	**00:0d:bc:6e:93:00**	DYNAMIC	Po 1
20	ff:ff:ff:ff:ff:ff	STATIC	00001
900	90:b1:1c:f4:29:f3	STATIC	Po 10
900	ff:ff:ff:ff:ff:ff	STATIC	00001
800	90:b1:1c:f4:29:f3	STATIC	Po 10
800	ff:ff:ff:ff:ff:ff	STATIC	00001
0	ff:ff:ff:ff:ff:ff	STATIC	00001
0	**90:b1:1c:f4:2c:bd**	LOCAL_DA	00001
0	**90:b1:1c:f4:29:f3**	LOCAL_DA	00001

To make sense out of the returned data, recall from the output of the *'show interfaces'* command above that MAC address **90:b1:1c:f4:2c:bd** is the system-wide MAC address for all the physical and virtual interfaces on switch Dell#1 and, as will be verified in section 5.2, MAC address **90:b1:1c:f4:29:f3** is the system-wide MAC address for all the physical and virtual interfaces on switch Dell#2. Furthermore, MAC address **00:0d:bc:6e:93:00** is the system MAC address for all the interfaces on the CR1 router and has been dynamically learned.

The last two entries are the most interesting with regard to peer routing. They are indicative that it is enabled on the switch and that the necessary forwarding information has been successfully imported into the L2 forwarding database. Specifically, it is the L2 MAC address of the SVI interfaces of the peer switch that must be present in the forwarding table for peer routing to work. Not only must it be present, but it must be considered a LOCAL_DA.

5.1.2 Taking a Detailed Look at the Contents of the L2 CAM.

The region indicates the type of entry. Regions are important because they determine the priority of the entries in the L2 CAM. For example, all entries in the LOCAL_DA region are searched before the entries in the STATIC region. Some MAC addresses may show up in more than one region. Entries in the same region are always grouped together in the CAM, and the order of the regions is always the same. The regions are as follows:

1. LOCAL_DA – This is a local destination address. The MAC address in the entry is the address associated with an IP address that is local to the switch. Loopback addresses do not have MAC addresses associated with them, so they do not require LOCAL_DA entries. **Any packet that matches a LOCAL_DA entry will then be forwarded to the layer-three CAM for final resolution. In other words, the packet will be processed and forwarded, as opposed to being sent to the peer switch.** In the case of a routed packet, the L3 CAM may forward the packet directly to an outgoing interface, while if the packet is destined for the local switch, the L3 CAM will forward it to the appropriate CPU (interface 00001 in the command output)

2. STATIC – This region contains entries that are statically defined in the MAC address table and special entries installed by particular protocols. Entries matching the broadcast address are installed in this region as well as multicast entries that are associated with protocols such as OSPF, RIP, or VRRP. If those protocols are not configured, the entries are not present.

3. DYNAMIC – This region matches any entries that are learned by the switch. This region should have the dynamic entries that are found in the MAC address table.

Going back to basics for a moment, when an end-station needs to communicate with a host that resides outside its local subnet, it forwards the frame to its default gateway. The source MAC address of that frame will be of the end-station, and, in a peer routing scenario, the destination MAC address will be of the SVI interface for that VLAN – that MAC address will exist in the L2 CAM as a LOCAL_DA.

5.2 Dell#2 Switch Configurations and Verification

Dell#2#sh run | find protocol
protocol spanning-tree pvst
 no disable
 vlan 1,20,800,900 bridge-priority 32768
*(Take note that this VLT peer is the **secondary** root bridge for all VLANs in keeping with the best practice that the STP root bridge for all VLANs should be the same as the VLT primary and the secondary root bridge should be the VLT secondary. Although, there is no such thing as a*

secondary root bridge in any spanning tree standard, what is being referred to here is the bridge that will be elected the root in the event of a failure due to its next-to-lowest bridge ID.)

!

Dell#2#sh vlan | find NUM

	NUM	Status	Description	Q Ports
*	1	Active		U Po10(Te 0/0-1)
	20	Active	OSPF PEERING VLAN	U Po1(Te 0/6)
				V Po10(Te 0/0-1)
	800	Active	CLIENT-VLAN-1	V Po10(Te 0/0-1)
				U Te 0/4
	900	Active	CLIENT-VLAN-2	V Po10(Te 0/0-1)

!

Dell#2#sh run int ma0/0

```
interface ManagementEthernet 0/0
 description Used_for_VLT_Keepalive
 ip address 10.10.10.2/24
 no shutdown
```

!

Dell#2#sh run int te0/0!

```
interface TenGigabitEthernet 0/0
 description VLTi LINK
 no ip address
 no shutdown
```

!

Dell#2#sh run int te0/1

```
interface TenGigabitEthernet 0/1
 description VLTi LINK
 no ip address
 no shutdown
```

!

Dell#2#sh run int po10

```
interface Port-channel 10
 description VLTi Link to Dell#2
 no ip address
 channel-member TenGigabitEthernet 0/0-1
 no shutdown
```

!

Dell#2#sh run int te0/4

```
interface TenGigabitEthernet 0/4
 description To_Access_Switch_A1_fa0/14
 no ip address
 port-channel-protocol LACP
  port-channel 2 mode active
```

```
 no shutdown
!
Dell#2#sh run int te0/6
interface TenGigabitEthernet 0/6
 description To_CR1__fa0/14
 no ip address
 port-channel-protocol LACP
  port-channel 1 mode active
 no shutdown
!
Dell#2#sh run int po1
interface Port-channel 1
 description port-channel_to_CR1
 no ip address
 switchport
 vlt-peer-lag port-channel 1
 no shutdown
!
Dell#2#sh run int po2
interface Port-channel 2
 description port-channel_to_access_switch
 no ip address
 portmode hybrid
 switchport
 vlt-peer-lag port-channel 2
 no shutdown
!
Dell#2#sh run int vlan 20
interface Vlan 20
 description OSPF PEERING VLAN
 ip address 192.168.20.2/29
 untagged Port-channel 1
 no shutdown
!
Dell#2#sh run int vlan 800
interface Vlan 800
 description CLIENT-VLAN-1
 ip address 192.168.8.2/24
 tagged Port-channel 2
 no shutdown
!
```

Dell#2#sh run | find vlt
vlt domain 1
 peer-link port-channel 10
 back-up destination 10.10.10.1
 primary-priority 55000
(This is the secondary VLT peer)
 system-mac MAC address 90:b1:1c:f4:01:01
 unit-id 1
peer routing
(Peer routing must be enabled on both VLT peers)
!
Dell#2#sh vlt brief
 VLT Domain Brief

Domain ID:	1
Role:	Secondary
Role Priority:	55000
ICL Link Status:	**Up**
HeartBeat Status:	**Up**
VLT Peer Status:	**Up**
Local Unit Id:	1
Version:	6(3)
Local System MAC address:	90:b1:1c:f4:29:f1
Remote System MAC address:	90:b1:1c:f4:2c:bb
Configured System MAC address:	90:b1:1c:f4:01:01
Remote system version:	6(3)
Delay-Restore timer:	90 seconds
Peer routing :	**Enabled**
Peer routing-Timeout timer:	**0 seconds**
Multicast peer routing timeout:	150 seconds

!
Dell#2#sh vlt backup-link
 VLT Backup Link

Destination:	10.10.10.1
Peer HeartBeat status:	**Up**
Destination VRF:	default
HeartBeat Timer Interval:	1
HeartBeat Timeout:	3
UDP Port:	34998
HeartBeat Messages Sent:	8
HeartBeat Messages Received:	8

!

Dell#2#show vlt detail

Local LAG Id	Peer LAG Id	Local Status	Peer Status	Active VLANs
1	1	UP	UP	20
2	2	UP	UP	1, 800, 900

!

Dell#2#sh run | find router

```
router ospf 1
 router-id 172.16.1.2
network 192.168.8.0/24 area 0
network 192.168.9.0/24 area 0
network 172.16.1.0/24 area 0
 network 192.168.20.0/29 area 0
 passive-interface default
 no passive-interface Vlan 20
```

!

Dell#2#show ip ospf neighbor

(Dell#2 is the BDR)

Neighbor ID	Pri	State	Dead Time	Address	Interface	Area
172.17.1.1	1	FULL/DR	00:00:31	192.168.20.1	Vl 20	0
172.15.1.1	1	FULL/DROTHER	00:00:33	192.168.20.3	Vl 20	0

!

Dell#2#show ip route ospf

	Destination	Gateway	Dist/Metric	Last Change
O	2.2.2.2/24	via 192.168.20.3, Vl 20	110/2	02:14:25
O	3.3.3.2/24	via 192.168.20.3, Vl 20	110/2	02:14:25
O	4.4.4.2/24	via 192.168.20.3, Vl 20	110/2	02:14:25
O	172.15.1.1/32	via 192.168.20.3, Vl 20	110/2	02:14:25
O	172.17.1.1/32	via 192.168.20.1, Vl 20	110/1	02:14:25

Dell#2#

(Dell#2 also learns of the routes to the loopback addresses on CR1 via OSPF.)

!

Dell#2#show interfaces | grep Hardware

```
Hardware is DellEth, address is 90:b1:1c:f4:29:f3
Hardware is DellEth, address is 90:b1:1c:f4:29:f3
Hardware is DellEth, address is 90:b1:1c:f4:29:f3
Hardware is DellEth, address is 90:b1:1c:f4:29:f3
```

(Output truncated. Every interface, physical and virtual, has the same system interface MAC address shown above.)

5.2.1 Verifying the Contents of Dell#2's System CAM

Dell#2#show cam mac stack-unit 0 port-set 0

VlanId	Mac Address	Region	Interface
900	ff:ff:ff:ff:ff:ff	STATIC	00001
800	**90:b1:1c:f4:2c:bd**	**STATIC**	**Po 10**
20	ff:ff:ff:ff:ff:ff	STATIC	00001
0	ff:ff:ff:ff:ff:ff	STATIC	00001
800	ff:ff:ff:ff:ff:ff	STATIC	00001
900	90:b1:1c:f4:2c:bd	STATIC	Po 10
20	00:0d:bc:6e:93:00	DYNAMIC	Po 1
20	90:b1:1c:f4:2c:bd	STATIC	Po 10
0	**90:b1:1c:f4:29:f3**	**LOCAL_DA**	**00001**
0	**90:b1:1c:f4:2c:bd**	**LOCAL_DA**	**00001**

(Take note that the MAC addresses of the SVI interfaces of both VLT peers are registered in the L2 CAM as LOCAL_DA addresses)

5.3 CR1 Configurations and Verification

CR1#show run | be interface Loopback2

```
interface Loopback2
 ip address 2.2.2.2 255.255.255.0
 ip ospf network point-to-point
!
interface Loopback3
 ip address 3.3.3.2 255.255.255.0
 ip ospf network point-to-point
!
interface Loopback4
 ip address 4.4.4.2 255.255.255.0
 ip ospf network point-to-point
```

(These Loopback interfaces were created for testing purposes. The OSPF network type of point-to-point is meant to simulate a true /24 network and not a /32 host address, which is how a loopback interface would appear in the routing table, as OSPF considers them STUB networks.

```
!
```

CR1#show run int port-channel 1

```
interface Port-channel1
 no switchport
 ip address 192.168.20.3 255.255.255.248
!
```

CR1#show run | be router

```
router ospf 1
 router-id 172.15.1.1
 log-adjacency-changes
 passive-interface default
 no passive-interface Port-channel1
 network 2.2.2.0 0.0.0.255 area 0
 network 3.3.3.0 0.0.0.255 area 0
network 4.4.4.0 0.0.0.255 area 0
```

(The above subnets correspond to loopback interfaces lo2, lo3 and lo4. These three loopback interfaces will be advertised to the VLT pair, Dell#1 and Dell#2)

```
 network 172.15.1.0 0.0.0.255 area 0
 network 192.168.20.0 0.0.0.7 area 0
!
```

CR1#show ip ospf neighbor

(CR1 is a DROTHER)

Neighbor ID	Pri	State	Dead Time	Address	Interface
172.16.1.2	1	FULL/BDR	00:00:31	192.168.20.2	Port-channel1
172.17.1.1	1	FULL/DR	00:00:38	192.168.20.1	Port-channel1

```
CR1#
!
```

CR1#show ip route

(Output Truncated)

```
    2.0.0.0/24 is subnetted, 1 subnets
C       2.2.2.0 is directly connected, Loopback2
    3.0.0.0/24 is subnetted, 1 subnets
C       3.3.3.0 is directly connected, Loopback3
O    192.168.8.0/24 [110/2] via 192.168.20.2, 02:02:34, Port-channel1
                    [110/2] via 192.168.20.1, 02:02:34, Port-channel1
```

(OSPF-learned route back to client subnet – VLAN 800)

```
    4.0.0.0/24 is subnetted, 1 subnets
C       4.4.4.0 is directly connected, Loopback4
O    192.168.9.0/24 [110/2] via 192.168.20.2, 02:02:34, Port-channel1
                    [110/2] via 192.168.20.1, 02:02:34, Port-channel1
```

(OSPF-learned route back to client subnet #2 – VLAN 900)

```
    172.17.0.0/24 is subnetted, 1 subnets
O       172.17.1.1 [110/1] via 192.168.20.1, 02:02:34, Port-channel1
    172.16.0.0/24 is subnetted, 1 subnets
O       172.16.1.2 [110/1] via 192.168.20.2, 02:02:34, Port-channel1
    192.168.20.0/29 is subnetted, 1 subnets
C       192.168.20.0 is directly connected, Port-channel1
```

(OSPF peering VLAN)

```
    10.0.0.0/24 is subnetted, 1 subnets
C       10.10.10.0 is directly connected, FastEthernet0/2
```

5.4 Access Switch A1 Configurations and Verification

A1#sh run | be spanning
spanning-tree mode pvst
spanning-tree vlan 1,800,900 priority 61440
(A1 access switch is configured to NOT be the STP root bridge)
!
interface Port-channeL2
description Port-Channel_to_Dell_VLT_Te0/4
switchport trunk encapsulation dot1q
 switchport mode trunk
 spanning-tree portfast trunk
(This is a VLT (Virtual Link Trunk) – a Split LAG)
!
interface Vlan800
description Client_VLAN
 ip address 192.168.8.100 255.255.255.0
!
ip route 0.0.0.0 0.0.0.0 192.168.8.2
*(This default route was put in place for testing purposes, as described in the next section. The access switch (A1) was used to generate ICMP test PINGs to a loopback interface on CR1. This default route points to **Dell#2's VLAN 800 SVI interface**. It's in place to ensure that routed test traffic will have Dell#2's MAC address as the destination address in the Ethernet frame's header)*
!
A1#sh ip ro static
S* 0.0.0.0/0 [1/0] via 192.168.8.2
(Default route pointing to Dell#2's VLAN 800 SVI interface)
!
A1#sh ip arp | in 192.168.8
Internet 192.168.8.100 - 000e.8364.6d80 ARPA Vlan800
Internet 192.168.8.1 55 90b1.1cf4.2cbd ARPA Vlan800
Internet 192.168.8.2 54 90b1.1cf4.29f3 ARPA Vlan800

6.0 Peer Routing Verification Test

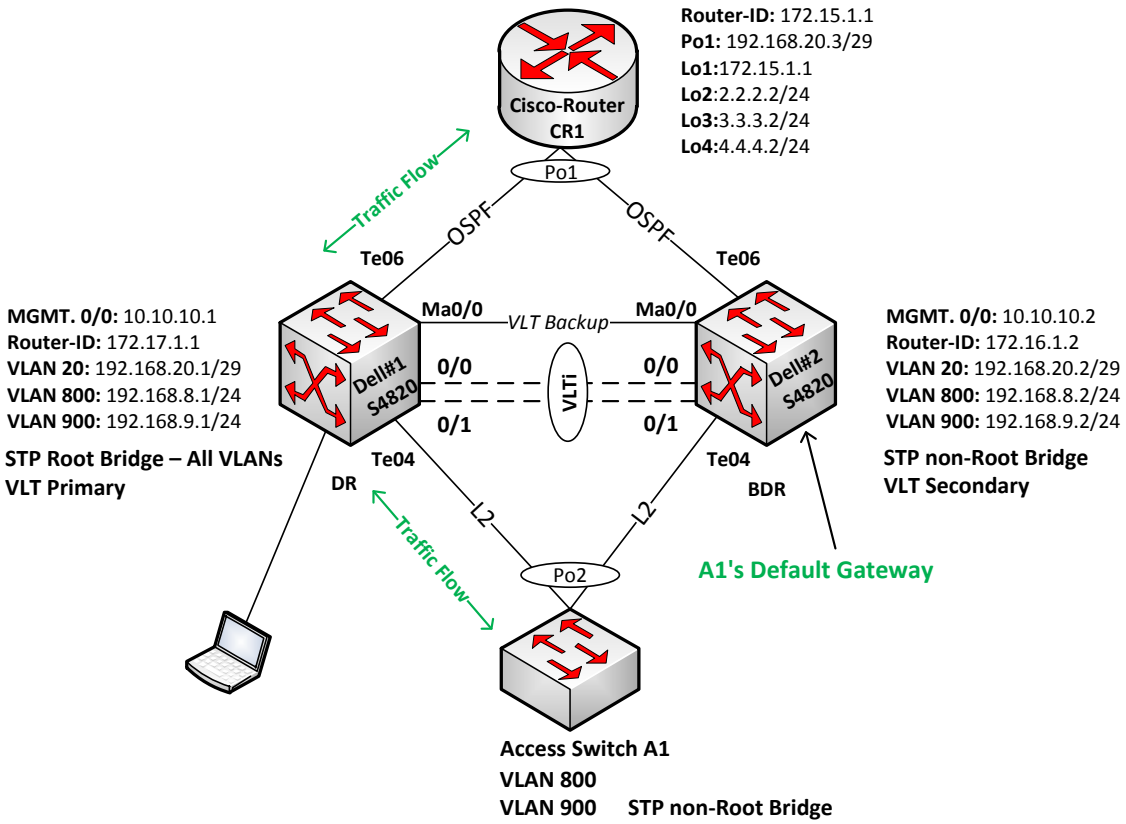


Figure 5 – Routed VLT Domain – Peer Routing Test

A quick review of some basic frame forwarding concepts in a switched network will provide the foundational understanding to appreciate the nature of this test. When an end-host needs to communicate with another host that resides outside its local subnet (VLAN), it forwards the traffic to its configured default gateway. The source MAC address will be that of the sending host, while the destination MAC address will be of the default gateway interface. The gateway in turn will source its own MAC address as it routes the frame onto the next layer-3 hop. Unlike the MAC address, the source and destination IP addresses do not change.

This test, which consists of PINGing a loopback interface on CR1 from access switch A1, is meant to showcase peer routing in action. To capture the frame forwarding sequence, ports Te0/4, Te0/6 and Po10 on Dell#1 were mirrored as part of three monitoring sessions, with the bidirectional traffic on each port bridged to a laptop that was running Wireshark protocol analyzer software.

The frame-by-frame sequence proves that, with peer routing configured on the VLT peer switches, all routed traffic received by boundary switch Dell#1, regardless of the fact that the frame's destination MAC address belonged to Dell#2, was routed without having to traverse the VLTi link (Po10) to the configured default gateway.

6.1 Relevant Switch Configurations

(Data taken from previous drawing and presented here for easier viewing)

Access Switch A1:

- **IP Address of VLAN 800:** 192.168.8.100/24
- **MAC address:** 000e.8364.6d80
- Port-channel 2 is a VLT
- Static Default: ip route 0.0.0.0 0.0.0.0 192.168.8.2
  (This default route was put in place for testing purposes. The access switch was used to generate ICMP test PINGs to a loopback interface on CR1. This default route points to Dell#2's VLAN 800 SVI interface. It's in place to ensure that routed test traffic will have Dell#2's MAC address as the destination address in the Ethernet frame's header)
- Internet 192.168.8.100 - 000e.8364.6d80 ARPA Vlan800
  (Local VLAN 800 interface entry in the ARP table)
- Internet 192.168.8.2 8 90:b1:1c:f4 :29 :f3 ARPA Vlan800
  (Dell#2's VLAN 800 SVI MAC address in the ARP table)

Dell#1 VLT Primary Switch:

- Dell#1 is the configured default gateway for A1:
- **IP address of VLAN 800:** 192.168.8.1/24
- **MAC address**: is 90:b1:1c:f4:2c:bd
  (As pointed out earlier, this MAC address is for every interface on Dell#1)

Dell#2 VLT Secondary Switch

- Dell#2 is the secondary VLT Peer switch
- **IP address of VLAN 800:** 192.168.8.2/24
- **MAC address**: is 90:b1:1c:f4 :29 :f3
  (As pointed out earlier, this MAC address is for every interface on the Dell#2 switch)

CR1 Router:

- **IP address of target subnet:** 4.4.4.0/24
- Port-channel 1 is a L3 VLT since an IP address is applied directly to it.

As can be clearly seen below, although the ICMP packets generated by A1 have Dell#2's MAC address in the destination address field of the Ethernet frame header (since Dell#2 is the default gateway), Dell#1 takes ownership of the frame after it receives it, replaces the source MAC address with its own, and then forwards it directly to the next L3 hop. See sniffer traces below. Dell#1 did not send the frame to Dell#2, which is the default gateway configured on A1.

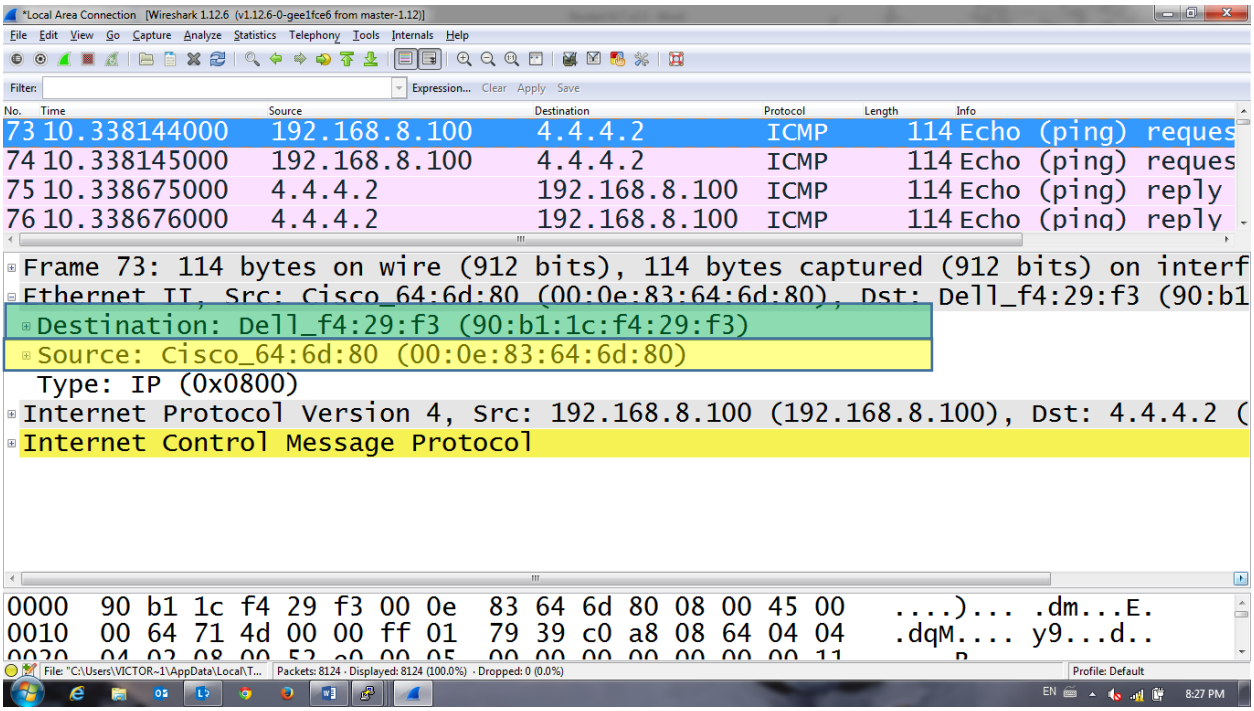


Figure 6a – The ICMP PING request destined for a loopback interface on CR1 is sent by A1 and received by Dell#1 as a result of the LACP hashing algorithm. Notice Dell#2's MAC address in the destination field.

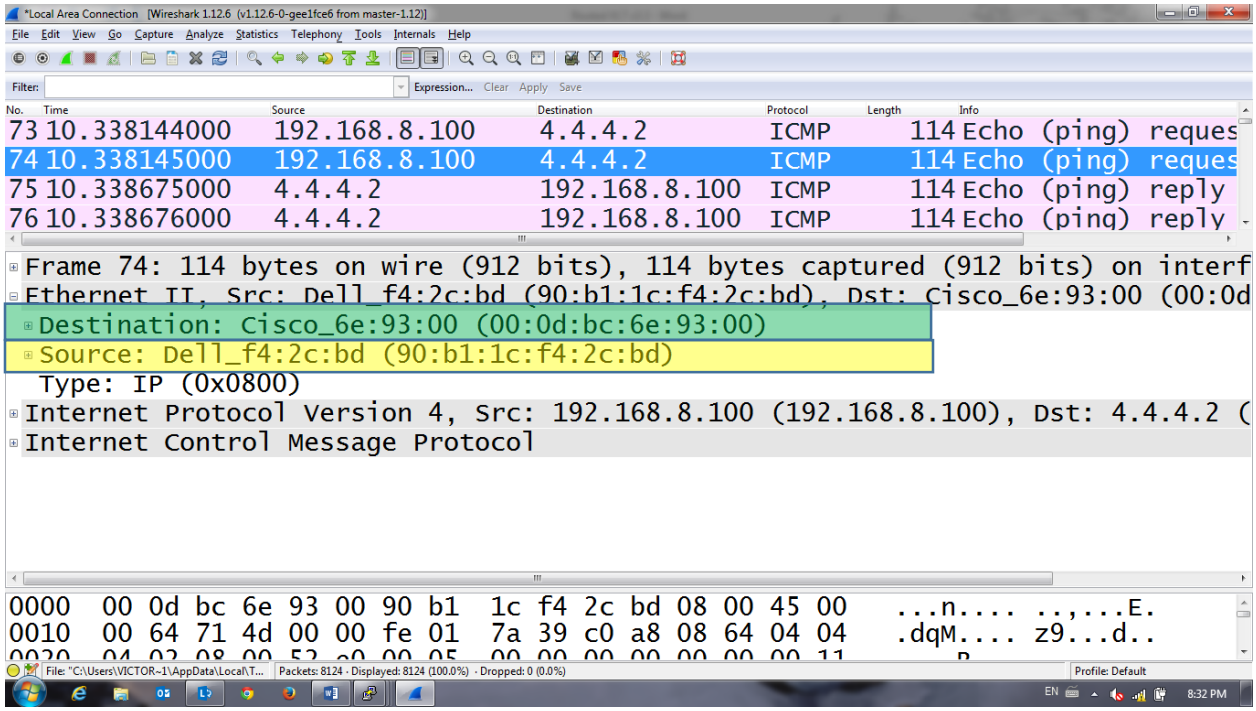


Figure 6b – Dell #1 creates a new frame with itself as the source MAC address and CR1 as the destination MAC address. Then it routes the frame to CR1 via its port-channel 1 connection. Note that it does NOT forward the frame to Dell#2, even though it is A1's default gateway.

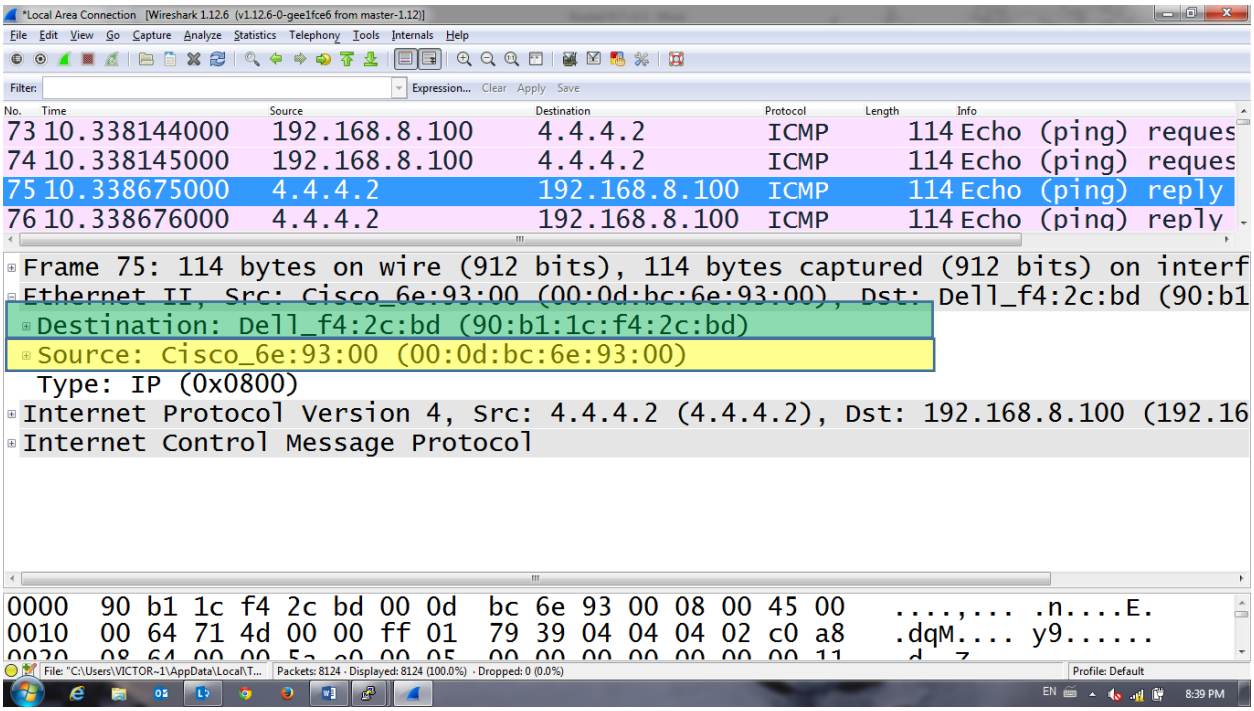


Figure 6c – CR1 responds to the PING request with a frame whose destination MAC address is for Dell#1 (2 routes to VALN 800 exist on CR1 and LACP hashed the traffic to Dell#1).

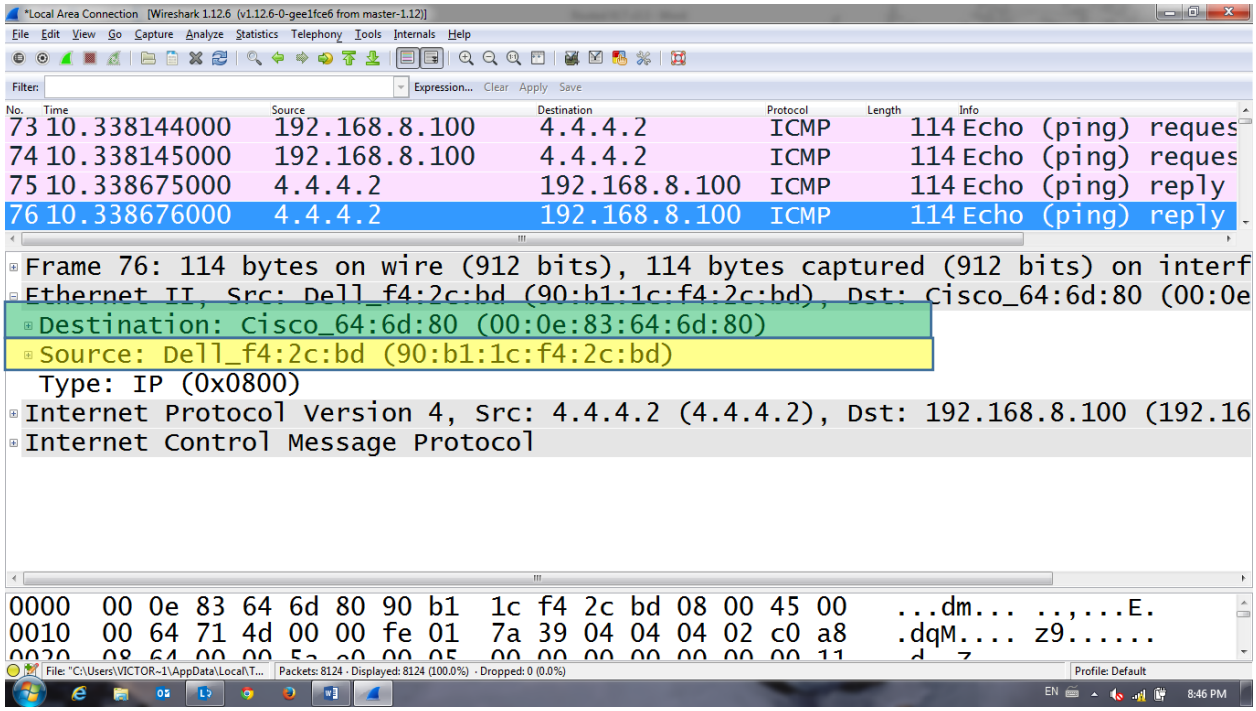


Figure 6d – Dell#1 receives the frame and forwards the PING reply to A1.

7.0 Peer Routing When a Peer Is Down

In the previous example, it was proven that Dell#1 routed traffic on behalf of its counterpart, even though it was not the default gateway. An interesting question arises when one of the VLT peers is down. Namely, what happens when a host has to renew its ARP entry for its default gateway at a time when the default gateway is unavailable?

Typically, a host will ARP for its default gateway and receive a reply from it, but only if both interfaces are on the same subnet. Otherwise, proxy ARP comes into play. Upon receiving the ARP reply, the host will make an entry in its ARP table and continue on with the business of sending traffic. With peer routing enabled, not only will a VLT peer route traffic on behalf of its counterpart, it will also answer ARP requests destined for that peer when its down. Talk about service!

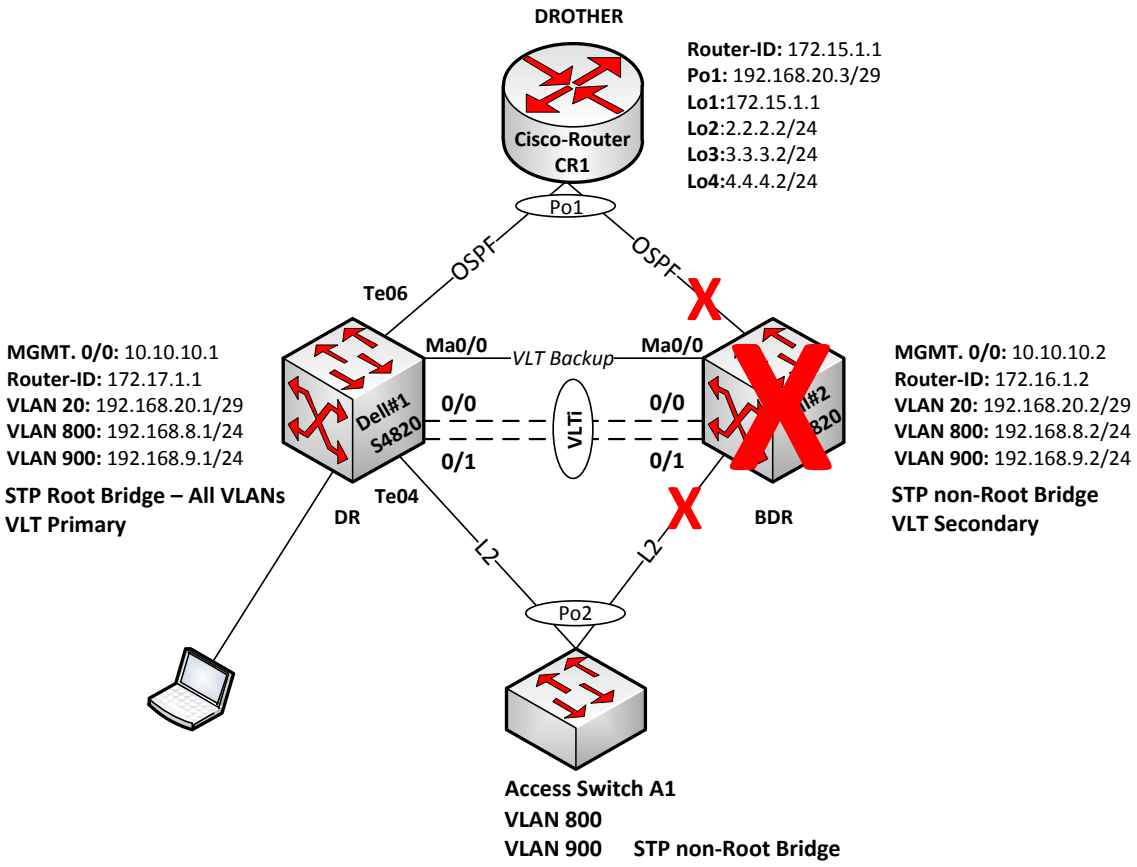


Figure 7 – Routed VLT Domain With a Peer Failure

In this instance, a test was conducted once again with a host whose default gateway is Dell#2's VLAN 800 interface, 192.168.8.2. The host's ARP cache was cleared and a PING test was initiated to the loopback interface on CR1 with the IP address of 4.4.4.2. As in the previous example, the source host is Access Switch A1.

A1#sh ip ro static
S* 0.0.0.0/0 [1/0] via 192.168.8.2
(Default gateway for this test is Dell#2's VLAN 800 interface)

A1#sh arp | in 192.168.8.2
Internet 192.168.8.2 62 90b1.1cf4.29f3 ARPA Vlan800
(Take note of the MAC address. It is Dell#2's interface address for VLAN 800)

VLT peer Dell#2 was powered down for this test.

Dell#1#sh vlt brief
VLT Domain Brief

Domain ID:	1
Role:	Primary
Role Priority:	4096
ICL Link Status:	**Down**
HeartBeat Status:	**Down**
VLT Peer Status:	**Down**
Local Unit Id:	0
Version:	6(3)
Local System MAC address:	90:b1:1c:f4:2c:bb
Remote System MAC address:	00:00:00:00:00:00
Configured System MAC address:	90:b1:1c:f4:01:01
Remote system version:	0(0)
Delay-Restore timer:	90 seconds
Peer-Routing :	**Enabled**
Peer-Routing-Timeout timer:	0 seconds
Multicast peer-routing timeout:	150 seconds

Dell#1#

Access Switch A1's ARP table was cleared, which forced it to immediately ARP for its configured default gateway. But its default gateway (Dell#2) was unavailable, so Dell#1 replied with Dell#2's MAC address.

A1#clear arp

Dell#1#
02:47:47 : cp-IP ARP: rcvd req src 192.168.8.100 00:0e:83:64:6d:80, dst 192.168.8.2 Port-channel 2 Vlan 800
02:47:47 : cp-IP ARP: sent rep src 192.168.8.2 90:b1:1c:f4:29:f3, dst 192.168.8.100 00:0e:83:64:6d:80 Vlan 800
(Dell#1 receives a unicast ARP request from A1 for Dell#2 and responds on its behalf)

A look at the Wireshark captures on Dell#1 will provide the necessary detail to appreciate what is occurring.

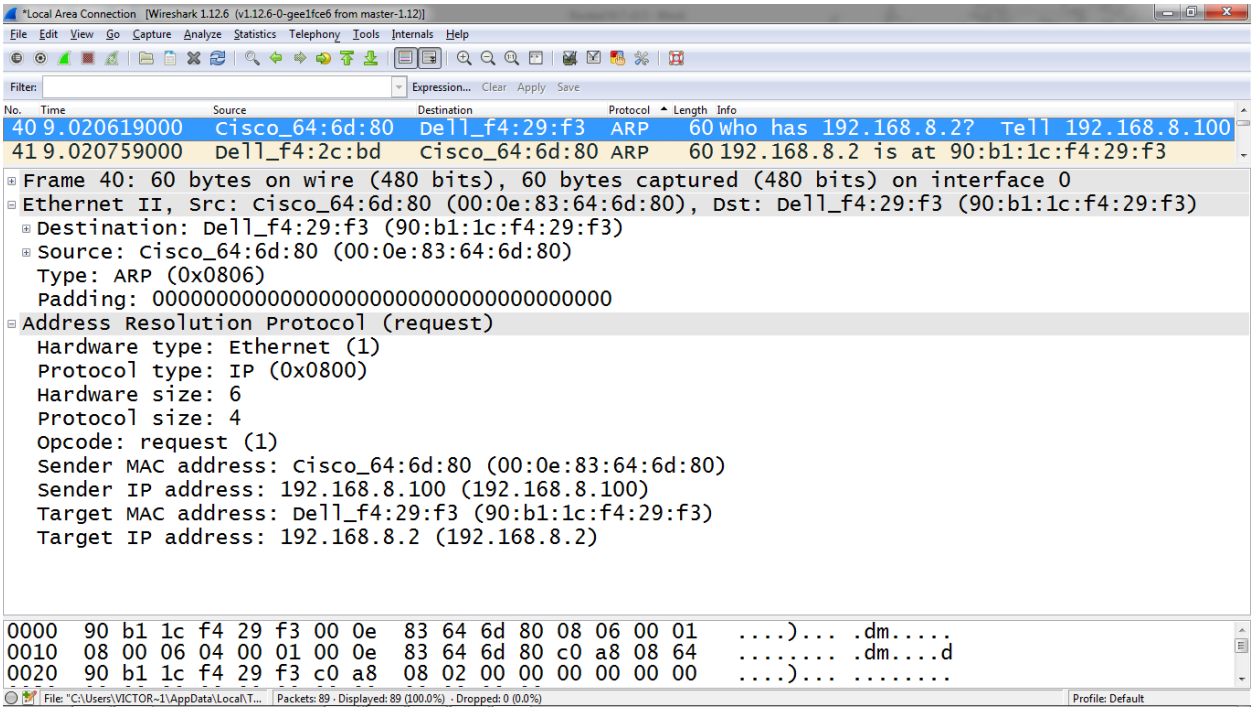


Figure 8a – Dell#1 receives the unicast ARP request from A1 for Dell#2's MAC address.

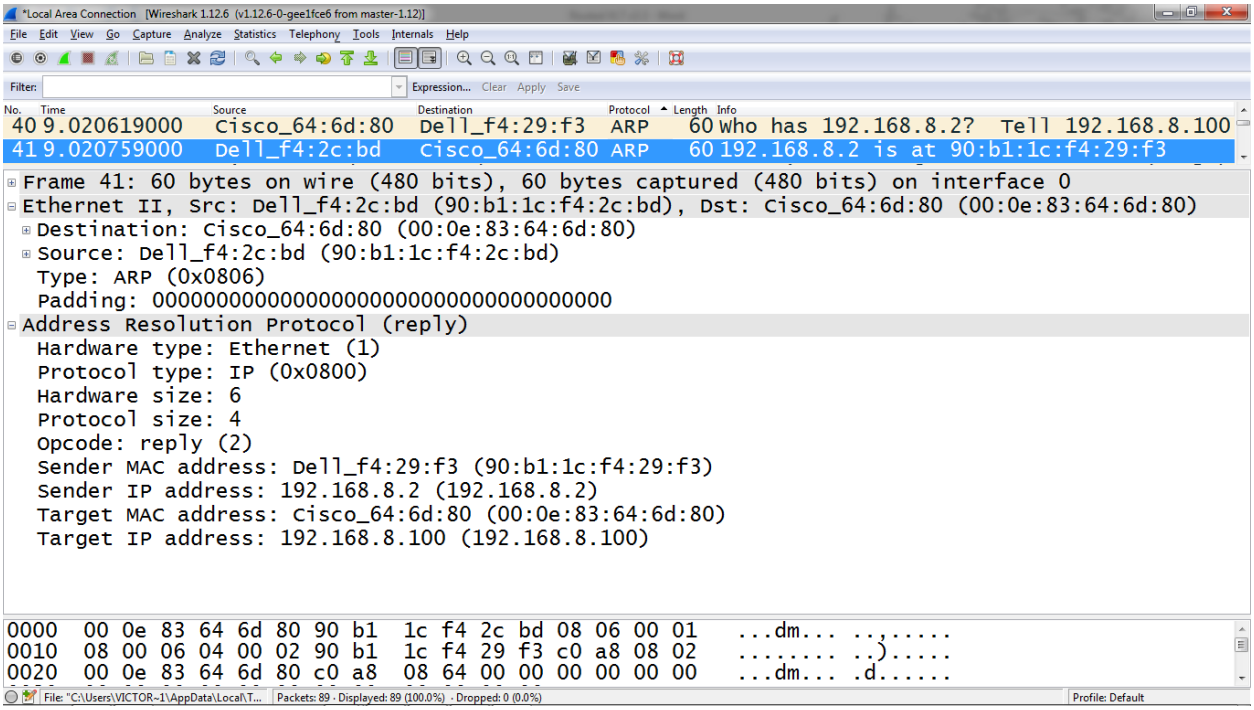


Figure 8b – Dell#1 responds to A1's ARP request with a reply that includes Dell#2's MAC address

With Dell#2 down, didn't Dell#1 clear its ARP and MAC address table entries for 192.168.8.2? Yes, as is seen in the following output.

Dell#1#sh arp ip 192.168.8.2

```
Protocol  Address        Age(min)  Hardware Address   Interface  VLAN          CPU
--------------------------------------------------------------------------------
Dell#1#
```

(No data for 192.168.8.2)

Dell#1#sh mac-address-table vlan 800

```
Codes: *N - VLT Peer Synced MAC
VlanId          Mac Address                Type          Interface     State
 800            00:0e:83:64:6d:80          Dynamic       Po 2          Active
 800            00:0e:83:64:6d:8d          Dynamic       Po 2          Active
Dell#1#
```

(The MAC address for Dell#2's VLAN 800 interface is also nowhere to be found)

So how did Dell#1 return the correct ARP information for 192.168.8.2? The answer can be found in Dell#1's system CAM!

Dell#1#sh cam mac stack-unit 0 port-set 0 address 90:b1:1c:f4:29:f3
!

```
VlanId          Mac Address          Region        Interface
 0              90:b1:1c:f4:29:f3    LOCAL_DA      00001
Dell#1#
```

(With peer routing enabled, the MAC address for the VLT peer remained in the system CAM table!)

8.0 Routed mVLT with VLT Proxy Gateway - A Case Study

A recent deployment at a large school district presented a set of requirements that fit nicely into the Routed mVLT with Proxy Gateway solution. The deployment involved two sites, a high school and an elementary school, about a mile apart from each other, with virtual server farms resident at each site. The schools were connected to each other via a single 10G "dark fiber" link.

Existing Environment at Each School:

- Juniper EX 4200 1G switches in the access layer of the campus (Virtual Chassis) and the 1G server farm.
- 2 x Juniper EX 4500 10G switches in a virtual chassis configuration at each school acting as a collapsed core (L3/L3 boundary) and for 10G server access.
- Unified Threat Management (UTM) HA appliance pair for perimeter security. ISP circuit terminates on the active node.
- One 10G dark fiber link between EX 4500 Virtual Chassis at each school
- Static Routing

The high-level requirements consisted of the following:

- L2 adjacency between sites to support vMotion of virtual workloads between schools.
- Certain VLANs were local to each site and some needed to be spanned across the dark fiber link
- Since the 10G dark fiber link is to be replaced by 2 x 1G fiber links, there was a desire to minimize crosslink traffic – traffic "tromboning."
- In the event of a failure of either the local UTM HA cluster or its link to the ISP, Internet traffic has to default to the other site.
- Replace static routing with OSPF and BGP for dynamic failover.

Figure 9 depicts the "interesting" part of the overall topology: the routed VLT pairs and their extension across geographically disparate sites as part of a routed mVLT domain. **Only the relevant configurations for VLT Proxy Gateway will be shown in section 8.2.** The VLT peers at each school consisted of Dell Networking s4810 10/40G switches, which replaced the Juniper EX4500s. The Juniper EX4200 access layer switches remained in place and were dually-attached to each of the VLT peers.

VLT Proxy Gateway functionality is ideally deployed as part of a fully-meshed (looped triangle) topology between VLT domains, although a partially-meshed (square) topology is also supported. In this case, there was only one link connecting the two VLT domains, which required some modifications to the deployment, as will be shown in the next sections.

Moreover, VLT Proxy Gateway can be deployed in a manner that leverages LLDP TLVs (dynamic) or static configurations. This example will cover the latter. Please see DNOS 9.8 (and above) Configuration Guides for a particular switch platform for further detail on the dynamic option.

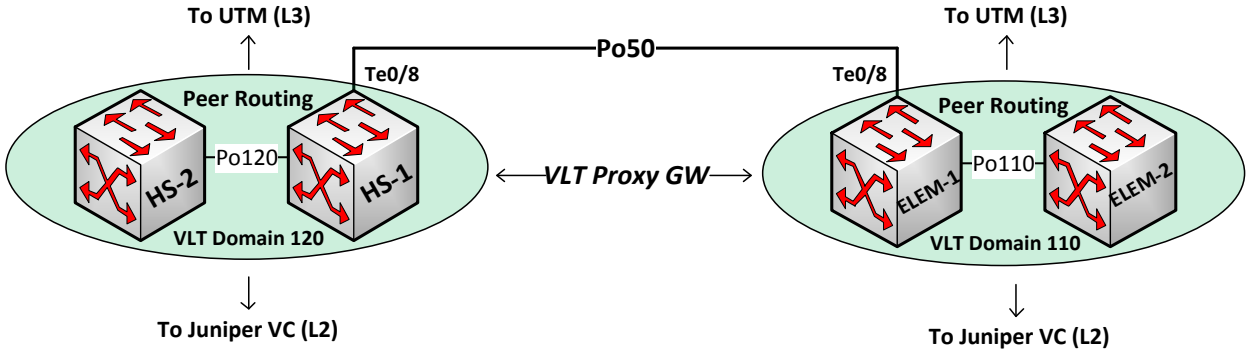


Figure 9 – VLT Proxy Gateway Case Study

8.1 Design Highlights

- All switches are configured for VLT and peer routing, exactly as described in the previous sections and as shown in Figure 9 above.
- Since there is only one inter-connecting link between the two VLT domains, OSPF should be configured in a point-to-point network design to remove any complexity or anomalous behavior that an OSPF broadcast network may present.
- Rapid STP (802.1w) is configured at each site, but disabled between sites, so each VLAN, including the ones that are spanned across the mVLT domain, has a local root bridge
- All VLT peers are L3/L2 boundary routers. OSPF point-to-point (/30) peering VLANs are used to establish the OSPF adjacencies, while all other VLAN interfaces are made passive.
- Port-channel 50, an L2 trunk with one member port, is used to link the two sites.
- VLT Proxy Gateway is configured statically. **This is required since there was only one link between sites.**

8.2 VLT Proxy Gateway-specific Configurations

(These are the configuration lines that need to be added in addition to the baseline VLT, peer routing and OSPF configurations that were covered in previous sections to build the mVLT domain and enable VLT Proxy Gateway.)

High School:

HS-1:

```
vlt domain 120
 peer-link port-channel 120
 back-up destination 10.1.1.3
 primary-priority 4096
 system-mac mac-address 02:01:e8:d8:93:e3
 unit-id 0
 peer-routing
!
 proxy-gateway static
  remote-mac-address 00:01:e8:8b:ff:4f
  remote-mac-address 00:01:e8:d8:93:04
```

(These MAC addresses are the system L2 interface addresses for each switch at the remote site, HS-1 and HS-2. They are entered under the "VLT domain" sub-configuration mode. These MAC addresses can be obtained by executing a "show interface" on each switch, as was shown in previous sections. These statements inform the local switch of the MAC addresses at the remote site for whom it has to proxy.

!

*These MAC addresses will be entered into the data table provided by the **My Station TCAM** and can only be viewed by executing a hardware information dump from the Broadcom shell – this is NOT recommended for anyone but Dell Networking Pro Support engineers to execute given the ease with which the data can be corrupted. These addresses will also appear as dynamic entries in the L2 CAM, not LOCAL_DA addresses as was the case with peer routing.)*

!

```
interface TenGigabitEthernet 0/8
 description "To ELEM-1 LightPath 10Gb"
 no ip address
!
 port-channel-protocol LACP
  port-channel 50 mode active
  no shutdown
```

(Physical 10G port for SMF fiber to Elementary school site)

```
interface Port-channel 50
 description "mVLT port channel to ELEM-1"
```

no ip address
switchport
no spanning-tree
(STP is disabled between sites)
vlt-peer-lag port-channel 50
(Note that the crosslink port channel is also a VLT port channel)
no shutdown
(This is an L2 trunk between sites)
!
interface Vlan 100
description OSPF Peering VLAN to HS-2
ip address 10.10.100.1/30
ip ospf network point-to-point
no shutdown
!
interface Vlan 101
description ospf peering vlan across VLTPG_Po50
ip address 10.10.101.1/30
tagged Port-channel 50
ip ospf network point-to-point
no shutdown
!
router ospf 1
router-id 4.4.4.4
network 10.10.100.0/30 area 0
network 10.10.101.0/30 area 0
!
HS-1#sh ip ospf nei
!

Neighbor ID	Pri	State	Dead Time	Address	Interface	Area
2.2.2.2	1	FULL/ -	00:00:39	10.10.100.2	Vl 100	0
3.3.3.3	1	FULL/ -	00:00:32	10.10.101.2	Vl 101	0

HS-2:

vlt domain 120
peer-link port-channel 120
back-up destination 10.1.1.2
primary-priority 24576
system-mac mac-address 02:01:e8:d8:93:e3
unit-id 1
peer-routing
!

proxy-gateway static
remote-mac-address 00:01:e8:8b:ff:4f
remote-mac-address 00:01:e8:d8:93:04
(Notice that the MAC addresses are the same that were entered on HS-1, its VLT peer)
!
interface Vlan 100
description OSPF peering VLAN to HS-1
ip address 10.10.100.2/30
ip ospf network point-to-point
no shutdown
!
router ospf 1
router-id 2.2.2.2
network 10.10.100.0/30 area 0
!
HS-2#sh ip ospf nei
!

Neighbor ID	Pri	State	Dead Time	Address	Interface	Area
4.4.4.4	1	FULL/ -	00:00:33	10.10.100.1	Vl 100	0

Elementary School:

ELEM-1:

vlt domain 110
(Notice that HS-1 and HS-2 are in a separate VLT domain – VLT domain 120)
peer-link port-channel 110
back-up destination 10.1.1.1
primary-priority 4096
system-mac mac-address 02:01:e8:d8:93:02
unit-id 0
peer-routing
!
proxy-gateway static
remote-mac-address 00:01:e8:d8:93:07
remote-mac-address 00:01:e8:d8:93:e5
(These MAC addresses are the system L2 interface addresses for each switch at the remote site, HS-1 and HS-2. They are entered under the "VLT domain" sub-configuration mode. These MAC addresses can be obtained by executing a "show interface" on each switch, as was shown in previous sections. These statements inform the local switch of the MAC addresses at the remote site for whom it has to proxy. The peer routing statement is used to inform it of its VLT peer's MAC address for whom it also has to proxy
!

*These MAC addresses will be entered into the data table provided by the **My Station TCAM** and can only be viewed by executing a hardware information dump from the Broadcom shell – this is NOT recommended for anyone but Dell Networking Pro Support engineers to execute given the ease with which the data can be corrupted. These addresses will also appear as dynamic entries in the L2 CAM, not LOCAL_DA addresses as was the case with peer routing.)*

```
!
interface TenGigabitEthernet 0/8
 description "To HS-1 LightPath 10Gb"
 no ip address
!
 port-channel-protocol LACP
  port-channel 50 mode active
  no shutdown
```

(Physical 10G port for SMF fiber to High School site)

```
!
interface Port-channel 50
 description "mVLT port channel to HS-1"
 no ip address
 switchport
no spanning-tree
```

(STP is disabled between sites)

```
 vlt-peer-lag port-channel 50
```

(Note that the crosslink port channel is also a VLT port channel)

```
 no shutdown
```

(This is an L2 trunk between sites)

```
!
interface Vlan 101
 description ospf peering vlan across VLTPG_Po50
 ip address 10.10.101.2/30
 tagged Port-channel 50
 ip ospf network point-to-point
 no shutdown
!
interface Vlan 102
 description ospf peering vlan to ELEM-2
 ip address 10.10.102.1/30
 ip ospf network point-to-point
 no shutdown
!
router ospf 1
 router-id 3.3.3.3
 network 10.10.101.0/30 area 0
 network 10.10.102.0/30 area 0
!
```

ELEM-1#sh ip ospf nei

!

Neighbor ID	Pri	State	Dead Time	Address	Interface	Area
4.4.4.4	1	FULL/ -	00:00:33	10.10.101.1	Vl 101	0
1.1.1.1	1	FULL/ -	00:00:34	10.10.102.2	Vl 102	0

ELEM-2:

vlt domain 110

(Notice that HS-1 and HS-2 are in a separate VLT domain – VLT domain 120)

peer-link port-channel 110
back-up destination 10.1.1.0
primary-priority 24576
system-mac mac-address 02:01:e8:d8:93:02
unit-id 1
peer-routing
!
proxy-gateway static
remote-mac-address 00:01:e8:d8:93:07
remote-mac-address 00:01:e8:d8:93:e5

(Note that the MAC addresses are the same ones that were entered on ELEM-1, its VLT peer)

!
interface Vlan 102
description ospf peering vlan to ELEM-1
ip address 10.10.102.2/30
ip ospf network point-to-point
no shutdown
!
router ospf 1
router-id 1.1.1.1
network 10.10.102.0/30 area 0
!
ELEM-2#sh ip ospf nei

!

Neighbor ID	Pri	State	Dead Time	Address	Interface	Area
3.3.3.3	1	FULL/ -	00:00:32	10.10.102.1	Vl 102	

9.0 VLT Proxy Gateway Forwarding When a Peer VLT Domain is Down

VLT Proxy Gateway functionality for a migrated host will NOT continue to function if that end-host needs to renew its ARP entry for its configured default gateway at a time when that default gateway is down or unreachable. For example, assume that VLT domain 1 exists at DC 1 and VLT domain 2 exists at DC 2. Also, workload X originally existed at DC1 and defaulted to VLT domain 1, but then it was migrated to site 2. Accordingly, VLT Proxy Gateway allowed VLT domain 2 to proxy on behalf of VLT domain 1.

Then DC 1 suffers a total outage and VLT domain 1 becomes completely unreachable by VLT domain 2. With VLT Proxy Gateway **statically** configured between VLT domains 1 and 2, workload X will continue to have connectivity and default to VLT domain 2 at DC 2. However, when workload X's ARP entry for its configured default gateway times out, it will cease to communicate because VLT domain 2 will NOT answer ARP requests on behalf of VLT domain 1 – in other words, there is no proxy ARP between VLT domains. Since one of the VLT peers of VLT domain 1 is the configured default gateway for the workload, it has to respond to the ARP request.

For the case of **dynamically** configured VLT Proxy Gateway, which leverages periodic exchanges of LLDP TLVs to share forwarding information between sites, the workload will become non-responsive shortly after the two sites lose complete visibility to each other and the remote MAC-address entries stored in the MY Station TCAM are flushed.

On the other hand, Peer routing will indeed still function, and the active/functioning VLT peer will reply to an end-host's ARP renewal request, even if the VLT peer is down, as clearly shown in section 7.0.

Summary

Dell Networking's Virtual Link Trunking technology can prove to be an indispensable tool to maximize data throughput and network availability. This can be particularly desirable for modern data centers that host demanding application workloads and deliver highly available cloud-based services. Virtualized data centers that want to take full advantage of available server CPU and memory capabilities to maximize guest OS consolidation ratios will require robust network I/O to match.

The ability to run a routing protocol over the layer-2 VLT construct offers added deployment flexibility and optimized traffic flows to maximize link efficiency and usage. It should be understood that Peer Routing and VLT Proxy Gateway are meant to be traffic optimization solutions, not High Availability (HA) solutions. Having some HA functionality in the event of either a VLT peer failure (Peer Routing) or a VLT domain failure when running static VLT Proxy Gateway are added bonuses. As of this writing Dell engineering is investigating the possibility of offering proxy ARP functionality between VLT domains.