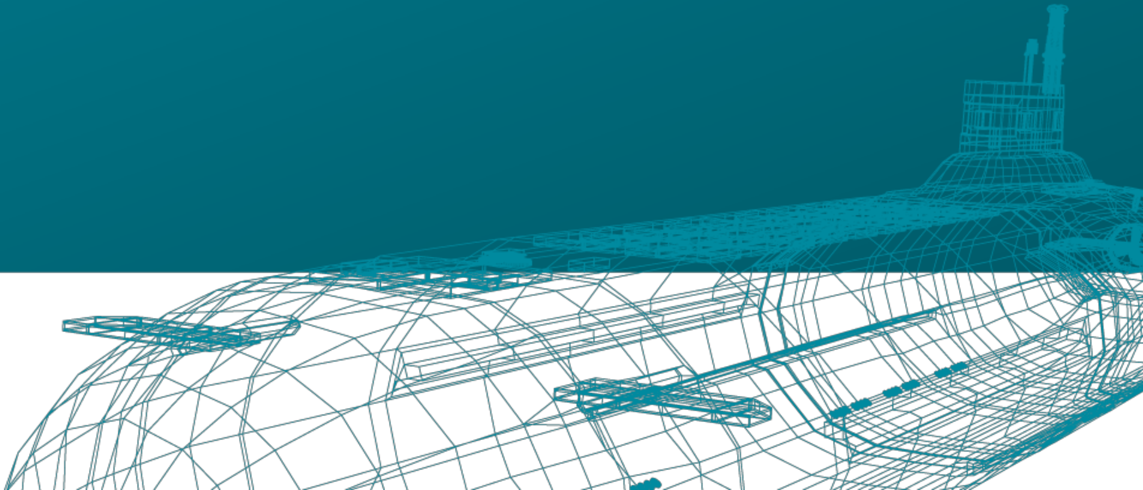
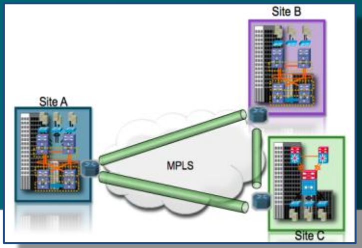
OTV – L2 Transport mezi DC

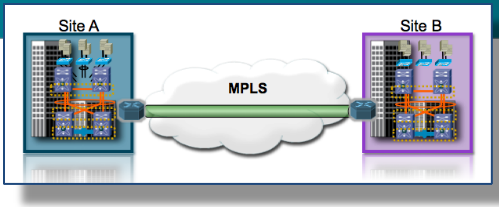
Roman Aprias, CCIE #19975 R&S SP
Team Leader, Network Architect
roman.aprias@alefnula.com

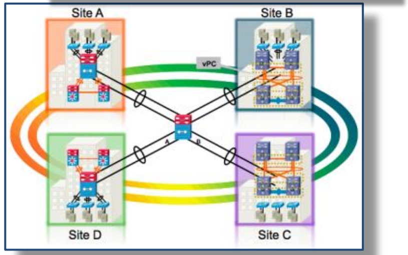


ALEF NULA

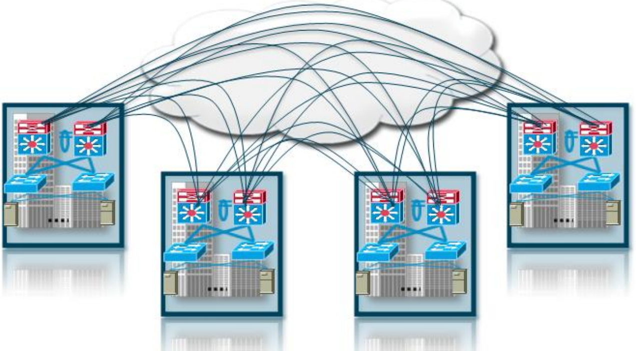
Traditional Layer 2 VPNs

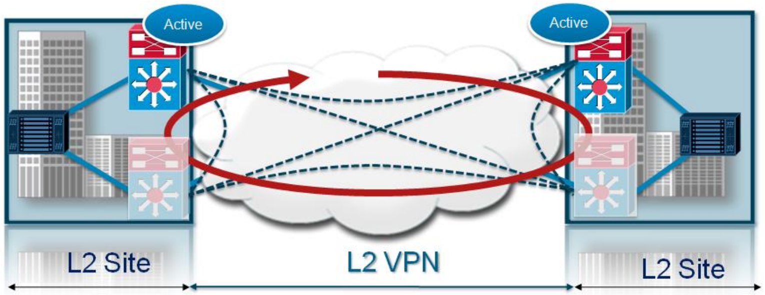






- › **Traditional Layer 2 VPNs**
  - – Dark fiber, DWDM, EoMPLS, VPLS
- › **Flooding Behavior**
  - – Traditional Layer 2 VPN technologies rely on flooding to propagate MAC reachability
  - – The flooding behavior causes failures to propagate to every site in the Layer 2 VPN
- › **Pseudo-Wires Maintenance**
  - – Before any learning can happen a full mesh of pseudo-wires/ tunnels must be in place
  - – For N sites, there will be N*(N-1)/2 pseudo-wires. Complex to add and remove sites
  - – Head-end replication for multicast and broadcast. Sub-optimal BW utilization
- › **Multi-homing**
  - – Require additional protocols to support Multi-homing
  - – STP is often extended across the sites of the Layer 2 VPN. Very difficult to manage as the number of sites grows
  - – Malfunctions on one site will likely impact all sites on the VPN





Overlay Transport Virtualization

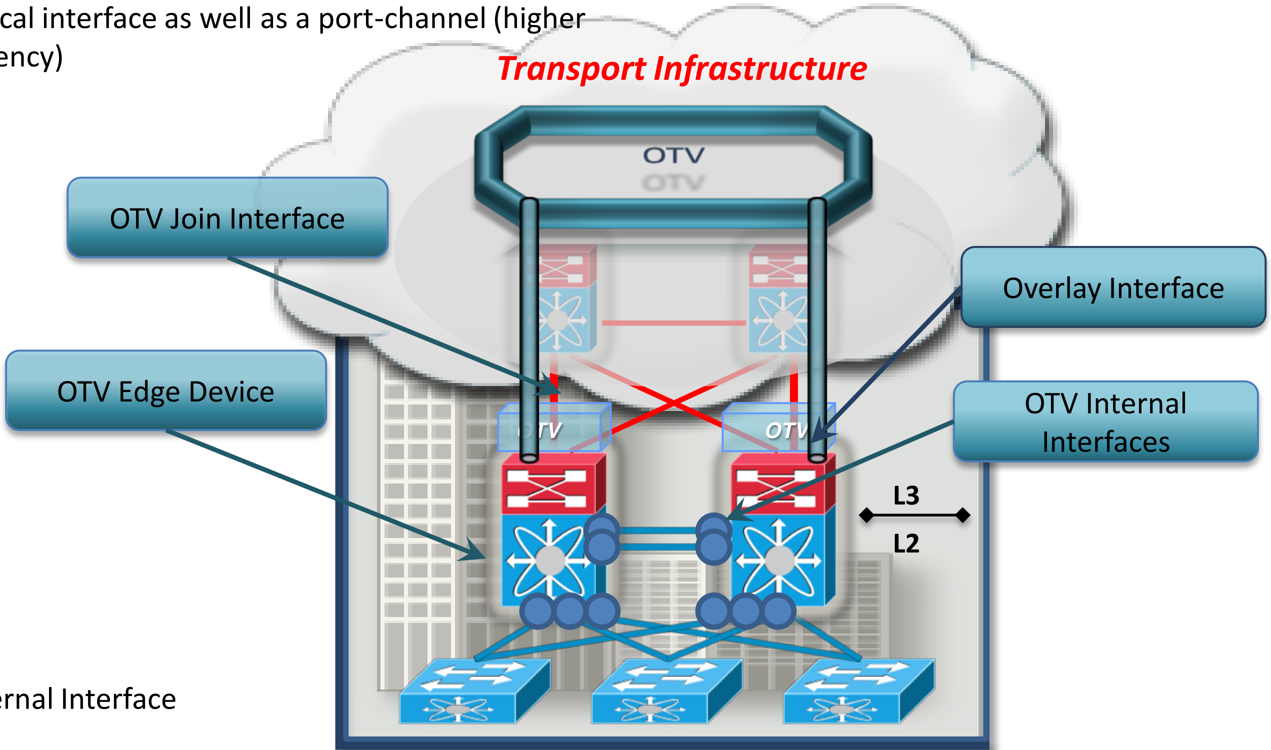
“MAC in IP” technique to extend L2 domains over any transport infrastructure

- **Flooding Based Learning => Control-Plane Based Learning**
  - Move to a Control Plane protocol that proactively advertises MAC addresses and their reachability instead of the current flooding mechanism
- **Pseudo-wires and Tunnels => Dynamic Encapsulation**
  - Not require static tunnel or pseudo-wire configuration
  - Offer optimal replication of traffic done closer to the destination, which translates into much more efficient bandwidth utilization in the core
- **Complex Dual-homing => Native Automated Multi-homing**
  - Allow load balancing of flows within a single VLAN across the active devices in the same site, while preserving the independence of the sites. STP confined within the site (each site with its own STP Root bridge)

ALEF NULA

OTV Terminology

- **Edge Device**
  - Is responsible for performing all the OTV functionality
  - Can be located at the Aggregation Layer as well as at the Core Layer depending on the network topology of the site
  - A given site can have multiple OTV Edge Devices (multi-homing)
- **Join Interface**
  - One of the uplink interfaces of the Edge Device
  - Point-to-point routed interface and it can be a single physical interface as well as a port-channel (higher resiliency)
- **Internal Interface**
  - Face the site and carry at least one of the VLANs extended through OTV
  - Behave as regular layer 2 interfaces. No OTV configuration is needed on the OTV Internal Interfaces
- **Overlay Interface**
  - Virtual interface where all the OTV configuration is placed
  - Logical multi-access multicast-capable interface
  - Encapsulates the site Layer 2 frames in IP unicast or multicast packets that are then sent to the other sites

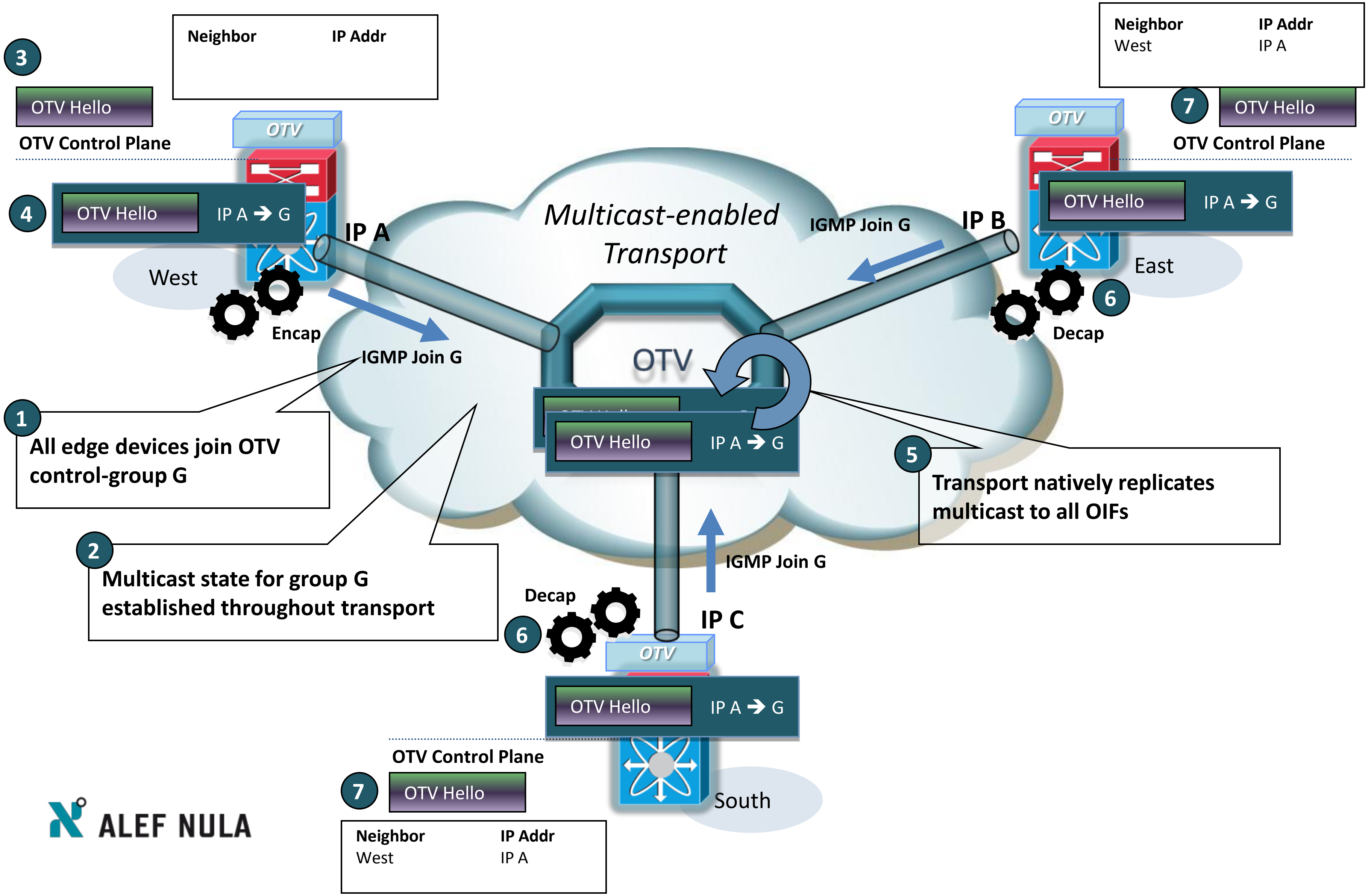


● = OTV Internal Interface

OTV Control Plane

- **Neighbor Discovery and Adjacency Formation**
  - Before any MAC address can be advertised the OTV Edge Devices must:
    - Discover each other
    - Build a neighbor relationship with each other
  - The neighbor relationship can be built over a transport infrastructure, that can be:
    - multicast-enabled
    - unicast-only
- **Building the MAC tables**
  - OTV proactively advertises MAC reachability (control-plane learning)
  - MAC addresses advertised in the background once OTV has been configured
  - No specific configuration is required
  - IS-IS is the OTV Control Protocol running between the Edge Devices. No need to learn how IS-IS works

OTV Control Plane
Neighbor Discovery (over Multicast Transport)



Neighbor	IP Addr

Neighbor	IP Addr
West	IP A

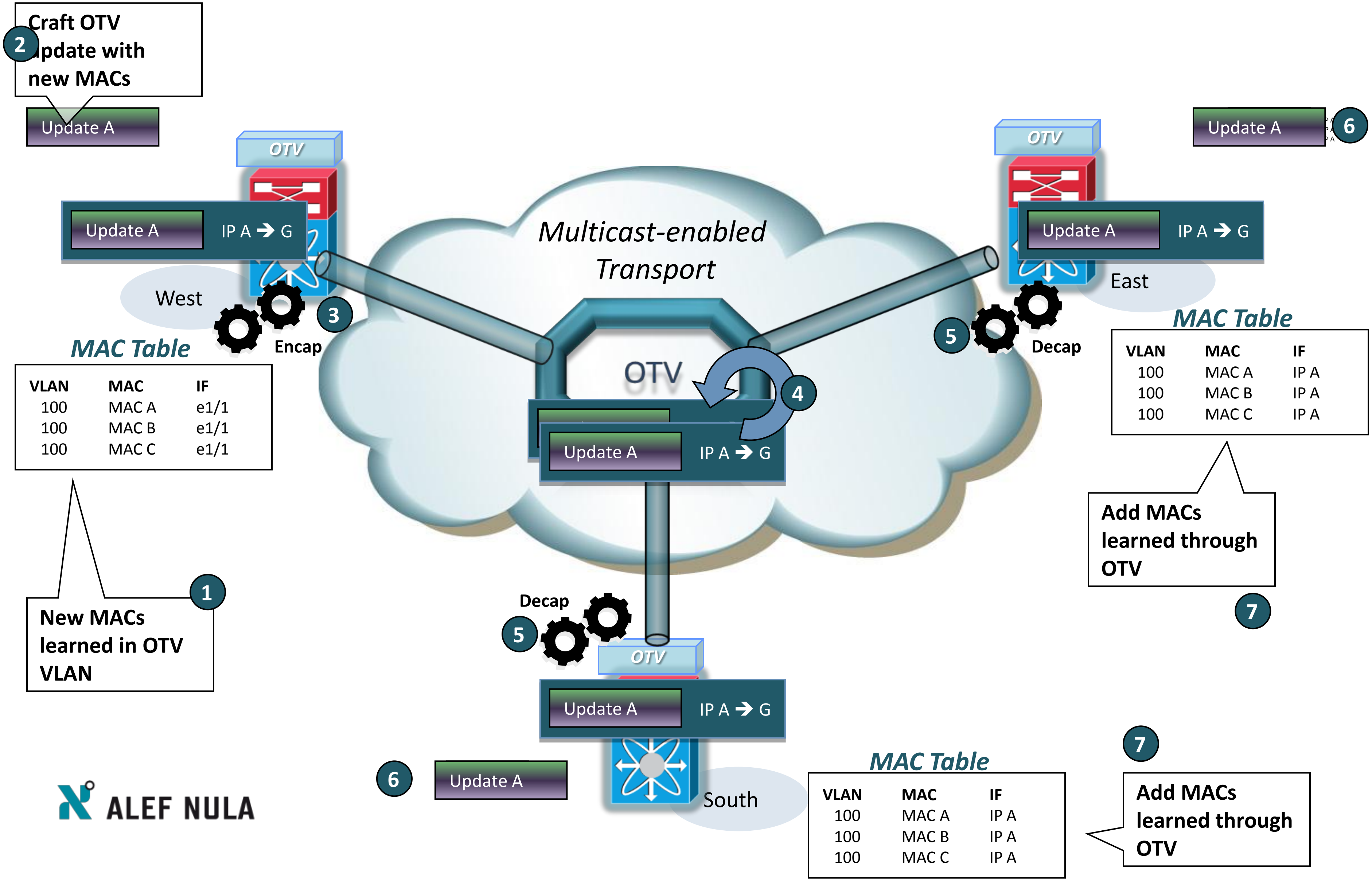
1 All edge devices join OTV control-group G

2 Multicast state for group G established throughout transport

5 Transport natively replicates multicast to all OIFs

Neighbor	IP Addr
West	IP A

OTV Control Plane
Route (MAC) Advertisements (over Multicast Transport)



2 Craft OTV update with new MACs

MAC Table (West)

VLAN	MAC	IF
100	MAC A	e1/1
100	MAC B	e1/1
100	MAC C	e1/1

1 New MACs learned in OTV VLAN

MAC Table (East)

VLAN	MAC	IF
100	MAC A	IP A
100	MAC B	IP A
100	MAC C	IP A

7 Add MACs learned through OTV

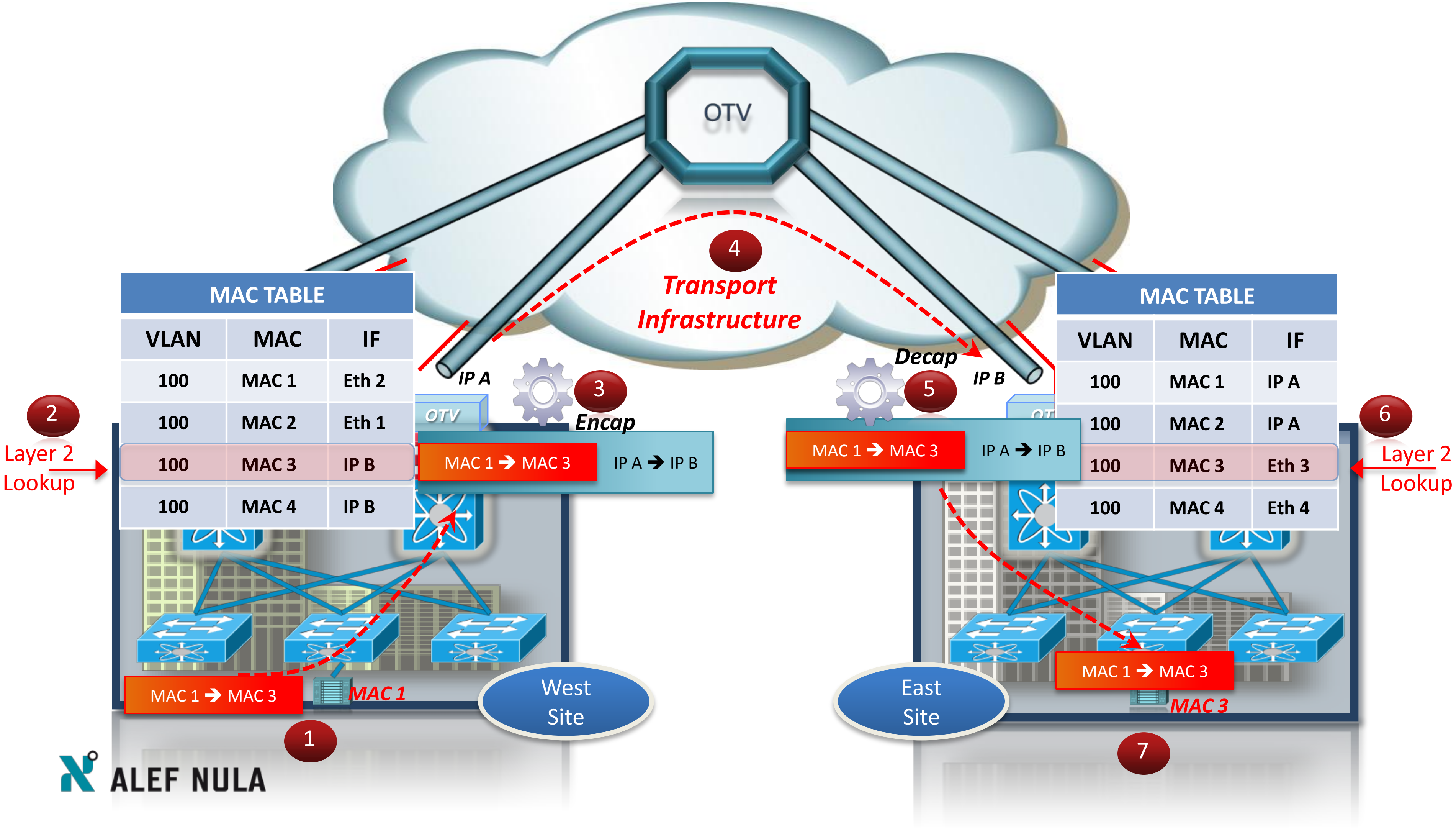
MAC Table (South)

VLAN	MAC	IF
100	MAC A	IP A
100	MAC B	IP A
100	MAC C	IP A

7 Add MACs learned through OTV

ALEF NULA

OTV Data Plane: Inter-Site Packet Flow



MAC TABLE

VLAN	MAC	IF
100	MAC 1	Eth 2
100	MAC 2	Eth 1
100	MAC 3	IP B
100	MAC 4	IP B

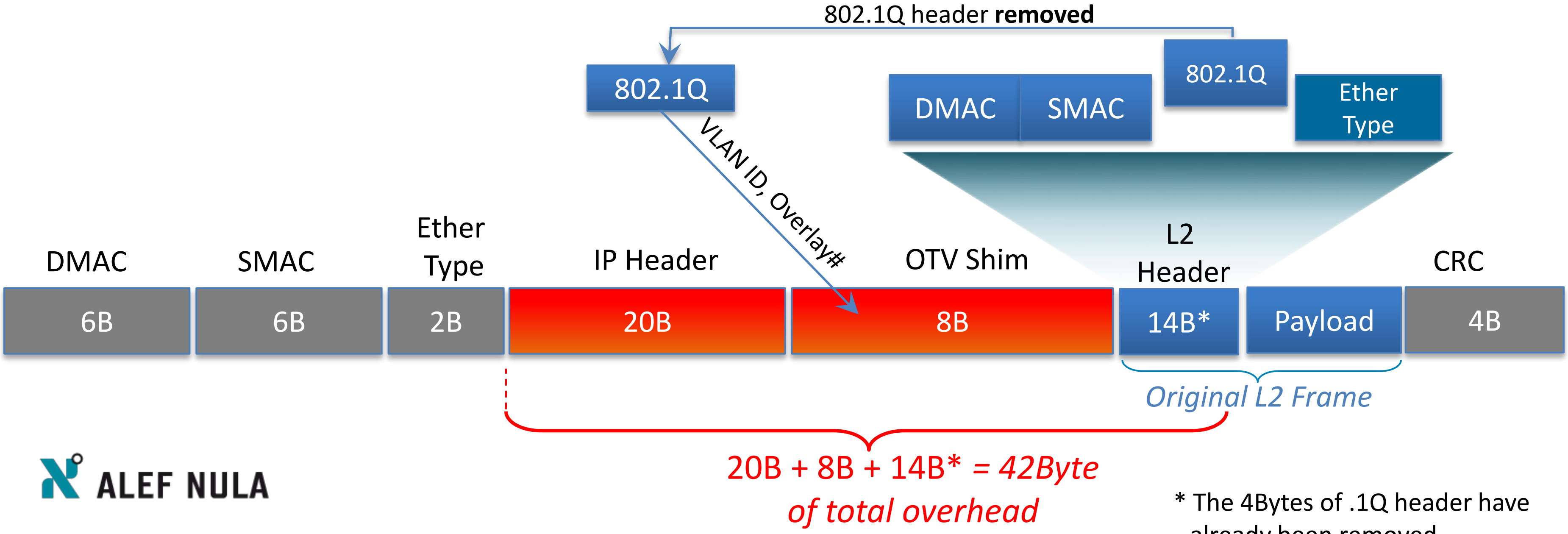
MAC TABLE

VLAN	MAC	IF
100	MAC 1	IP A
100	MAC 2	IP A
100	MAC 3	Eth 3
100	MAC 4	Eth 4

ALEF NULA

OTV Data Plane - Encapsulation

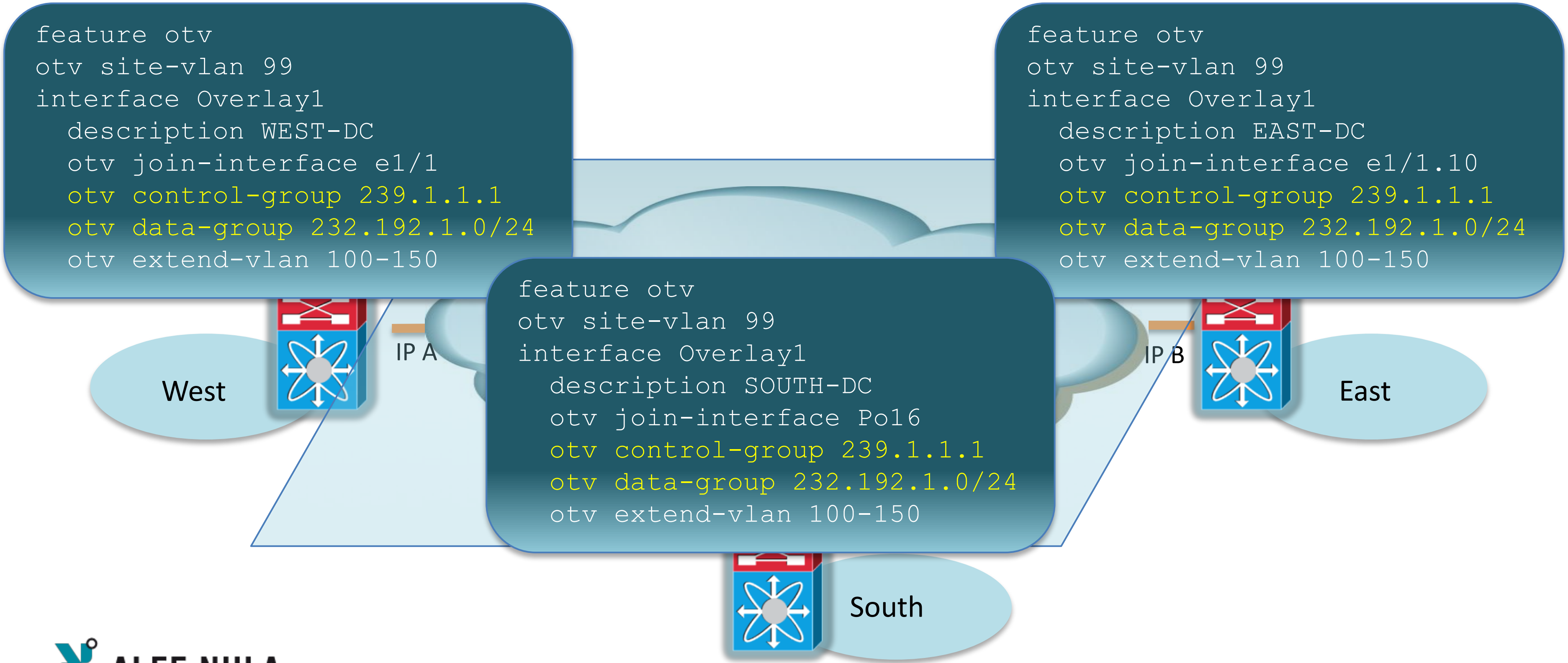
- OTV encapsulation adds 42 Bytes to the packet IP MTU size
- Outer IP Header and OTV Shim Header in addition to original L2 Header stripped off of the .1Q header
- The outer OTV shim header contains information about the overlay (VLAN, overlay number)
- The 802.1Q header is removed from the original frame and the VLAN field copied over into the OTV shim header



* The 4Bytes of .1Q header have already been removed

Configuration
OTV over a Multicast Transport

- Minimal configuration required to get OTV up and running



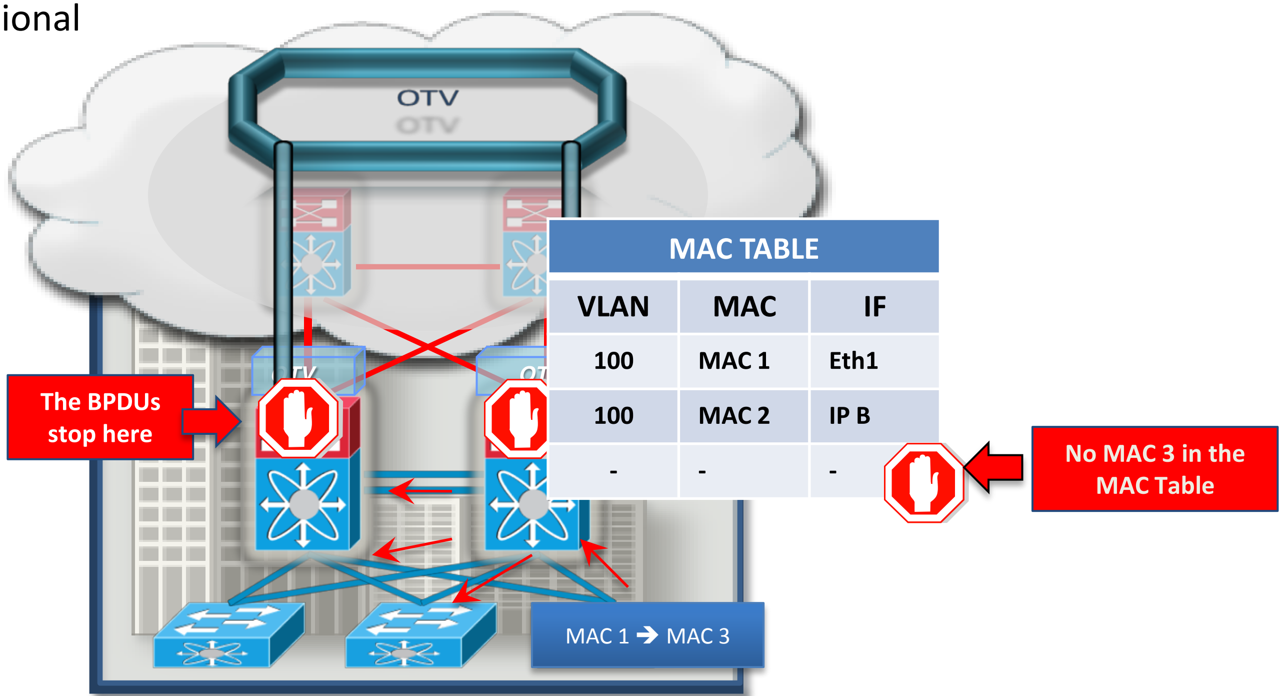
```
feature otv
otv site-vlan 99
interface Overlay1
  description WEST-DC
  otv join-interface e1/1
  otv control-group 239.1.1.1
  otv data-group 232.192.1.0/24
  otv extend-vlan 100-150
```

```
feature otv
otv site-vlan 99
interface Overlay1
  description EAST-DC
  otv join-interface e1/1.10
  otv control-group 239.1.1.1
  otv data-group 232.192.1.0/24
  otv extend-vlan 100-150
```

```
feature otv
otv site-vlan 99
interface Overlay1
  description SOUTH-DC
  otv join-interface Po16
  otv control-group 239.1.1.1
  otv data-group 232.192.1.0/24
  otv extend-vlan 100-150
```

OTV - Failure Isolation

- **Spanning-tree**
  - OTV is site transparent: no changes to the STP topology, each site keeps its own STP domain
  - This functionality is built-in into OTV and no additional configuration is required
  - An Edge Device will send and receive BPDUs ONLY on the OTV Internal Interfaces
- **Unknown Unicast**
  - No requirements to forward unknown unicast frames
  - OTV does not forward unknown unicast frames to the overlay. This is achieved without any additional configuration
  - The assumption here is that the end-points connected to the network are not silent or uni-directional



MAC TABLE

VLAN	MAC	IF
100	MAC 1	Eth1
100	MAC 2	IP B
-	-	-

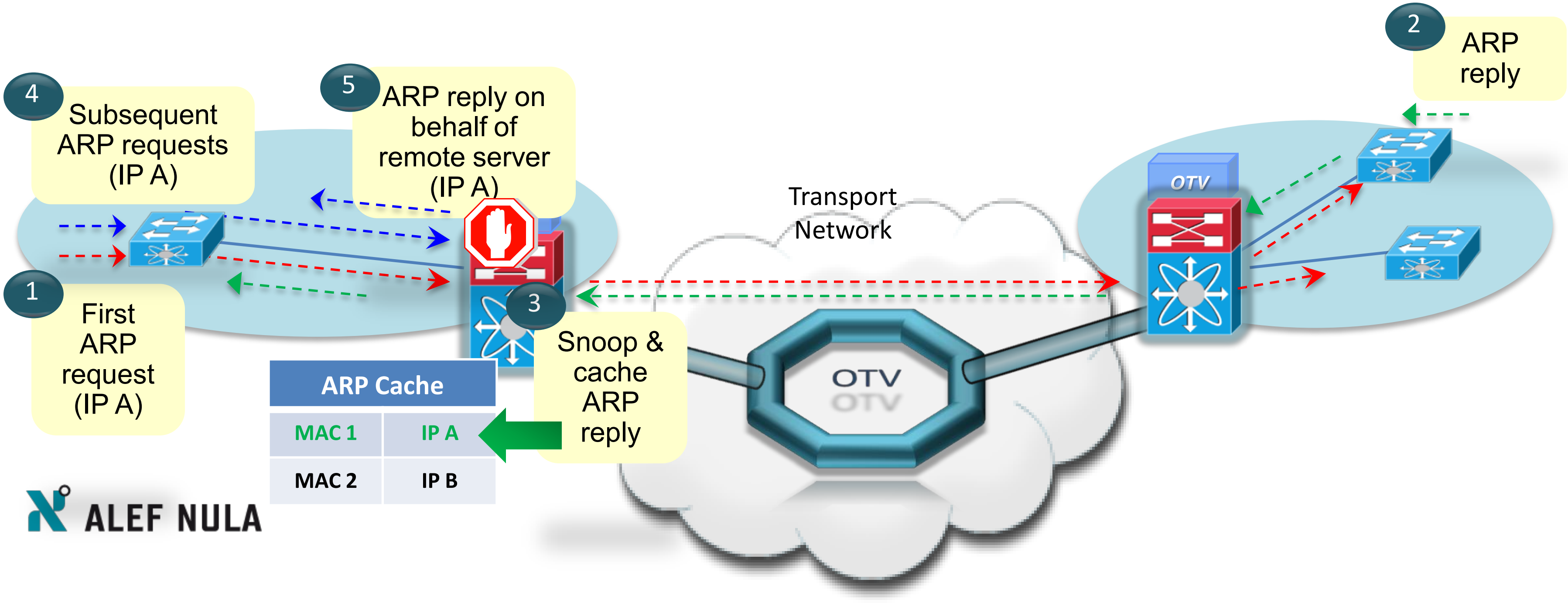
The BPDUs stop here

No MAC 3 in the MAC Table

MAC 1 → MAC 3

Controlling ARP traffic
ARP Neighbor-Discovery (ND) Cache

- An ARP cache is maintained by every OTV edge device and is populated by snooping ARP replies
- Initial ARP requests are broadcasted to all sites, but subsequent ARP requests are suppressed at the Edge Device and answered locally
- ARP traffic spanning multiple sites can thus be significantly reduced



1 First ARP request (IP A)

2 ARP reply

3 Snoop & cache ARP reply

4 Subsequent ARP requests (IP A)

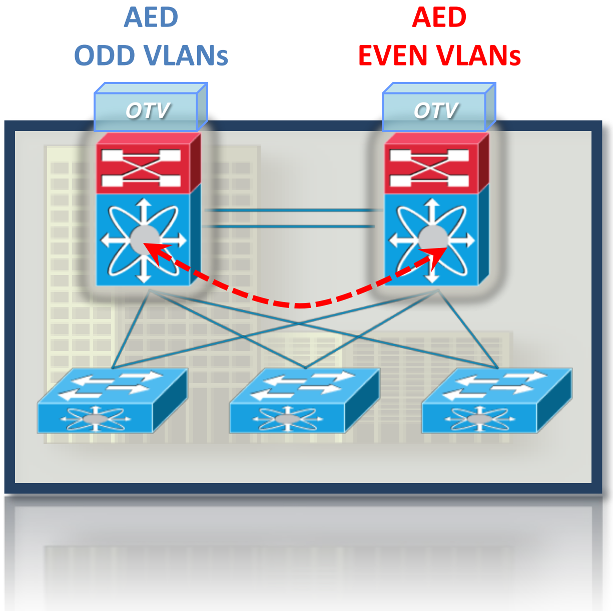
5 ARP reply on behalf of remote server (IP A)

ARP Cache	
MAC 1	IP A
MAC 2	IP B

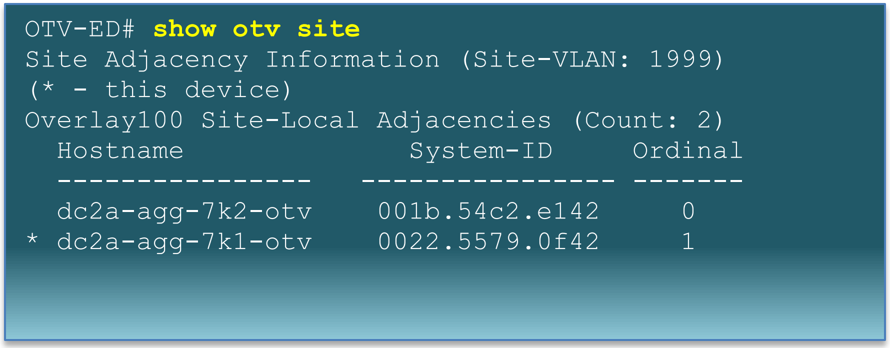
ALEF NULA

OTV Automated Multi-homing

- Fully automated and it does not require additional protocols and configuration
- The Edge Devices within a site discover each other over the “otv site-vlan”
- Authoritative Edge Device (AED)
  - MAC addresses advertisement for its VLANs
  - Forwarding its VLANs’ traffic inside and outside the site
  - Achieved via a very deterministic algorithm (not configurable, even & odd vlans)



Internal peering for AED election



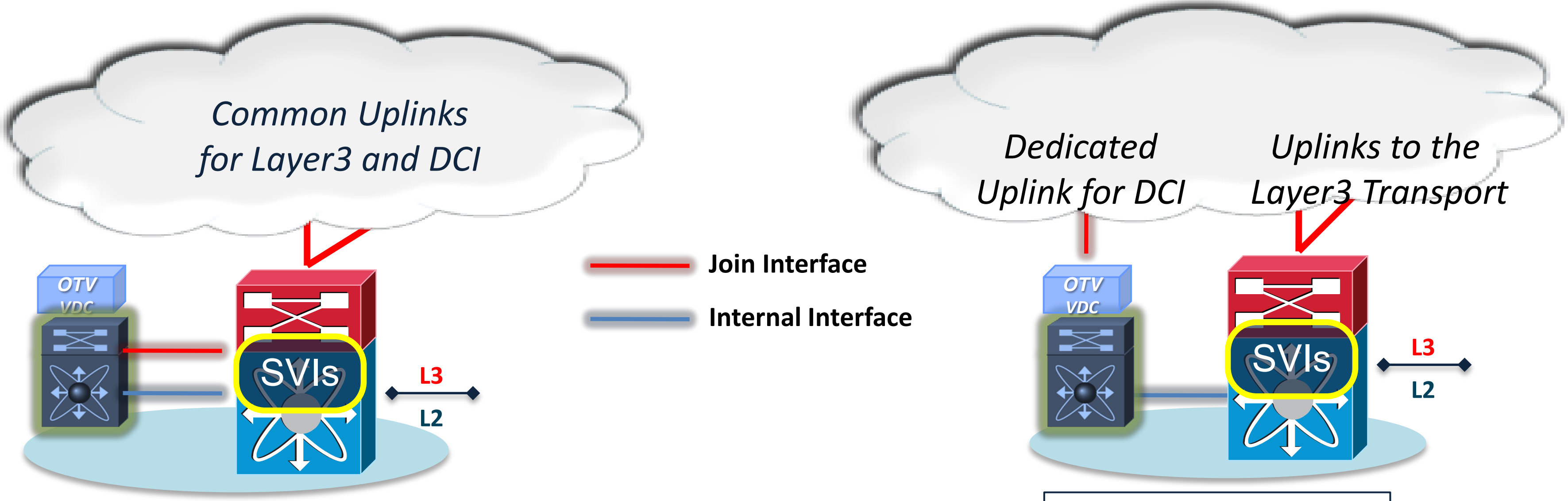
```
OTV-ED# show otv site
Site Adjacency Information (Site-VLAN: 1999)
(* - this device)
Overlay100 Site-Local Adjacencies (Count: 2)
   Hostname                 System-ID      Ordinal
   ----------------    ----------------  -------
   dc2a-agg-7k2-otv     001b.54c2.e142        0
*  dc2a-agg-7k1-otv     0022.5579.0f42        1
```

OTV - Broadcast & Multicast Traffic

- **Broadcast**
  - Broadcast frames are sent to all remote OTV edge devices by leveraging the same ASM multicast group in the transport already used for the OTV control protocol. (handled exactly the same way as the OTV Hello messages)
- **Multicast**
  - The site multicast groups are mapped to a SSM group range in the core
  - Source ED communicates the mapping information (including the source VLAN) to the other Eds
  - Receiver ED joins SSM group
  - The source ED adds the Overlay interface to the *Outbound Interface List* (OIL).
  - The right number of SSM groups to be used depends on a tradeoff between the amount of multicast state to be maintained in the core and the optimization of Layer 2 multicast traffic delivery

OTV and SVI Separation

- Guideline: The current OTV implementation on the Nexus 7000 enforces the separation between SVI routing and OTV encapsulation for a given VLAN
  - Can be achieved with having two separate devices to perform these two functions
  - An alternative, cleaner and less intrusive solution is the use of Virtual Device Contexts (VDCs) available with Nexus 7000 platform:
    - A dedicated OTV VDC to perform the OTV functionalities
    - The Aggregation-VDC used to provide SVI routing support

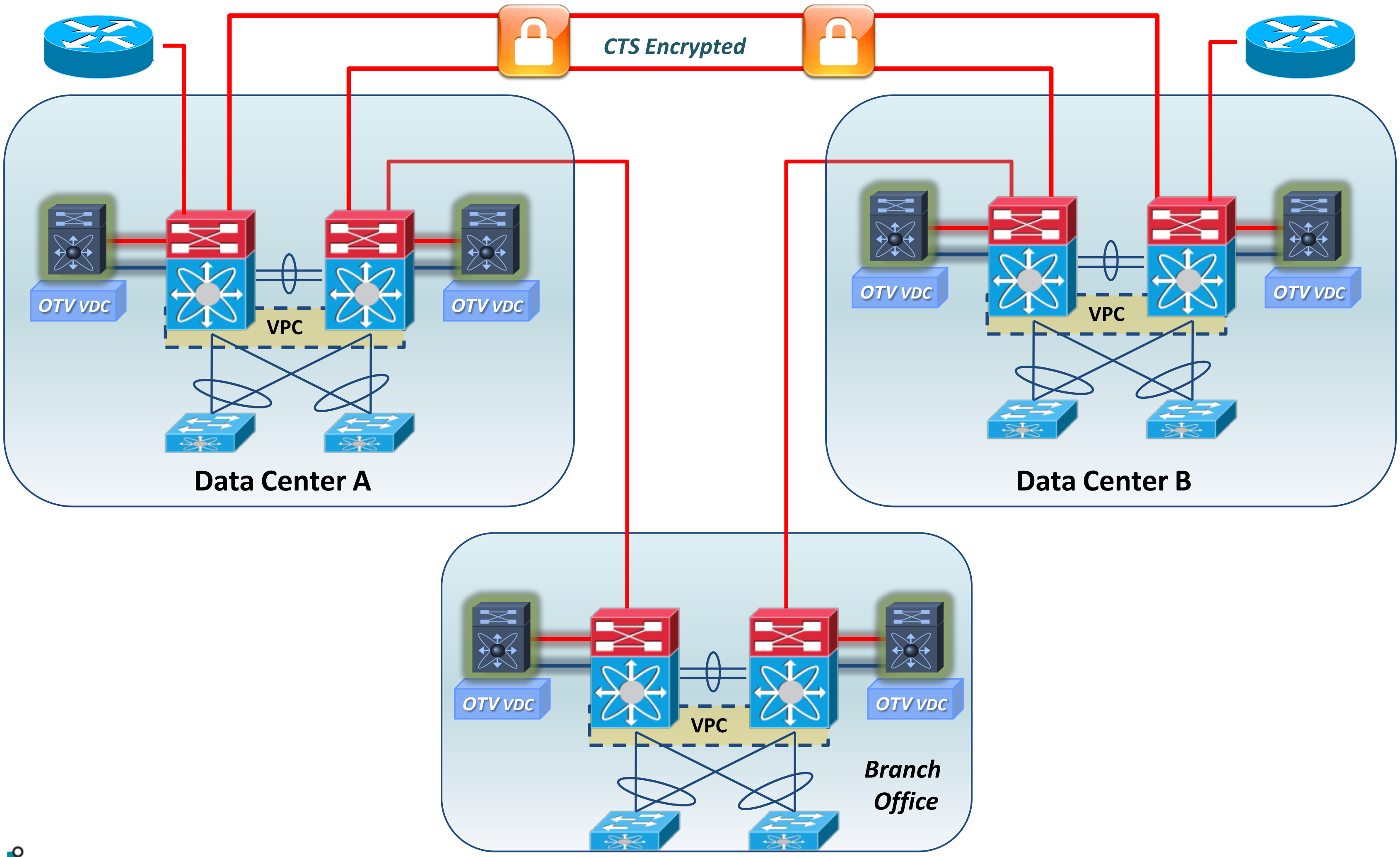


Common Uplinks for Layer3 and DCI
OTV VDC
SVIs
L3
L2
Join Interface
Internal Interface
Dedicated Uplink for DCI
Uplinks to the Layer3 Transport
OTV VDC
SVIs
L3
L2

OTV Appliance on a Stick

Inline OTV Appliance

OTV Design – Collapsed Core



ALEF NULA

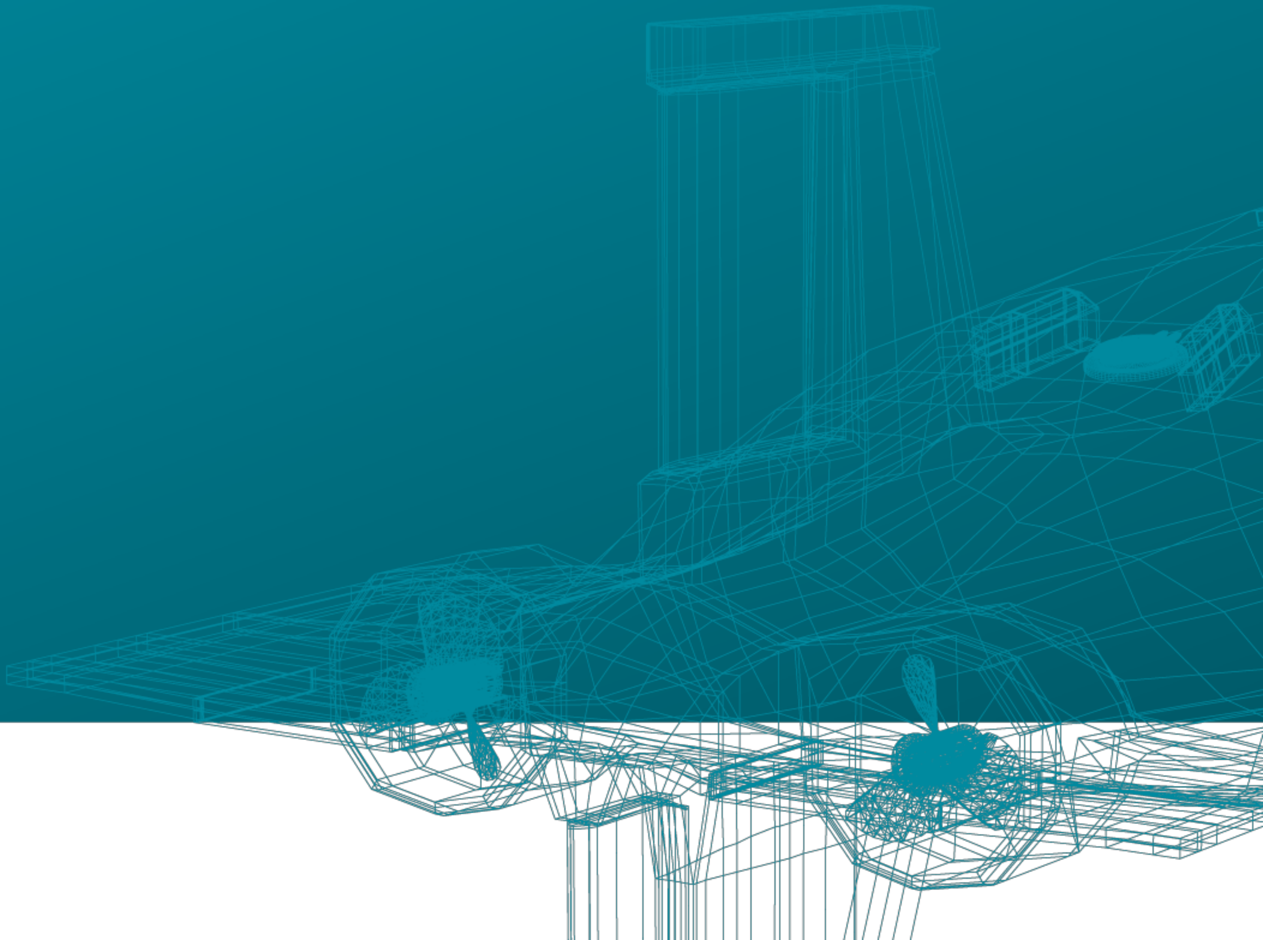
OTV Current Limits

Feature	Maximum Limits
Number of OTV overlays	3
Number of OTV-connected sites	3
Number of edge devices in all sites	6
Number of edge devices per site	2
Number of VLANs per overlay	128
Number of OTV-extended VLANs across all configured overlays	128
Number of MAC Addresses across all the extended VLANs in all configured overlays	12000
Number of multicast data groups	1000
Number of multicast data groups per site	100

OTV – Documentation, links

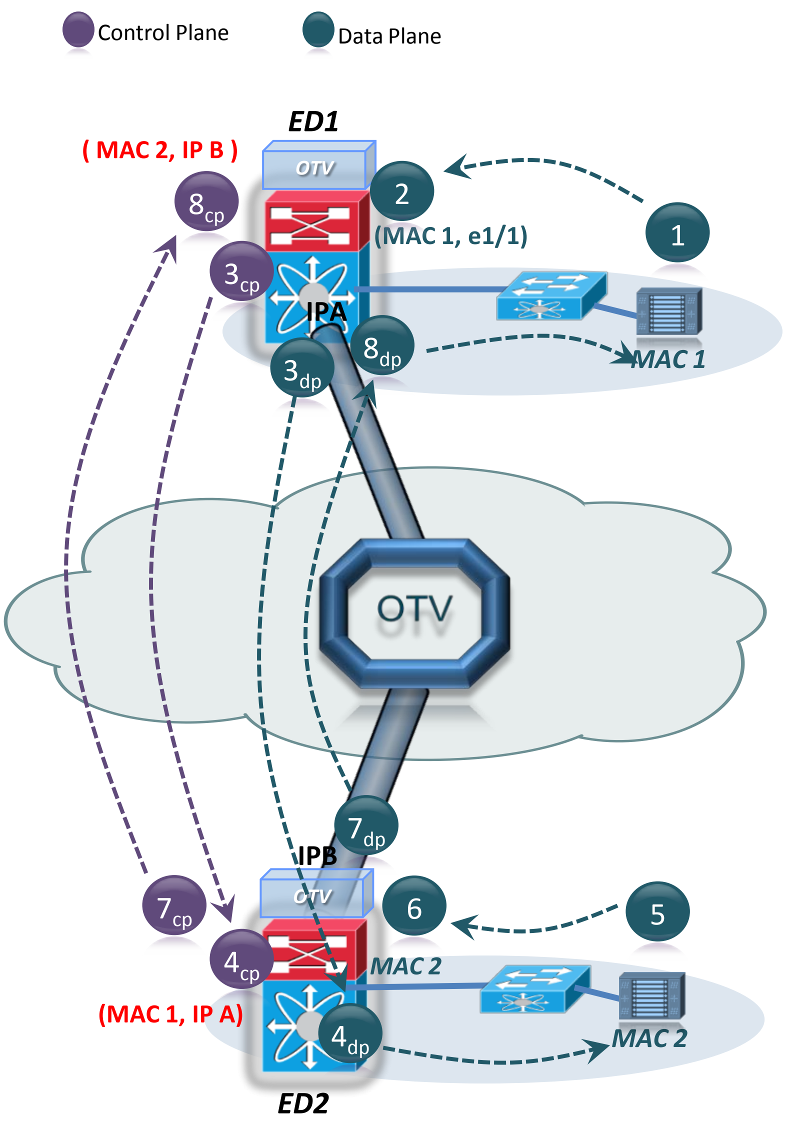
- **DCI page**
  - http://www.cisco.com/en/US/netsol/ns975/index.html
- **OTV whitepaper**
  - http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/whitepaper/DCI3_OTV_Intro.html
- **RFC Draft**
  - http://tools.ietf.org/html/draft-hasmit-otv-01
- **Cisco Overlay Transport Virtualization Technology Introduction and Deployment Considerations**
  - http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/whitepaper/DCI3_OTV_Intro.html
- **Cisco Nexus 7000 Series NX-OS OTV Configuration Guide**
  - http://www.cisco.com/en/US/docs/switches/datacenter/sw/5_x/nx-os/otv/configuration/guide/b_Cisco_Nexus_7000_Series_NX-OS_OTV_Configuration_Guide__Release_5.x.html
- **Nexus 7000 - OTV - Design and Configuration Example**
  - http://docwiki.cisco.com/wiki/Nexus_7000_-_OTV_-_Design_and_Configuration_Example
- **OTV and Long Distance Live Vmotion Whitepapers**
  - Cisco, VMware and Netapp: http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-591960.pdf
  - Cisco, VMware and EMC: http://media.vceportal.com/documents/WhitePaper_Application_Mobility.pdf

Děkuji za pozornost



ALEF NULA

A Packet Walk is Worth a Million Words
Establishing OTV Unicast Communication



1 – Server 1 sends a broadcast ARP for MAC 2

2 – ARP broadcast is received by ED1, which learns MAC 1 on its internal interface

3_{cp} – ED1 advertises MAC 1 in an OTV Update sent via the multicast control group

4_{cp} – ED2 receives the update and stores MAC1 in MAC table, next-hop is ED1

3_{dp} – ED1 encapsulates broadcast in the core IP multicast group so all the EDs in the overlay receive it

4_{dp} – ED2 decapsulates the frame and forwards the ARP broadcast request into the site

5 – Server 2 receives the ARP and replies with a unicast ARP reply to MAC 1

6 – ED2 learns MAC 2 on its internal interface

7_{cp} – ED2 advertises MAC 2 in IS-IS LSP sent via the multicast control group

8_{cp} – ED1 receives the update and stores MAC2 in MAC table, next-hop is ED2

7_{dp} – ED2 knows that MAC 1 is reachable via IP A so encapsulates the packet and sends it unicast to ED1's IP address (IP A)

8_{dp} – Core delivers packet to ED1, ED1 decapsulates and forwards it into the site to MAC 1

OTV Control Plane
CLI Verification

- Establishment of control plane adjacencies between OTV Edge Devices:

```
dc1-agg-7k1# show otv adjacency

Overlay Adjacency database
Overlay-Interface Overlay100   :
Hostname         System-ID        Dest Addr     Up Time
Adj-State
dc2-agg-7k1    001b.54c2.efc2   20.11.23.2    15:08:53    UP
dc1-agg-7k2    001b.54c2.e1c3   20.12.23.2    15:43:27    UP
dc2-agg-7k2    001b.54c2.e142   20.22.23.2    14:49:11    UP
```

- Unicast MAC reachability information:

```
dc1-agg-7k1# show otv route
OTV Unicast MAC Routing Table For Overlay100
VLAN MAC-Address     Metric  Uptime    Owner      Next-hop(s)
---- -------------- ------  --------- ---------  -----------
2001 0000.0c07.ac01  1       3d15h     site       Ethernet1/1
2001 0000.1641.d70e  1       3d15h     site       Ethernet1/2
2001 0000.49f3.88ff  42      2d22h     overlay    dc2-agg-7k1
2001 0000.49f3.8900  42      2d22h     overlay    dc2-agg-7k2
```