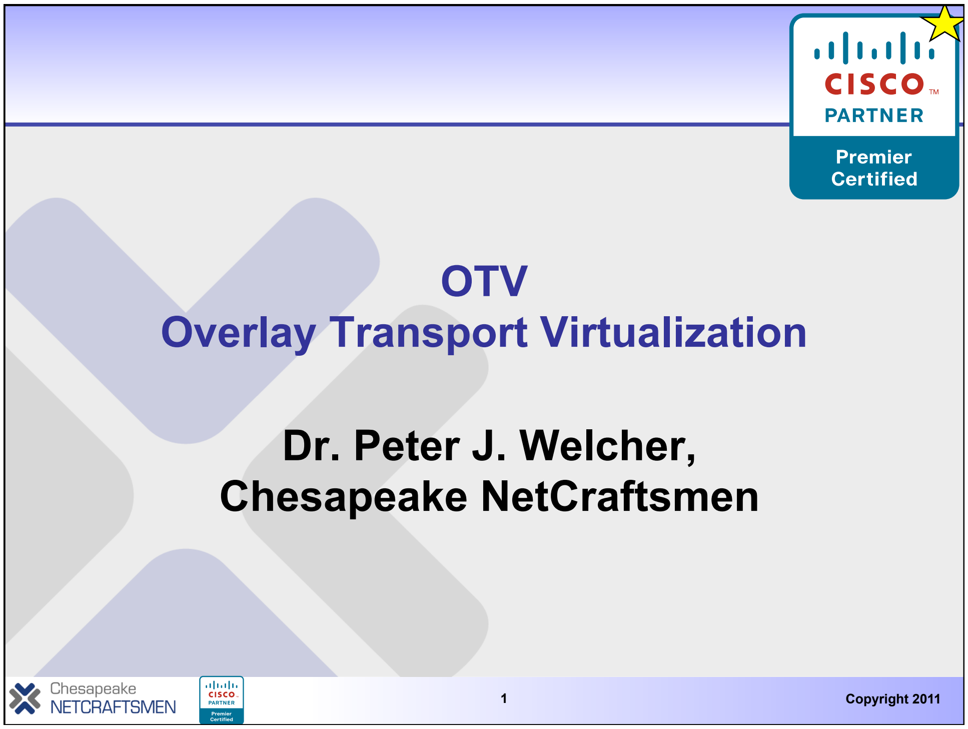
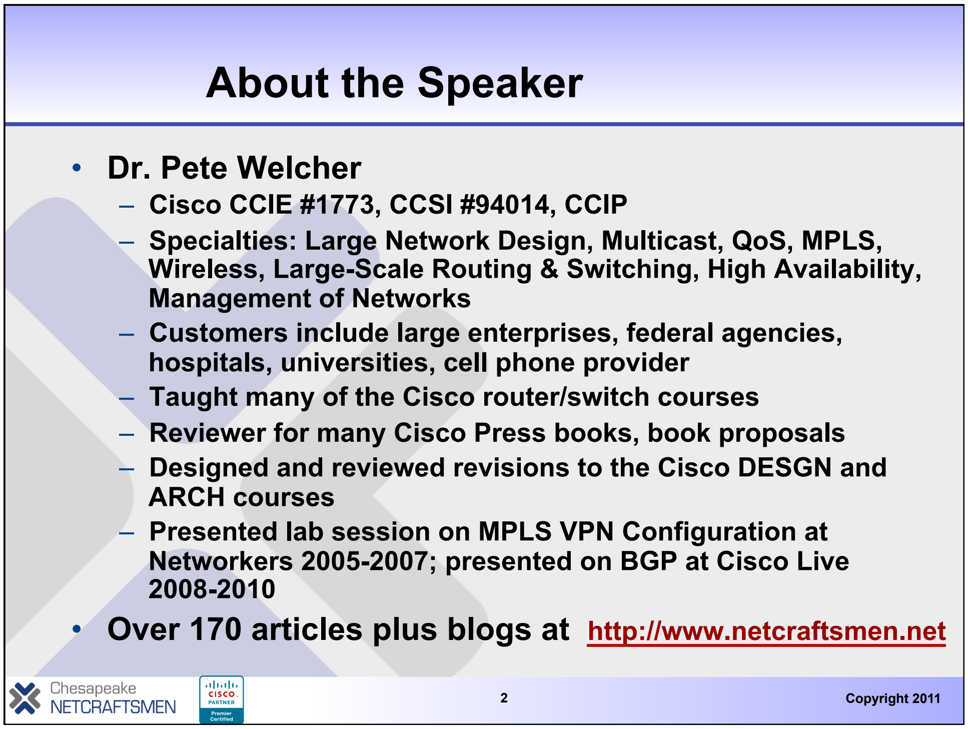


OTV
Overlay Transport Virtualization

Dr. Peter J. Welcher,
Chesapeake NetCraftsmen

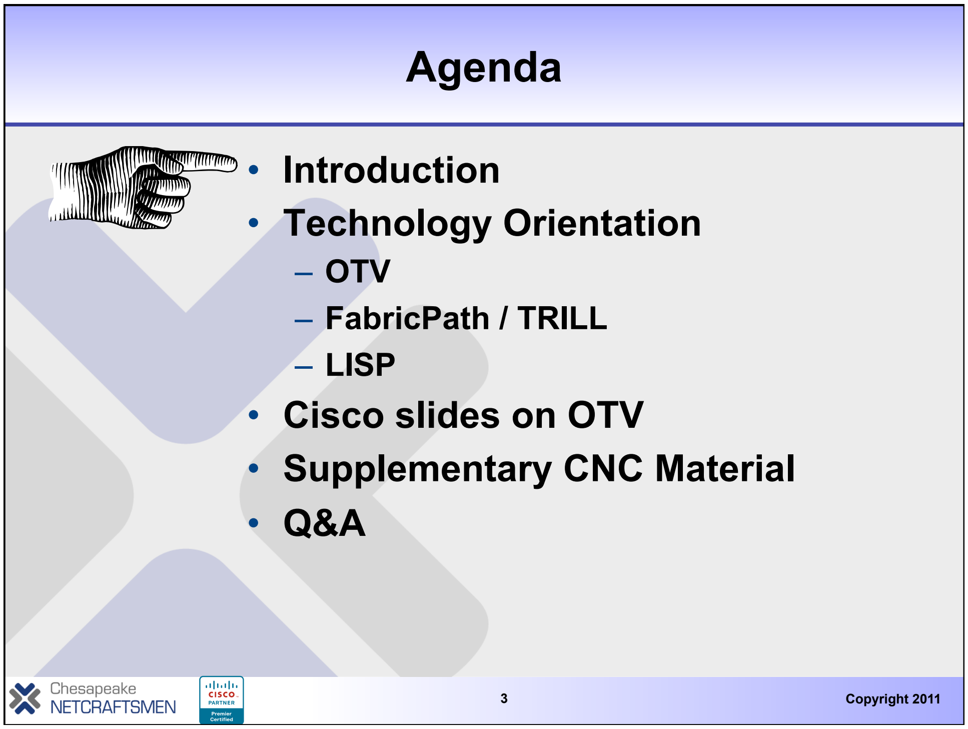
Chesapeake NETCRAFTSMEN · CISCO PARTNER Premier Certified · 1 · Copyright 2011



About the Speaker

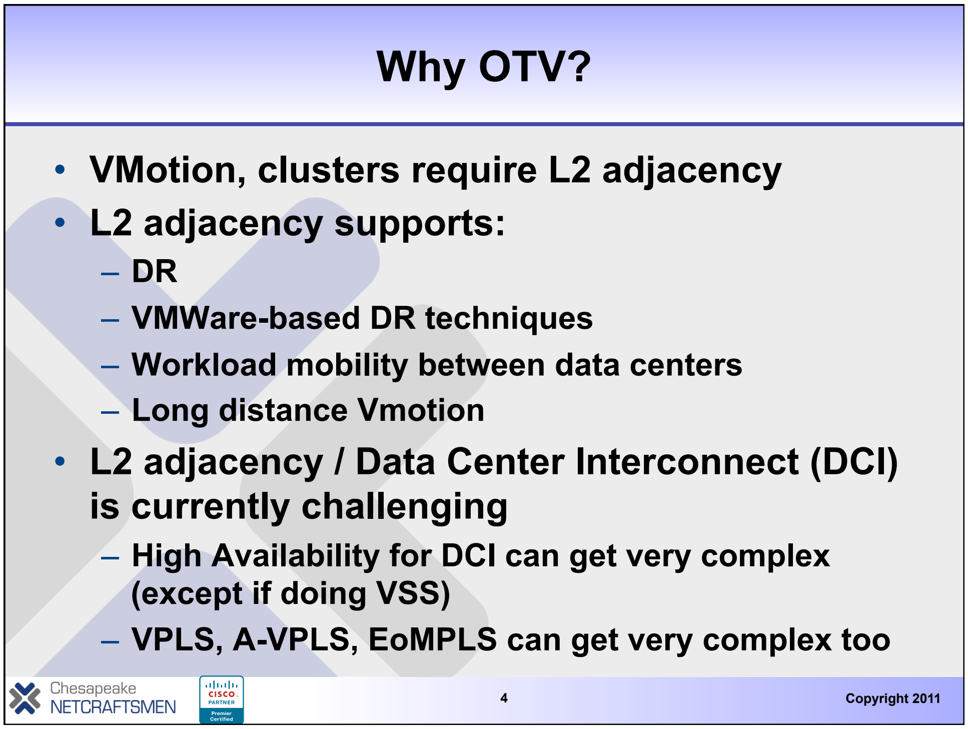
- Dr. Pete Welcher
  - Cisco CCIE #1773, CCSI #94014, CCIP
  - Specialties: Large Network Design, Multicast, QoS, MPLS, Wireless, Large-Scale Routing & Switching, High Availability, Management of Networks
  - Customers include large enterprises, federal agencies, hospitals, universities, cell phone provider
  - Taught many of the Cisco router/switch courses
  - Reviewer for many Cisco Press books, book proposals
  - Designed and reviewed revisions to the Cisco DESGN and ARCH courses
  - Presented lab session on MPLS VPN Configuration at Networkers 2005-2007; presented on BGP at Cisco Live 2008-2010
- Over 170 articles plus blogs at http://www.netcraftsmen.net

Chesapeake NETCRAFTSMEN · CISCO PARTNER Premier Certified · 2 · Copyright 2011



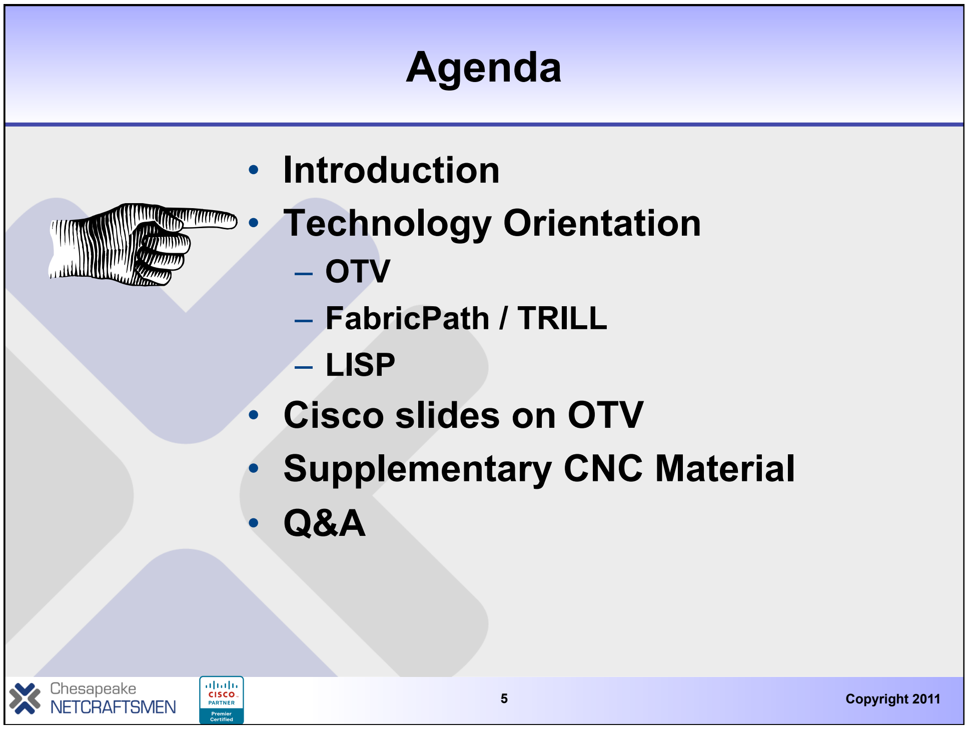
Agenda

- Introduction
- Technology Orientation
  - OTV
  - FabricPath / TRILL
  - LISP
- Cisco slides on OTV
- Supplementary CNC Material
- Q&A



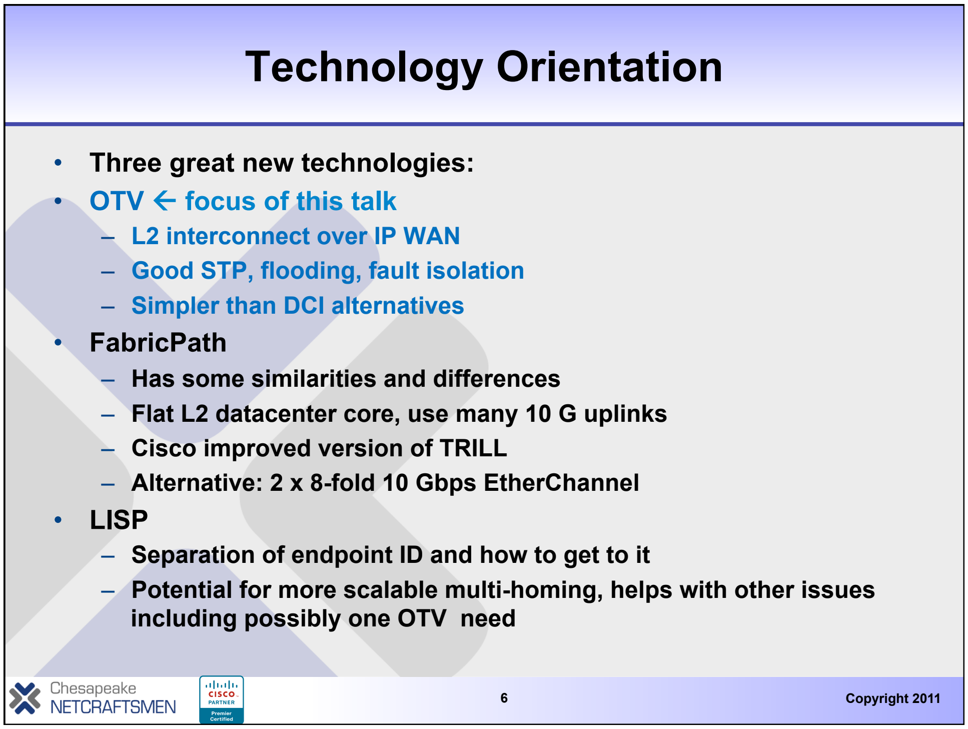
Why OTV?

- VMotion, clusters require L2 adjacency
- L2 adjacency supports:
  - DR
  - VMWare-based DR techniques
  - Workload mobility between data centers
  - Long distance Vmotion
- L2 adjacency / Data Center Interconnect (DCI) is currently challenging
  - High Availability for DCI can get very complex (except if doing VSS)
  - VPLS, A-VPLS, EoMPLS can get very complex too



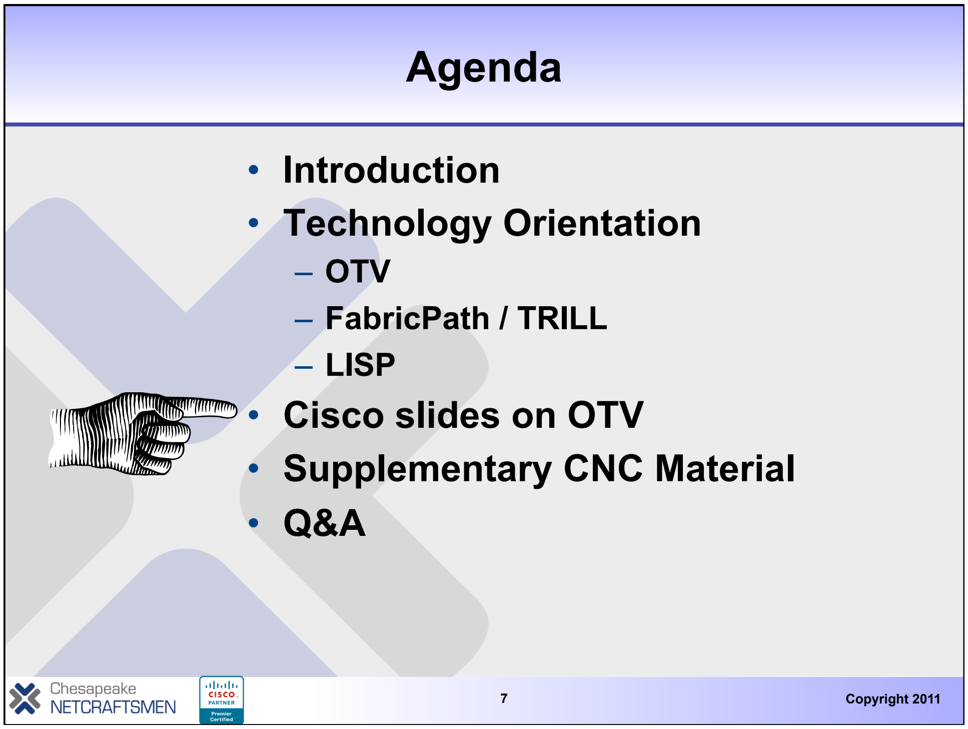
Agenda

- Introduction
- Technology Orientation
  - OTV
  - FabricPath / TRILL
  - LISP
- Cisco slides on OTV
- Supplementary CNC Material
- Q&A



Technology Orientation

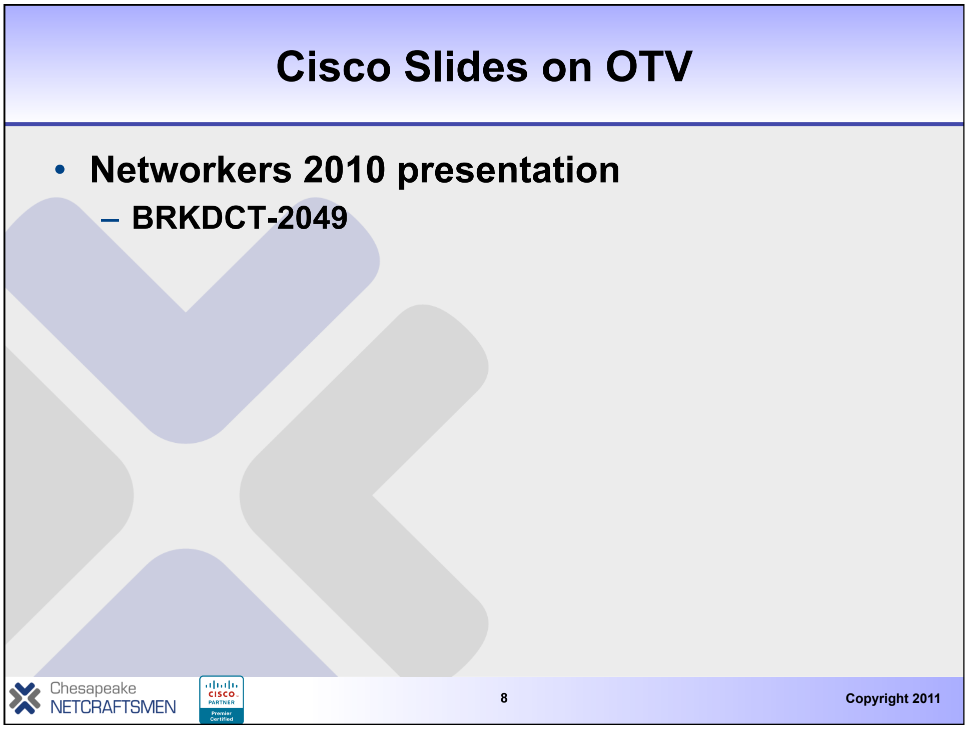
- Three great new technologies:
- OTV ← focus of this talk
  - L2 interconnect over IP WAN
  - Good STP, flooding, fault isolation
  - Simpler than DCI alternatives
- FabricPath
  - Has some similarities and differences
  - Flat L2 datacenter core, use many 10 G uplinks
  - Cisco improved version of TRILL
  - Alternative: 2 x 8-fold 10 Gbps EtherChannel
- LISP
  - Separation of endpoint ID and how to get to it
  - Potential for more scalable multi-homing, helps with other issues including possibly one OTV need



Agenda

- Introduction
- Technology Orientation
  - OTV
  - FabricPath / TRILL
  - LISP
- Cisco slides on OTV
- Supplementary CNC Material
- Q&A

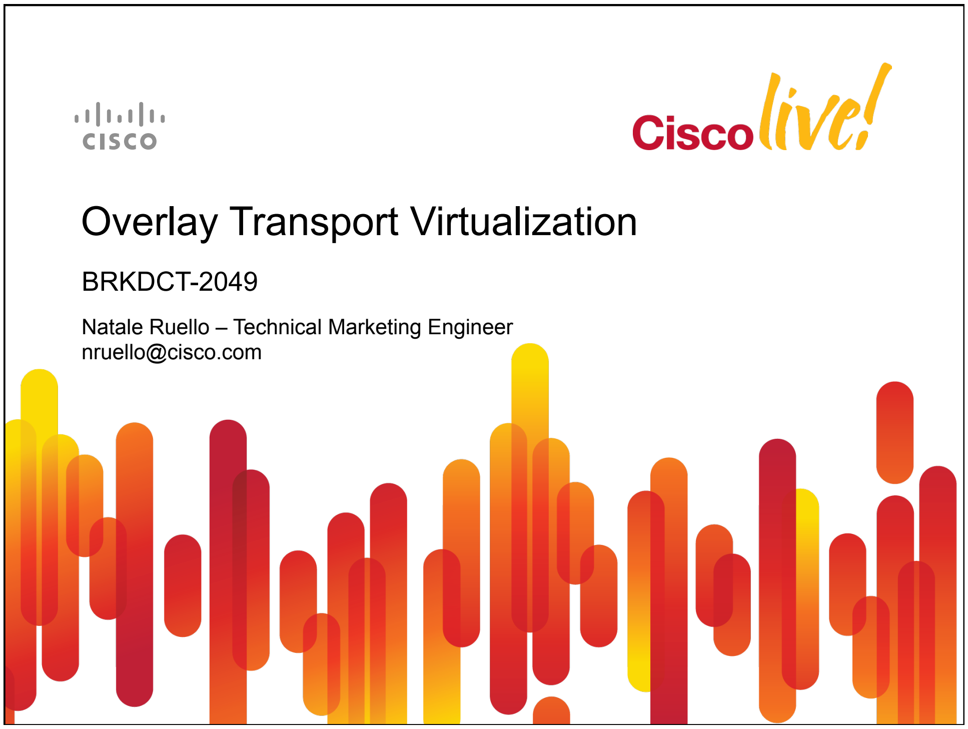
Chesapeake NETCRAFTSMEN 7 Copyright 2011



Cisco Slides on OTV

- Networkers 2010 presentation
  - BRKDCT-2049

Chesapeake NETCRAFTSMEN 8 Copyright 2011

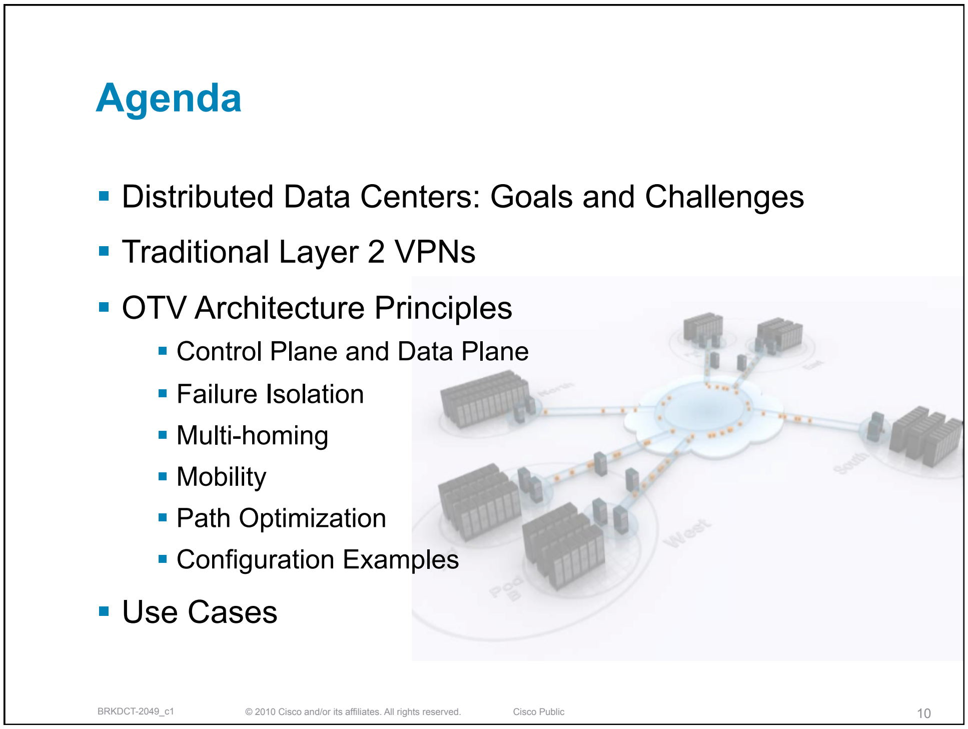


Cisco live!

Overlay Transport Virtualization

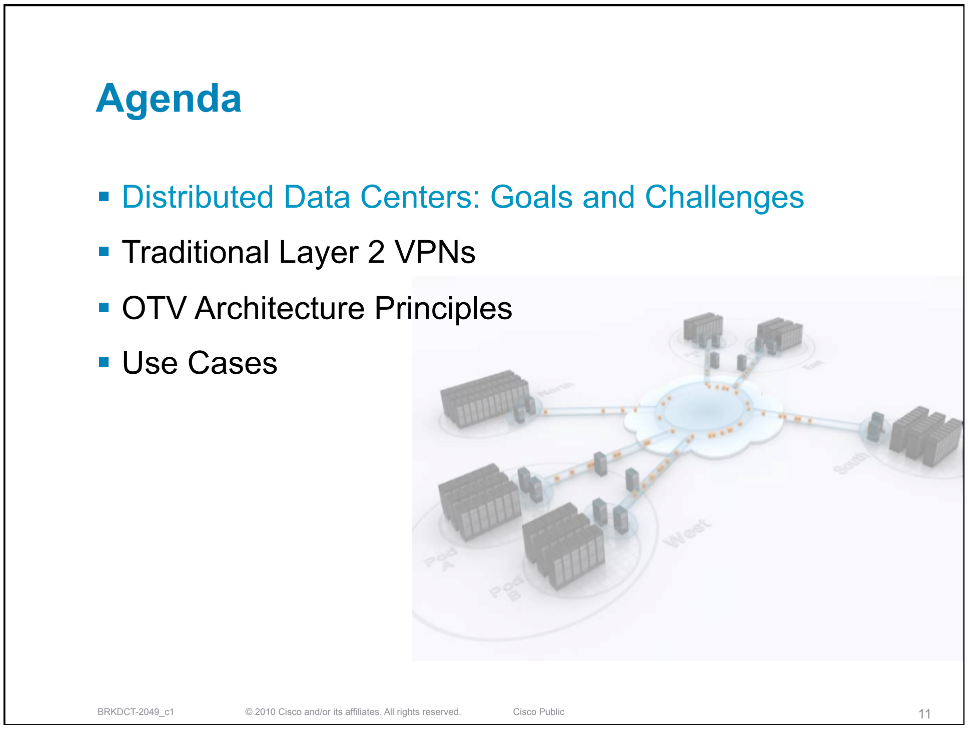
BRKDCT-2049

Natale Ruello – Technical Marketing Engineer
nruello@cisco.com



Agenda

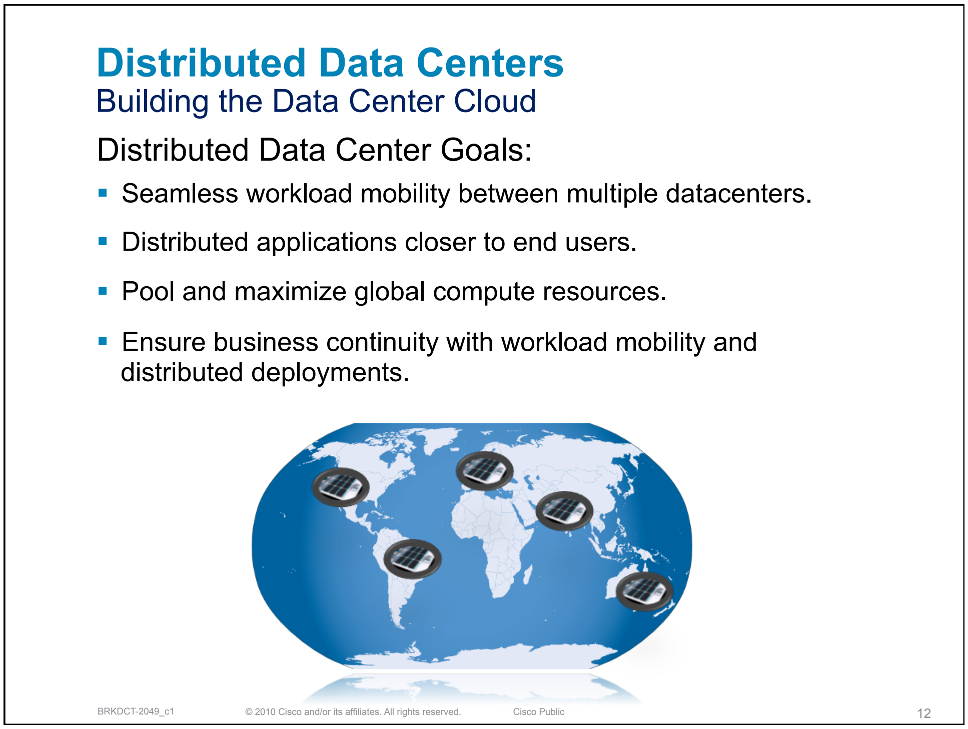
- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
  - Control Plane and Data Plane
  - Failure Isolation
  - Multi-homing
  - Mobility
  - Path Optimization
  - Configuration Examples
- Use Cases



Agenda

- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
- Use Cases

North
East
South
West
Pod A
Pod B

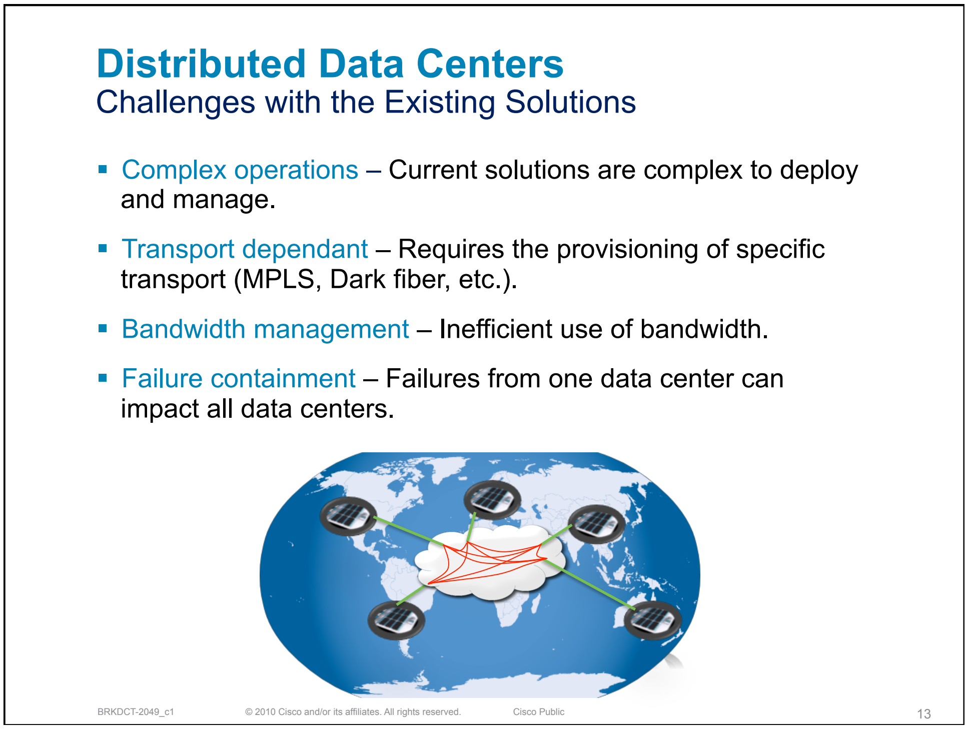


Distributed Data Centers

Building the Data Center Cloud

Distributed Data Center Goals:

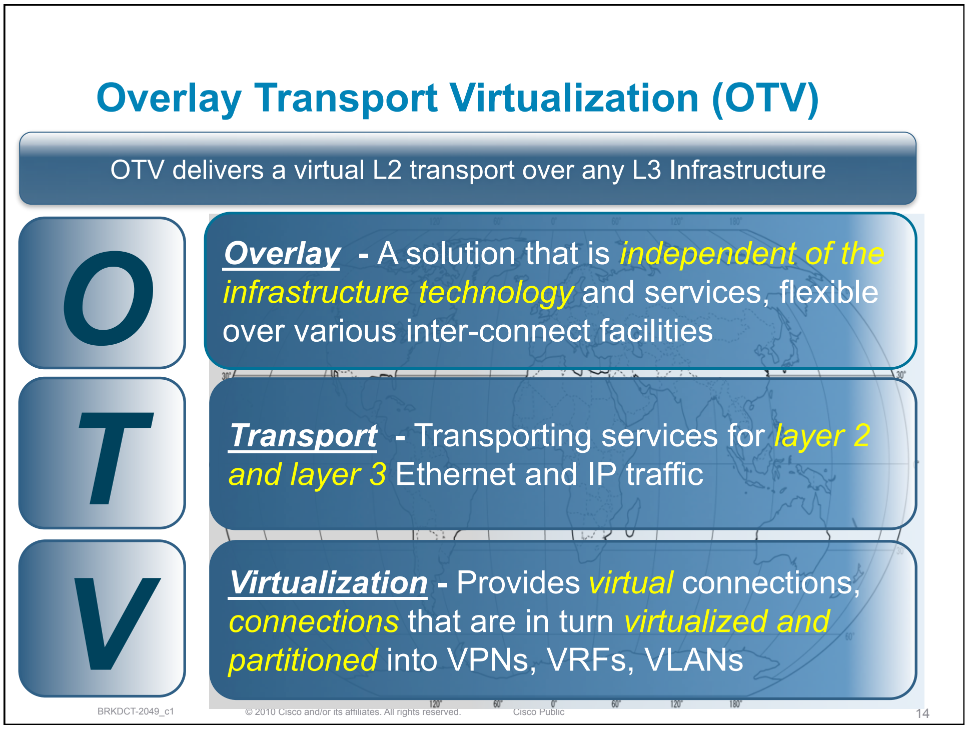
- Seamless workload mobility between multiple datacenters.
- Distributed applications closer to end users.
- Pool and maximize global compute resources.
- Ensure business continuity with workload mobility and distributed deployments.



Distributed Data Centers

Challenges with the Existing Solutions

- Complex operations – Current solutions are complex to deploy and manage.
- Transport dependant – Requires the provisioning of specific transport (MPLS, Dark fiber, etc.).
- Bandwidth management – Inefficient use of bandwidth.
- Failure containment – Failures from one data center can impact all data centers.



Overlay Transport Virtualization (OTV)

OTV delivers a virtual L2 transport over any L3 Infrastructure

O

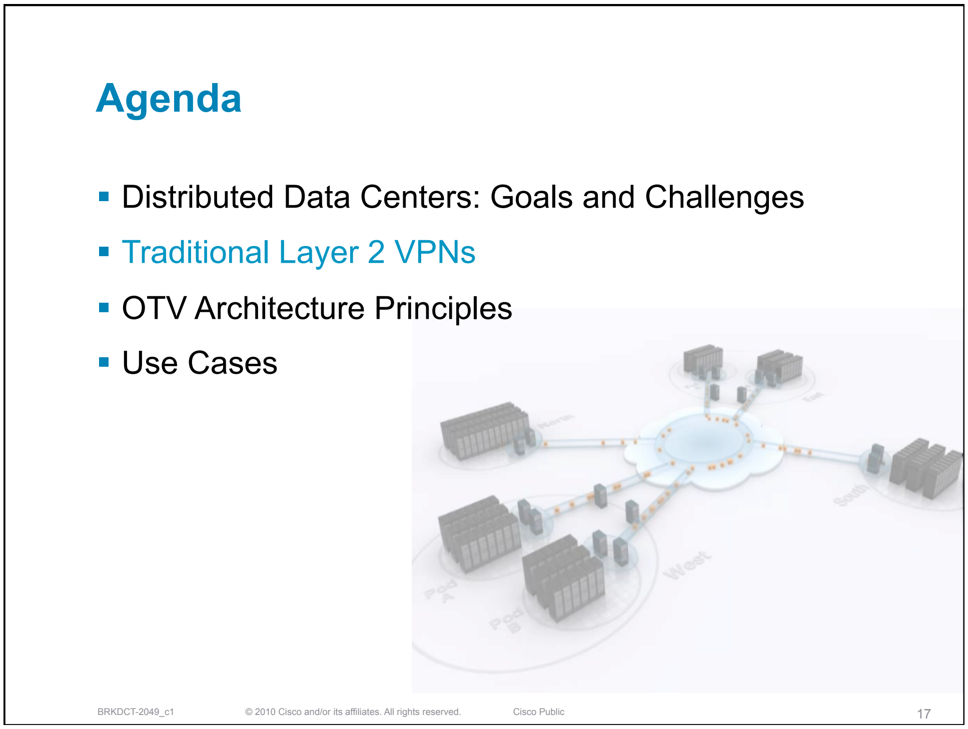
Overlay **-** A solution that is *independent of the infrastructure technology* and services, flexible over various inter-connect facilities

T

Transport **-** Transporting services for *layer 2 and layer 3* Ethernet and IP traffic

V

Virtualization **-** Provides *virtual* connections, *connections* that are in turn *virtualized and partitioned* into VPNs, VRFs, VLANs

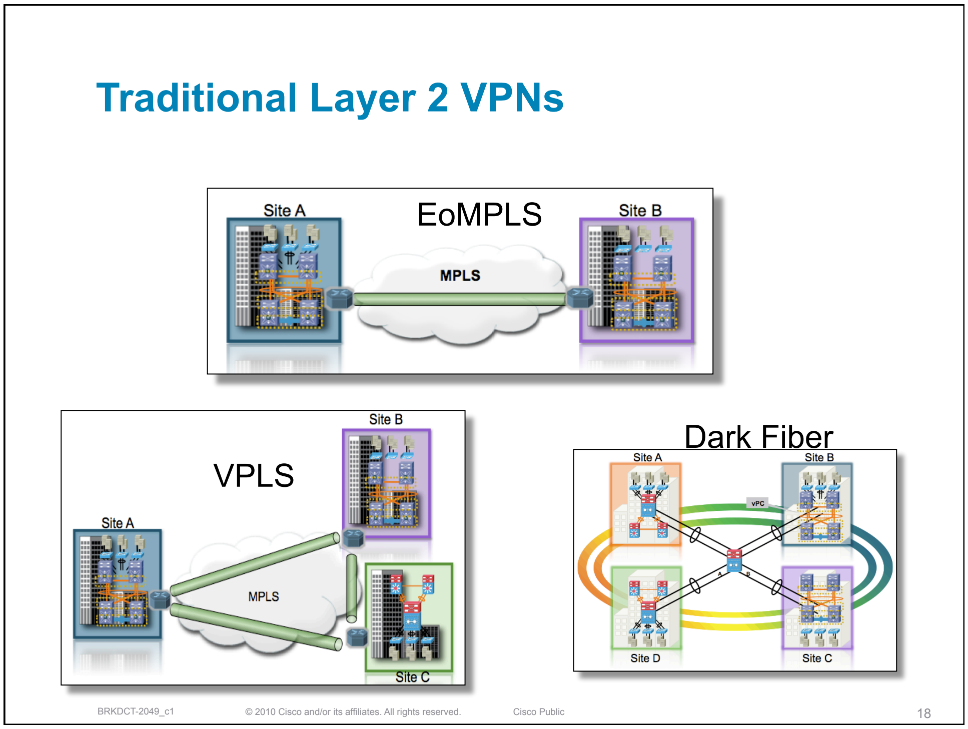


Agenda

- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
- Use Cases

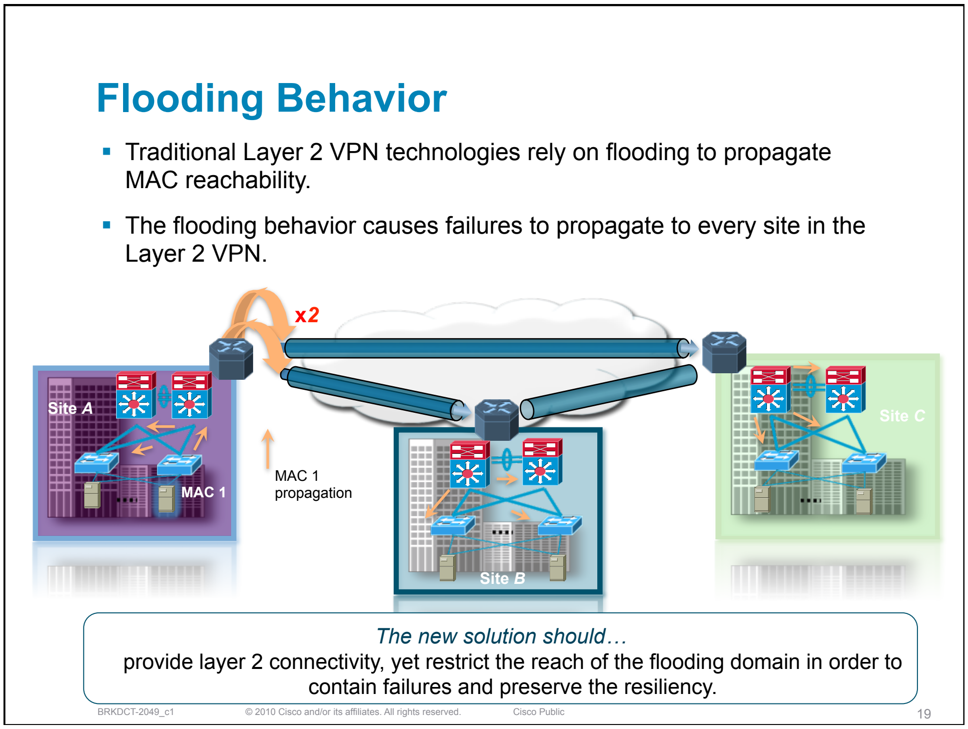
North
East
South
West
Pod A
Pod B

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 17



Traditional Layer 2 VPNs

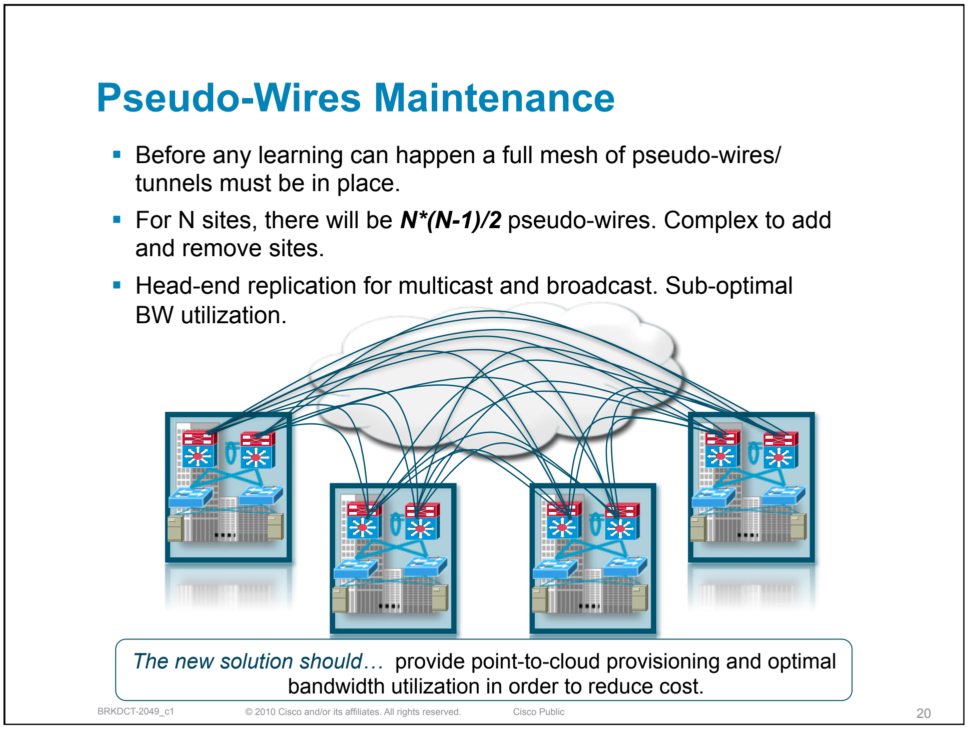
BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 18



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.

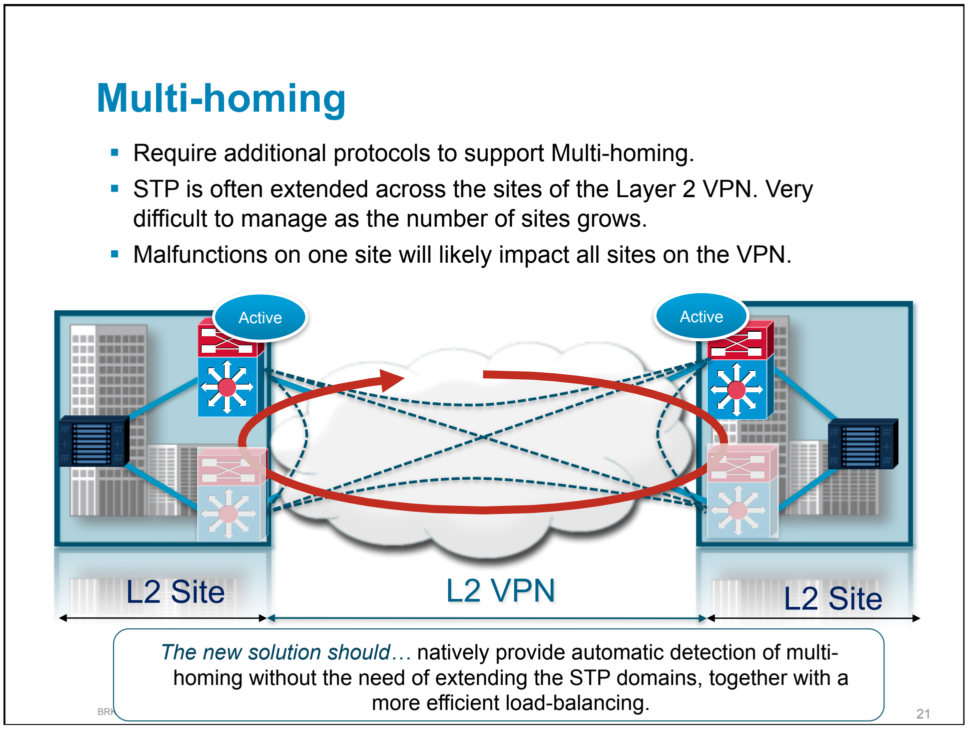
The new solution should…
provide layer 2 connectivity, yet restrict the reach of the flooding domain in order to contain failures and preserve the resiliency.



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.

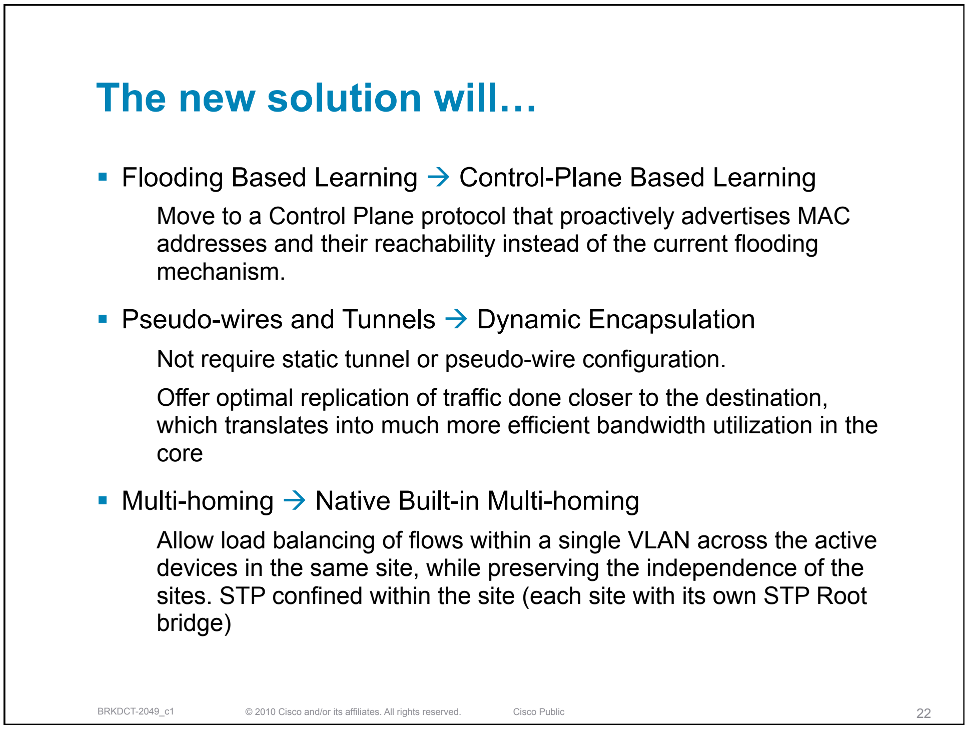
The new solution should… provide point-to-cloud provisioning and optimal bandwidth utilization in order to reduce cost.



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.

The new solution should… natively provide automatic detection of multi-homing without the need of extending the STP domains, together with a more efficient load-balancing.



The new solution will…

- 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.

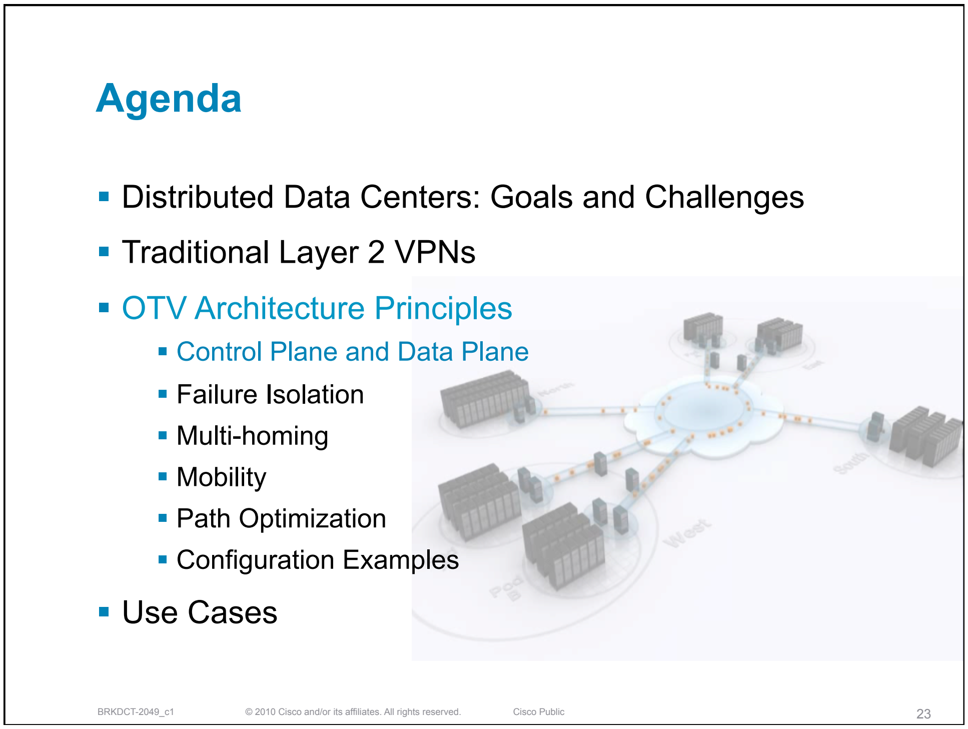
- Pseudo-wires and Tunnels → Dynamic Encapsulation

  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

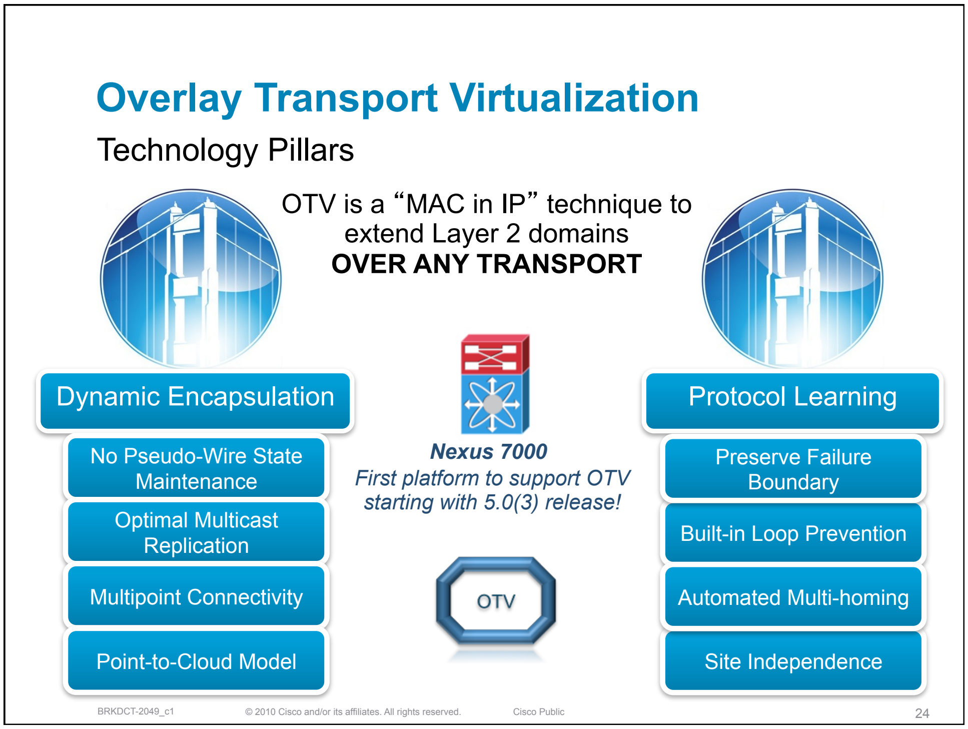
- Multi-homing → Native Built-in Multi-homing

  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)



Agenda

- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
  - Control Plane and Data Plane
  - Failure Isolation
  - Multi-homing
  - Mobility
  - Path Optimization
  - Configuration Examples
- Use Cases



Overlay Transport Virtualization

Technology Pillars

OTV is a “MAC in IP” technique to extend Layer 2 domains
OVER ANY TRANSPORT

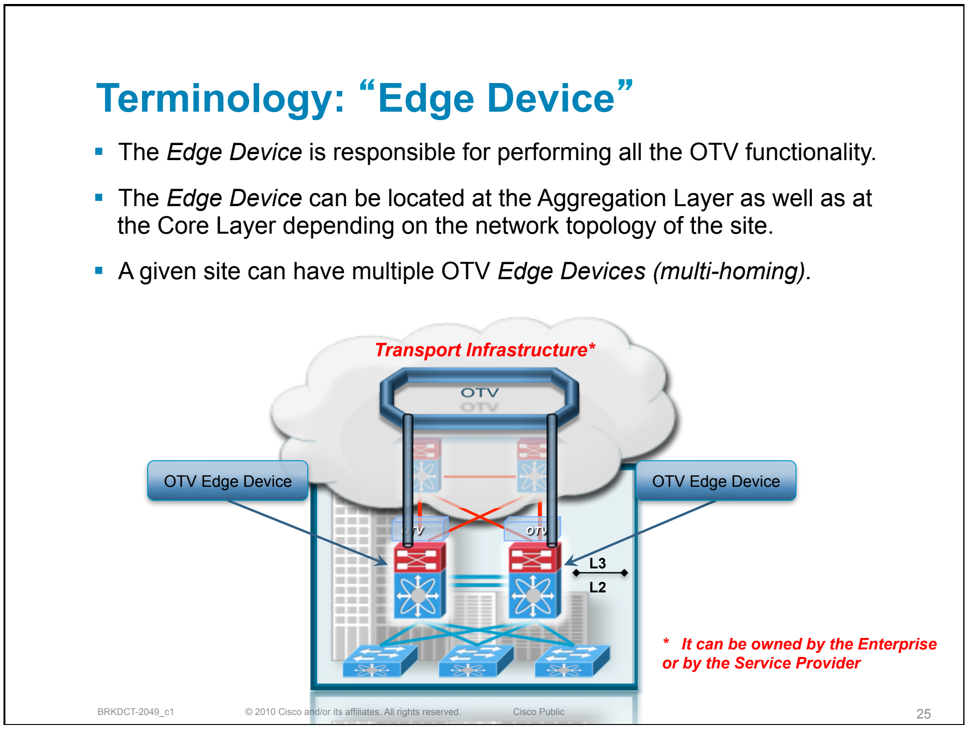
Dynamic Encapsulation

- No Pseudo-Wire State Maintenance
- Optimal Multicast Replication
- Multipoint Connectivity
- Point-to-Cloud Model

Nexus 7000
First platform to support OTV starting with 5.0(3) release!

Protocol Learning

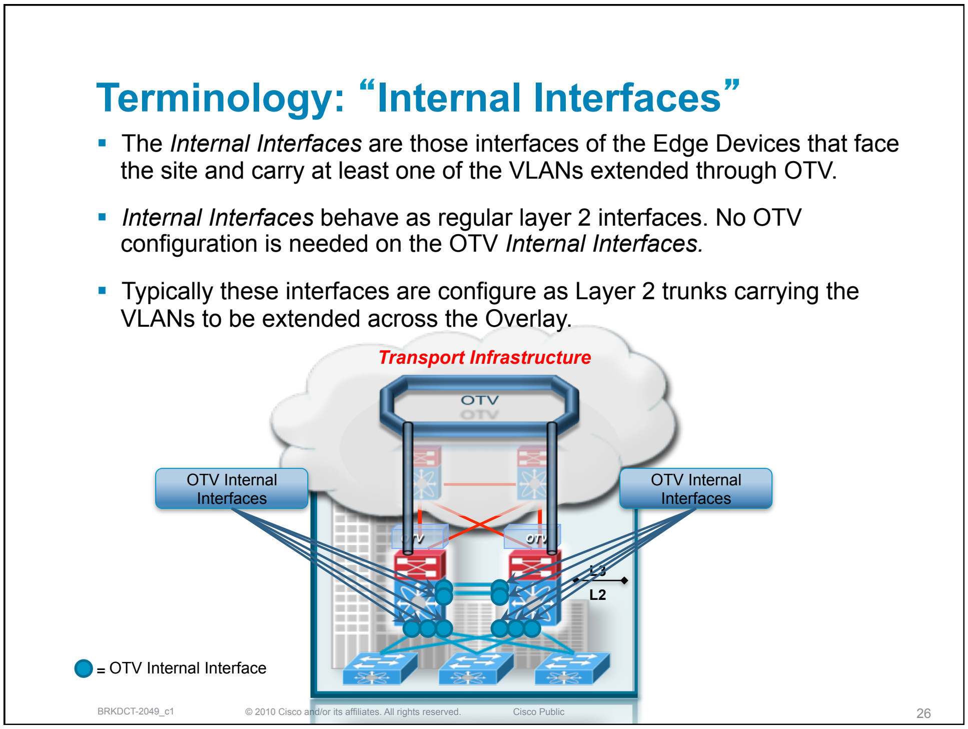
- Preserve Failure Boundary
- Built-in Loop Prevention
- Automated Multi-homing
- Site Independence



Terminology: “Edge Device”

- The *Edge Device* is responsible for performing all the OTV functionality.
- The *Edge Device* 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).*

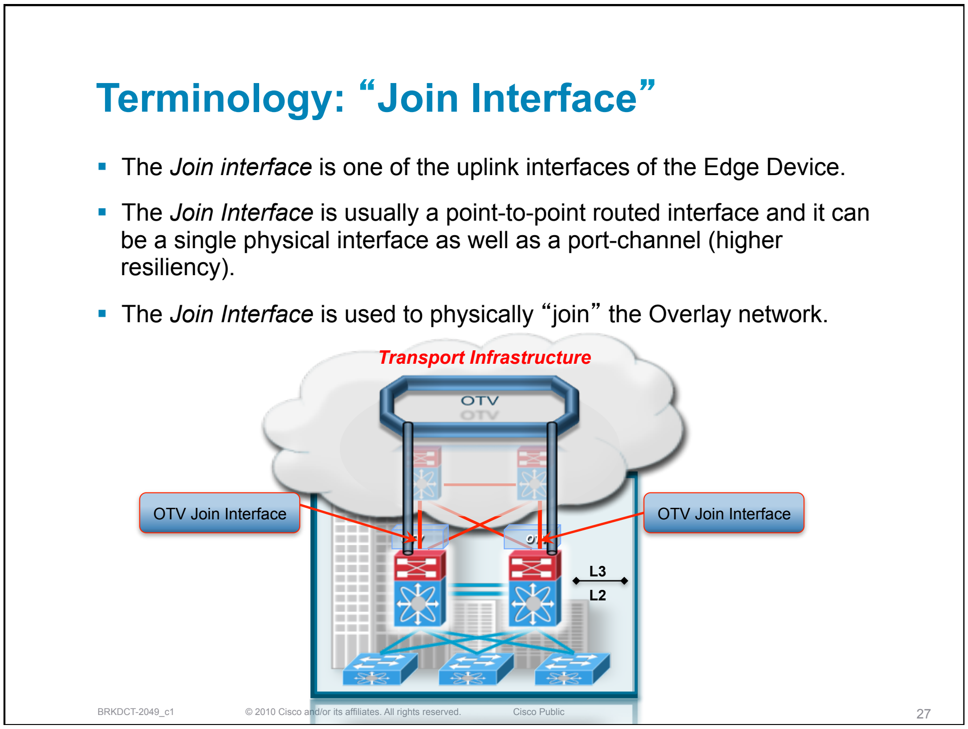
It can be owned by the Enterprise or by the Service Provider



Terminology: “Internal Interfaces”

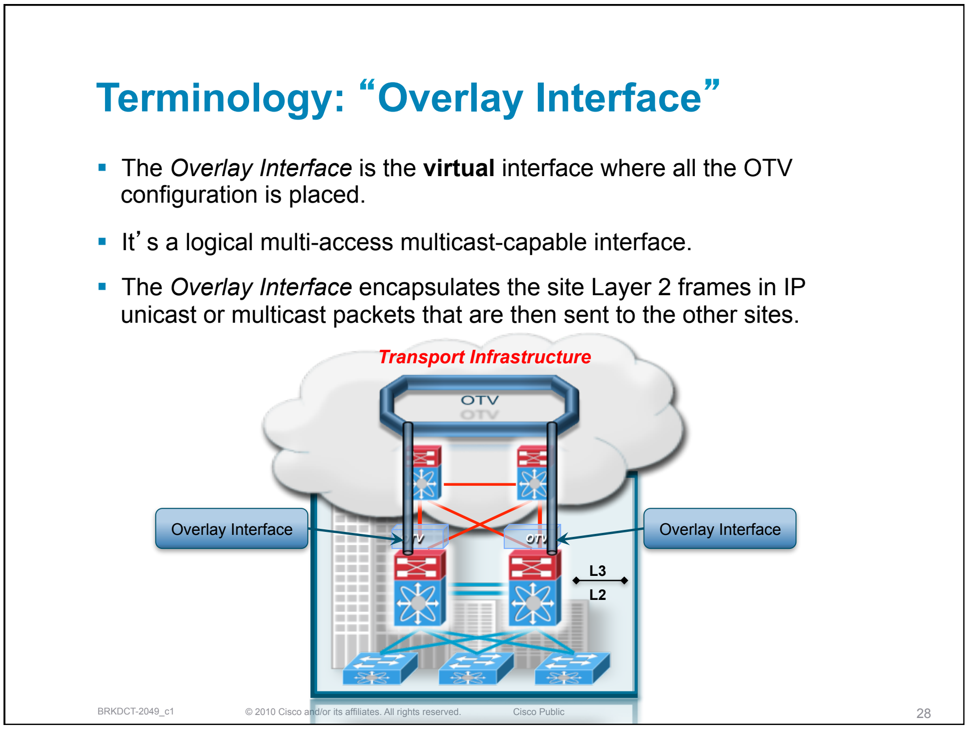
- The *Internal Interfaces* are those interfaces of the Edge Devices that face the site and carry at least one of the VLANs extended through OTV.
- *Internal Interfaces* behave as regular layer 2 interfaces. No OTV configuration is needed on the OTV *Internal Interfaces.*
- Typically these interfaces are configure as Layer 2 trunks carrying the VLANs to be extended across the Overlay.

● = OTV Internal Interface



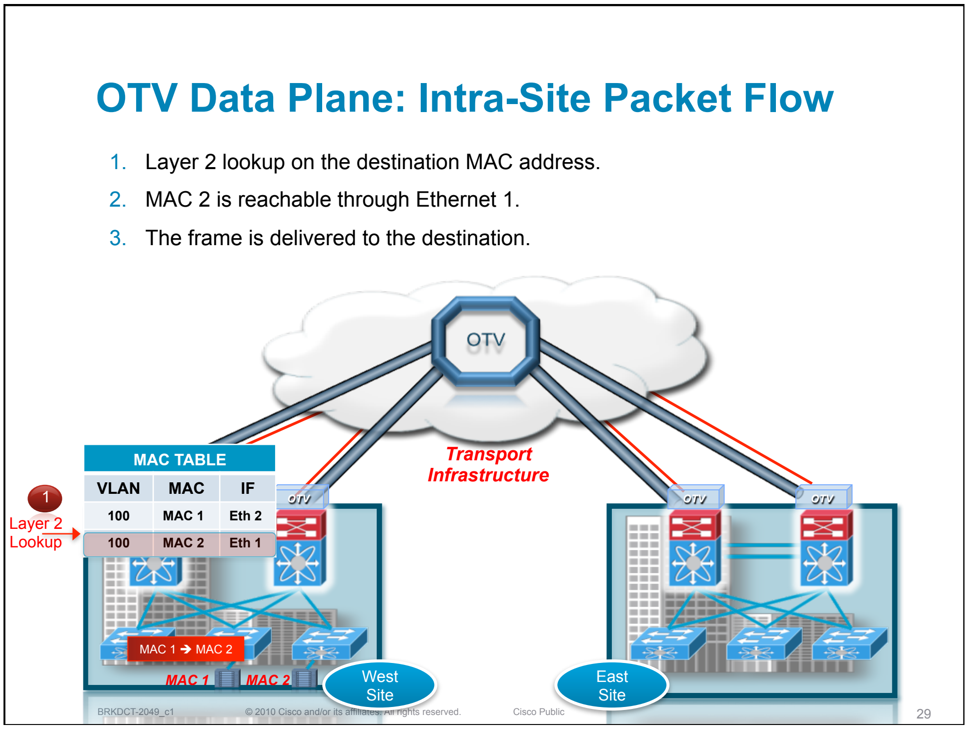
Terminology: “Join Interface”

- The *Join interface* is one of the uplink interfaces of the Edge Device.
- The *Join Interface* is usually a point-to-point routed interface and it can be a single physical interface as well as a port-channel (higher resiliency).
- The *Join Interface* is used to physically “join” the Overlay network.



Terminology: “Overlay Interface”

- The *Overlay Interface* is the **virtual** interface where all the OTV configuration is placed.
- It’s a logical multi-access multicast-capable interface.
- The *Overlay Interface* encapsulates the site Layer 2 frames in IP unicast or multicast packets that are then sent to the other sites.

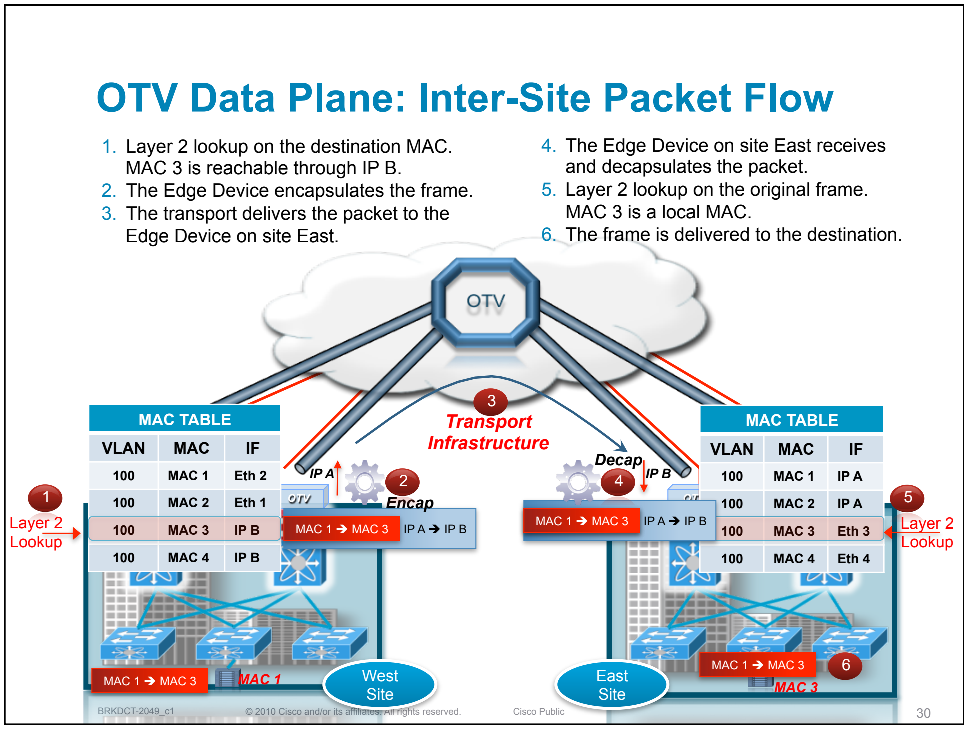


OTV Data Plane: Intra-Site Packet Flow

1. Layer 2 lookup on the destination MAC address.
2. MAC 2 is reachable through Ethernet 1.
3. The frame is delivered to the destination.

MAC TABLE

VLAN	MAC	IF
100	MAC 1	Eth 2
100	MAC 2	Eth 1



OTV Data Plane: Inter-Site Packet Flow

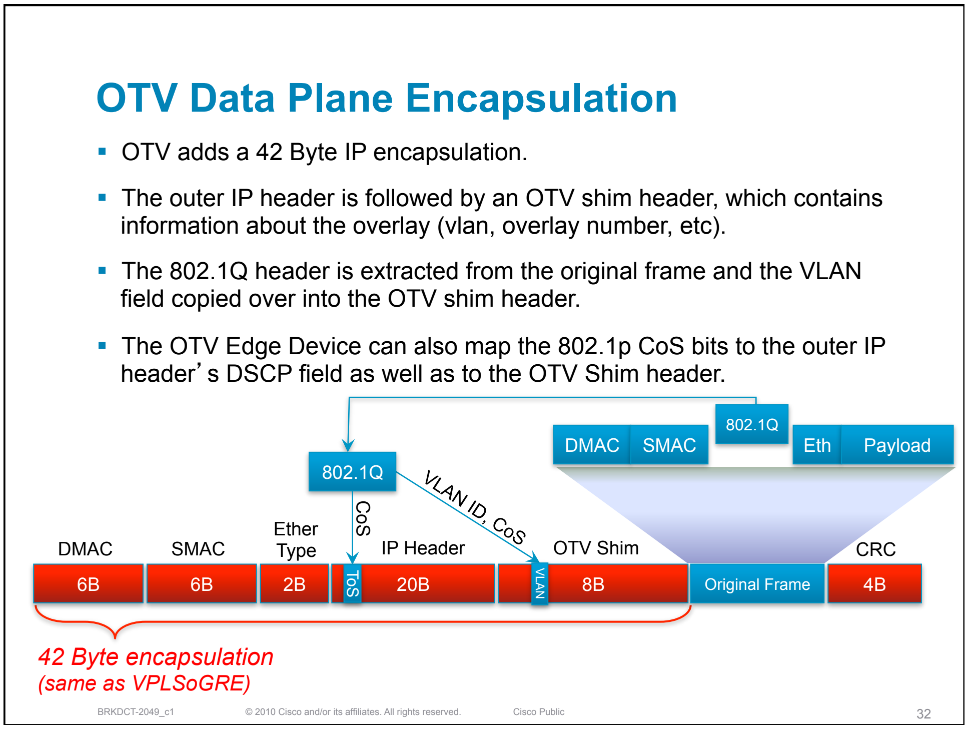
1. Layer 2 lookup on the destination MAC. MAC 3 is reachable through IP B.
2. The Edge Device encapsulates the frame.
3. The transport delivers the packet to the Edge Device on site East.
4. The Edge Device on site East receives and decapsulates the packet.
5. Layer 2 lookup on the original frame. MAC 3 is a local MAC.
6. The frame is delivered to the destination.

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

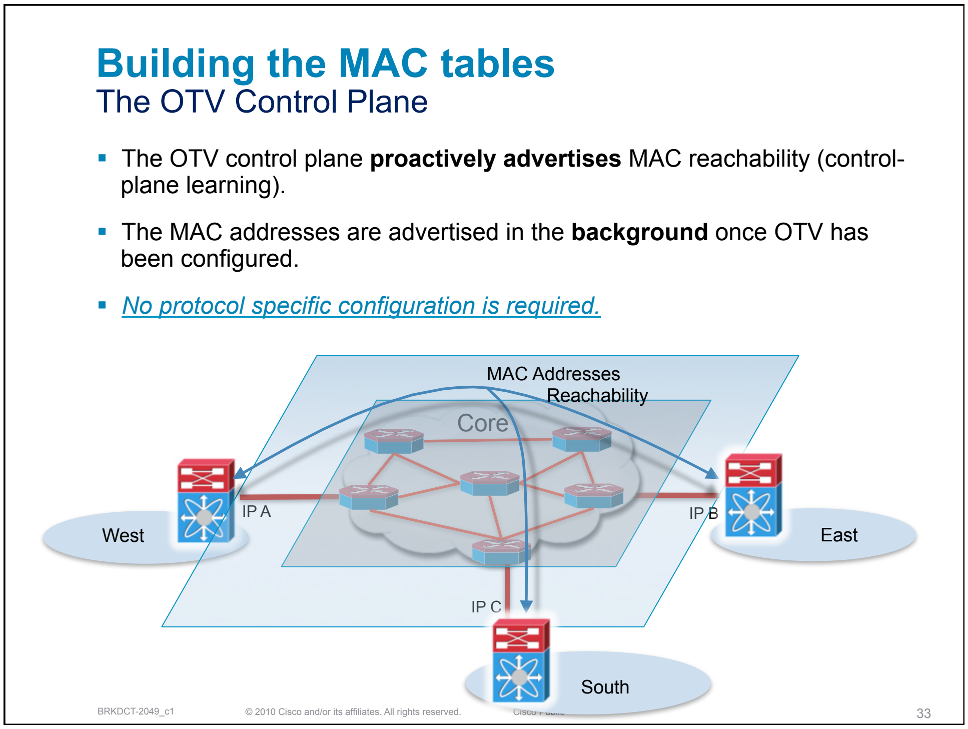
VLAN	MAC	IF
100	MAC 1	IP A
100	MAC 2	IP A
100	MAC 3	Eth 3
100	MAC 4	Eth 4



OTV Data Plane Encapsulation

- OTV adds a 42 Byte IP encapsulation.
- The outer IP header is followed by an OTV shim header, which contains information about the overlay (vlan, overlay number, etc).
- The 802.1Q header is extracted from the original frame and the VLAN field copied over into the OTV shim header.
- The OTV Edge Device can also map the 802.1p CoS bits to the outer IP header’s DSCP field as well as to the OTV Shim header.

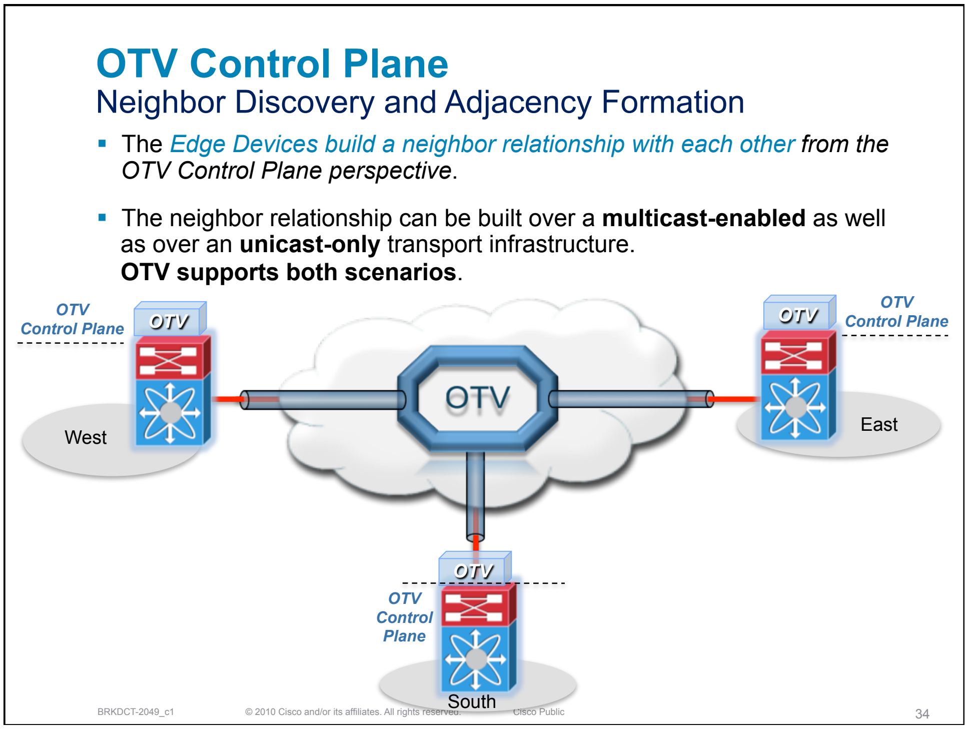
42 Byte encapsulation (same as VPLSoGRE)



Building the MAC tables

The OTV Control Plane

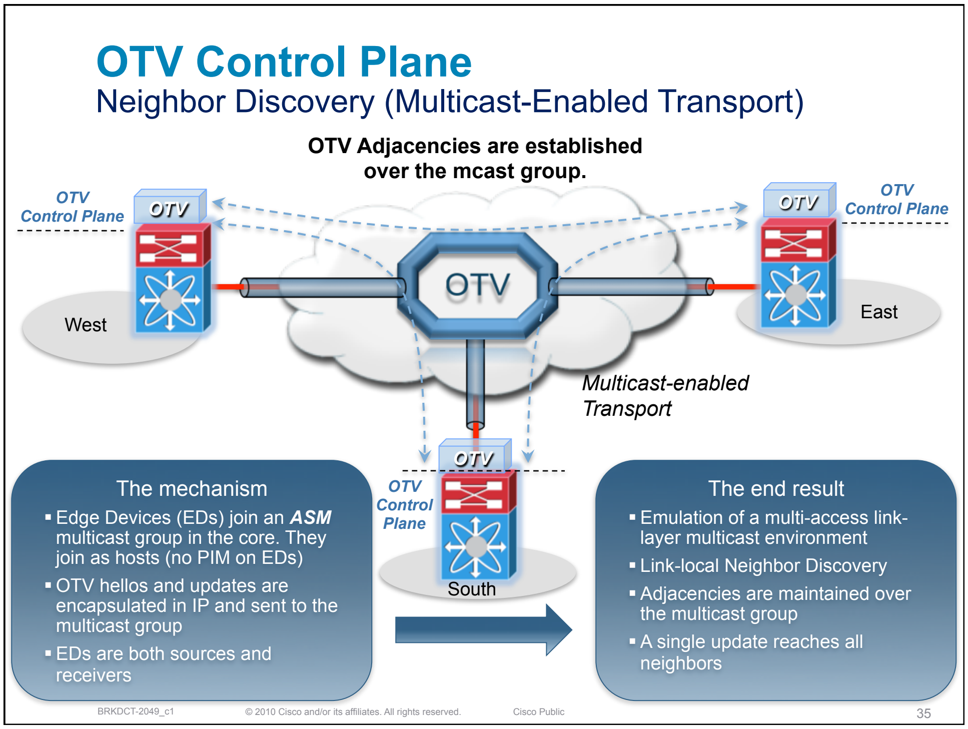
- The OTV control plane **proactively advertises** MAC reachability (control-plane learning).
- The MAC addresses are advertised in the **background** once OTV has been configured.
- *No protocol specific configuration is required.*



OTV Control Plane

Neighbor Discovery and Adjacency Formation

- The *Edge Devices build a neighbor relationship with each other from the OTV Control Plane perspective*.
- The neighbor relationship can be built over a **multicast-enabled** as well as over an **unicast-only** transport infrastructure.
OTV supports both scenarios.



OTV Control Plane

Neighbor Discovery (Multicast-Enabled Transport)

OTV Adjacencies are established over the mcast group.

The mechanism

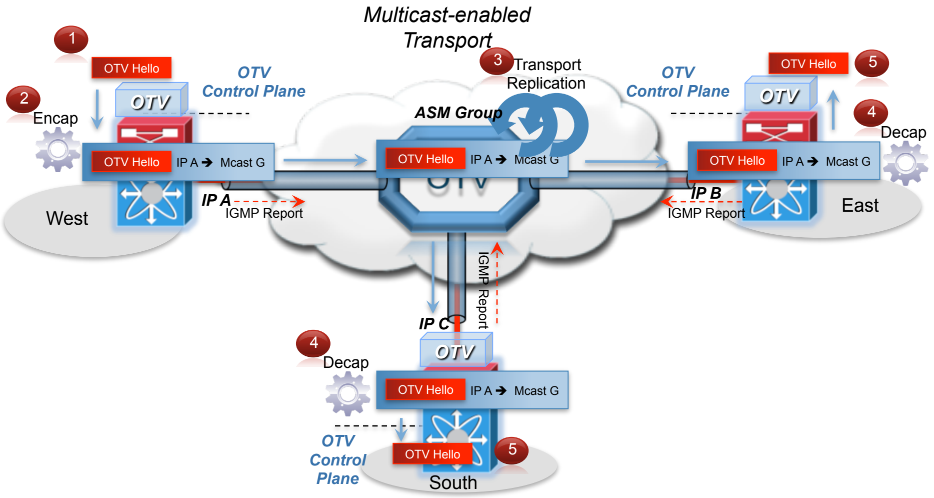
- Edge Devices (EDs) join an ***ASM*** multicast group in the core. They join as hosts (no PIM on EDs)
- OTV hellos and updates are encapsulated in IP and sent to the multicast group
- EDs are both sources and receivers

The end result

- Emulation of a multi-access link-layer multicast environment
- Link-local Neighbor Discovery
- Adjacencies are maintained over the multicast group
- A single update reaches all neighbors

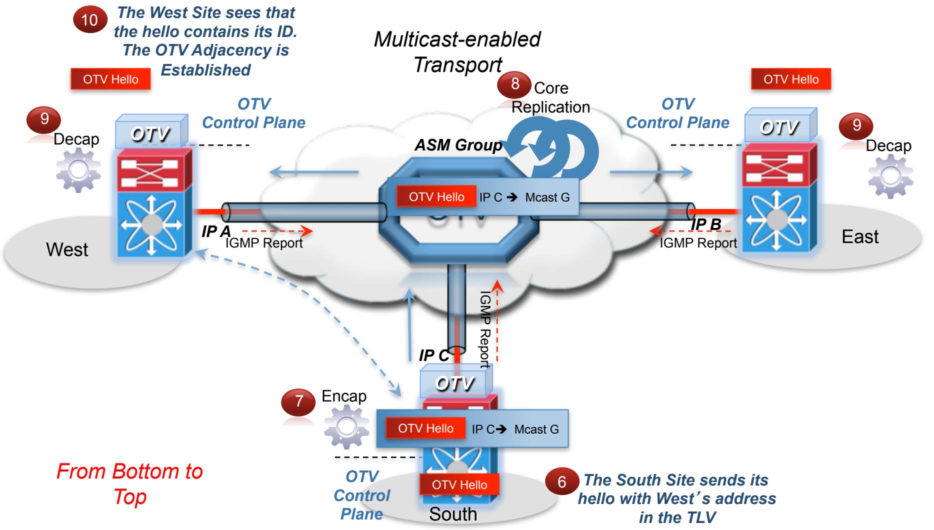
OTV Control Plane

Neighbor Discovery (Multicast-Enabled Transport – 1)



OTV Control Plane

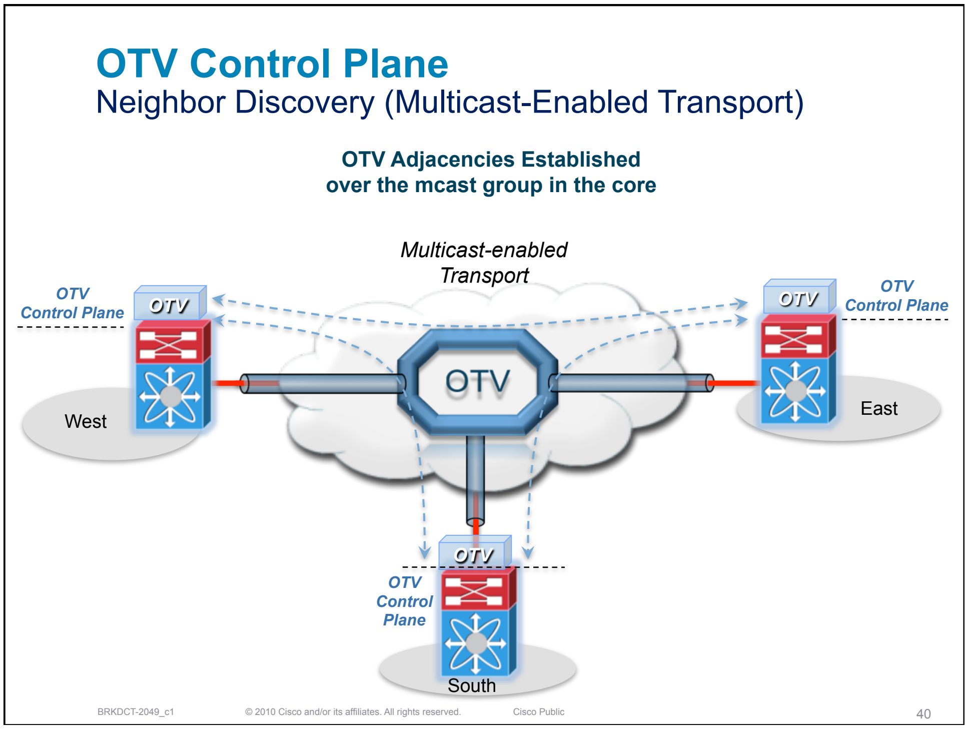
Neighbor Discovery (Multicast-Enabled Transport – 2)



The West Site sees that the hello contains its ID. The OTV Adjacency is Established

The South Site sends its hello with West's address in the TLV

From Bottom to Top

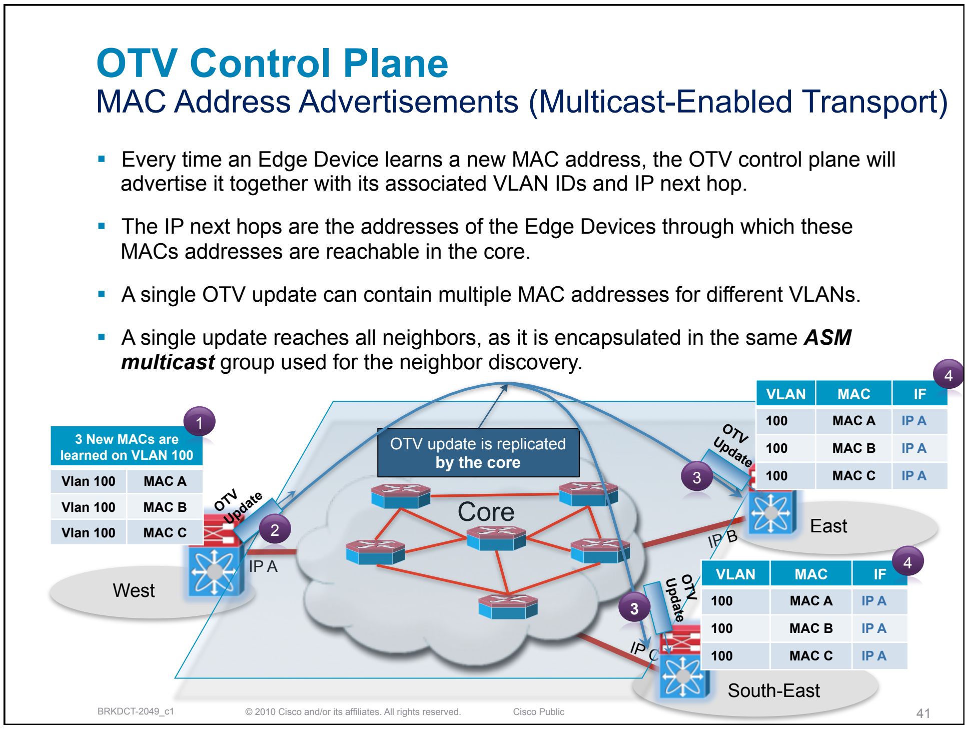


OTV Control Plane

Neighbor Discovery (Multicast-Enabled Transport)

OTV Adjacencies Established over the mcast group in the core

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 40



OTV Control Plane

MAC Address Advertisements (Multicast-Enabled Transport)

- Every time an Edge Device learns a new MAC address, the OTV control plane will advertise it together with its associated VLAN IDs and IP next hop.
- The IP next hops are the addresses of the Edge Devices through which these MACs addresses are reachable in the core.
- A single OTV update can contain multiple MAC addresses for different VLANs.
- A single update reaches all neighbors, as it is encapsulated in the same ***ASM multicast*** group used for the neighbor discovery.

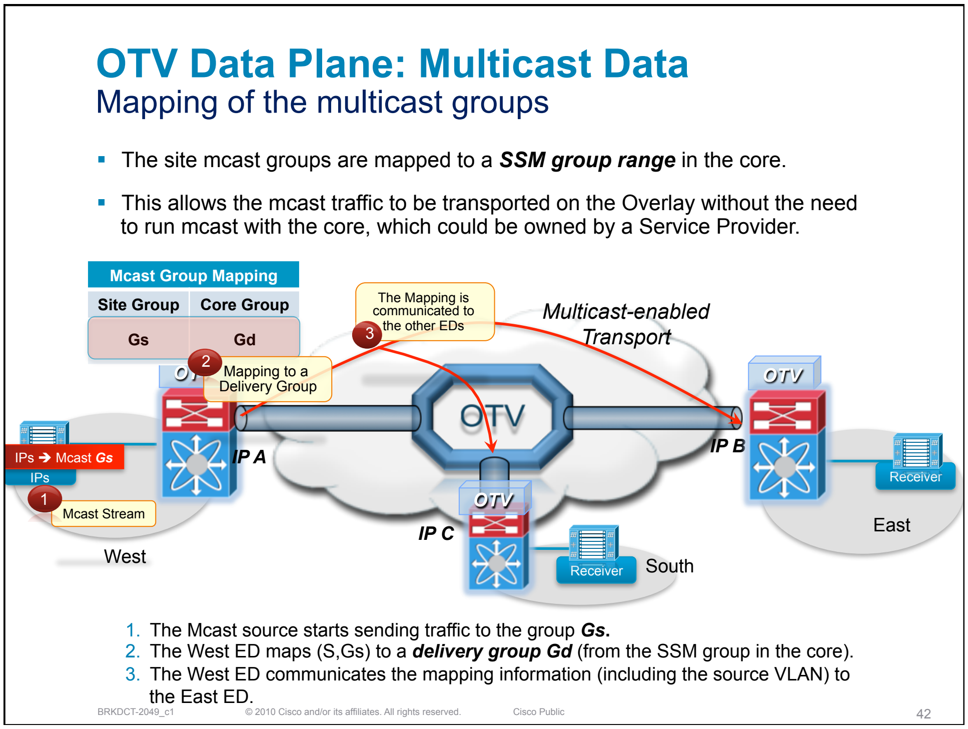
1
3 New MACs are learned on VLAN 100
2
OTV Update
IP A
West
OTV update is replicated by the core
Core
3
OTV Update
IP B
East
4
3
OTV Update
IP C
South-East
4

Vlan 100	MAC A
Vlan 100	MAC B
Vlan 100	MAC C

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

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

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 41



OTV Data Plane: Multicast Data

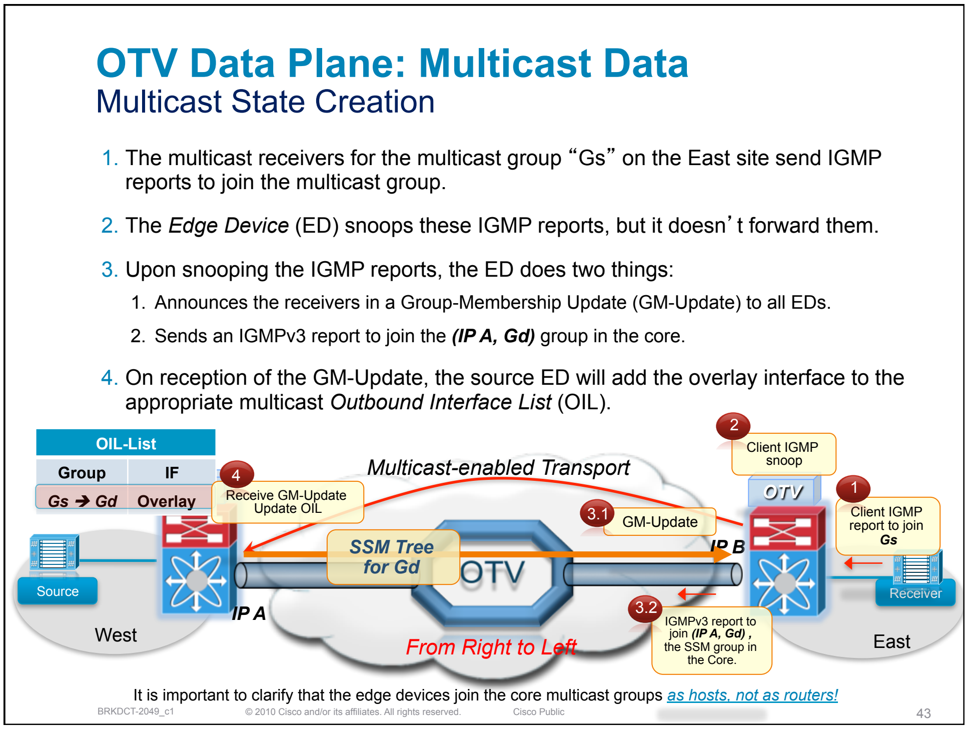
Mapping of the multicast groups

- The site mcast groups are mapped to a ***SSM group range*** in the core.
- This allows the mcast traffic to be transported on the Overlay without the need to run mcast with the core, which could be owned by a Service Provider.

Mcast Group Mapping	
Site Group	Core Group
Gs	Gd

The Mapping is communicated to the other EDs
3
Multicast-enabled Transport
2
Mapping to a Delivery Group
OTV
OTV
OTV
OTV
IP A
IP B
IP C
IPs ➔ Mcast Gs
IPs
1
Mcast Stream
West
Receiver
East
Receiver
South

1. The Mcast source starts sending traffic to the group ***Gs.***
2. The West ED maps (S,Gs) to a ***delivery group Gd*** (from the SSM group in the core).
3. The West ED communicates the mapping information (including the source VLAN) to the East ED.



OTV Data Plane: Multicast Data

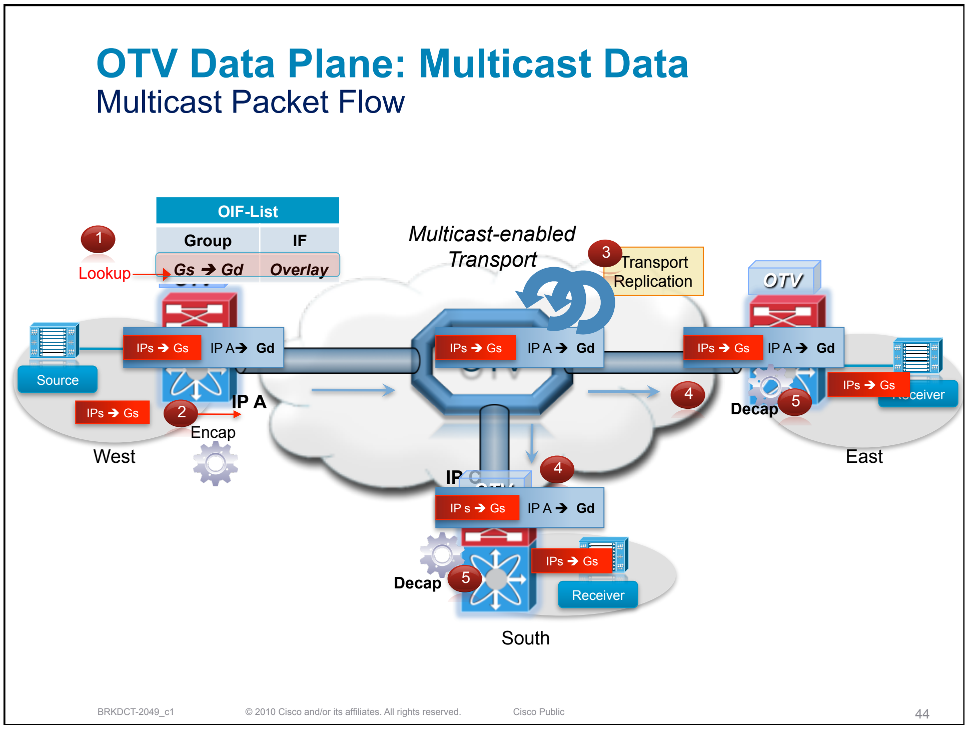
Multicast State Creation

1. The multicast receivers for the multicast group “Gs” on the East site send IGMP reports to join the multicast group.
2. The *Edge Device* (ED) snoops these IGMP reports, but it doesn’t forward them.
3. Upon snooping the IGMP reports, the ED does two things:
   1. Announces the receivers in a Group-Membership Update (GM-Update) to all EDs.
   2. Sends an IGMPv3 report to join the ***(IP A, Gd)*** group in the core.
4. On reception of the GM-Update, the source ED will add the overlay interface to the appropriate multicast *Outbound Interface List* (OIL).

OIL-List	
Group	IF
Gs ➔ Gd	Overlay

2
Client IGMP snoop
4
Receive GM-Update Update OIL
Multicast-enabled Transport
OTV
OTV
3.1
GM-Update
1
Client IGMP report to join Gs
SSM Tree for Gd
IP B
Source
West
IP A
From Right to Left
3.2
IGMPv3 report to join (IP A, Gd) , the SSM group in the Core.
Receiver
East

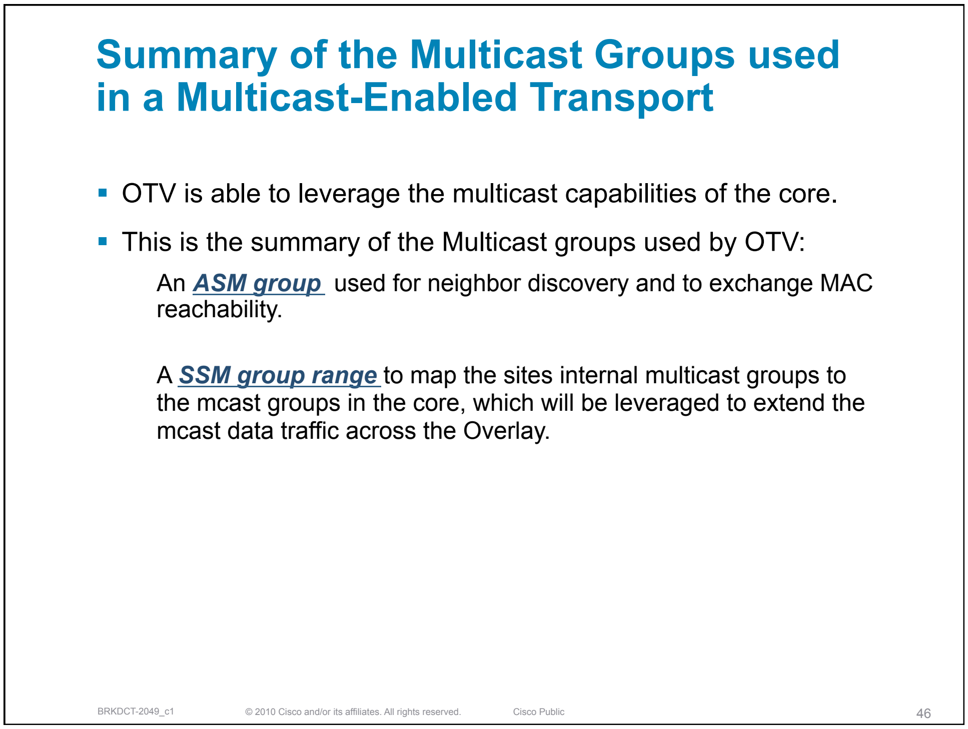
It is important to clarify that the edge devices join the core multicast groups *as hosts, not as routers!*



OTV Data Plane: Multicast Data

Multicast Packet Flow

OIF-List	
Group	IF
Gs → Gd	*Overlay*

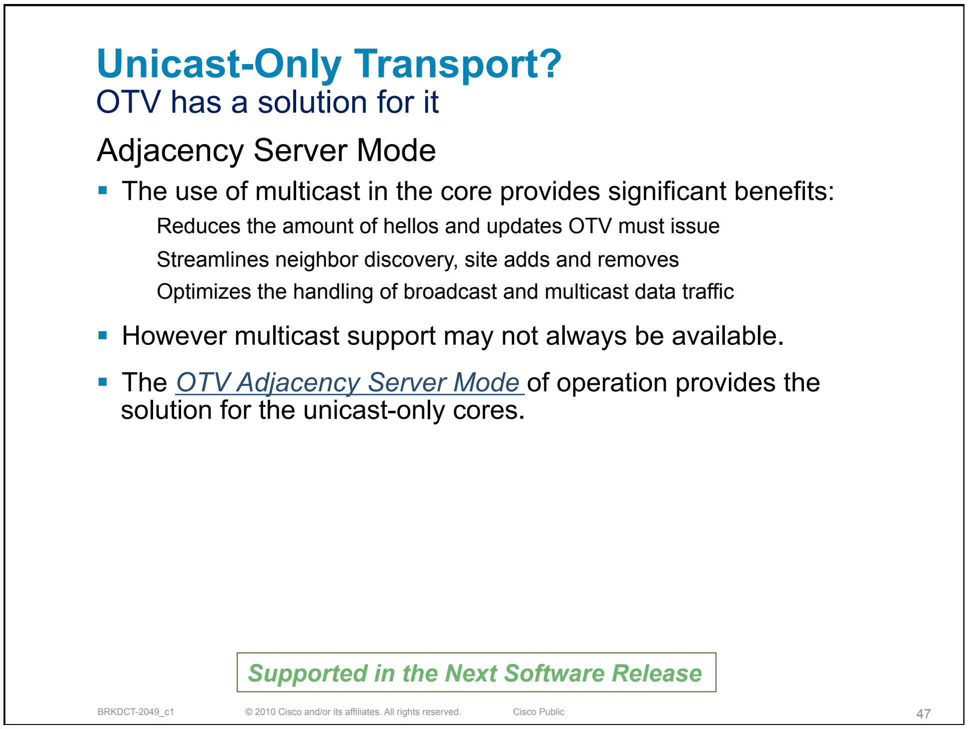


Summary of the Multicast Groups used in a Multicast-Enabled Transport

- OTV is able to leverage the multicast capabilities of the core.
- This is the summary of the Multicast groups used by OTV:

An ***ASM group*** used for neighbor discovery and to exchange MAC reachability.

A ***SSM group range*** to map the sites internal multicast groups to the mcast groups in the core, which will be leveraged to extend the mcast data traffic across the Overlay.



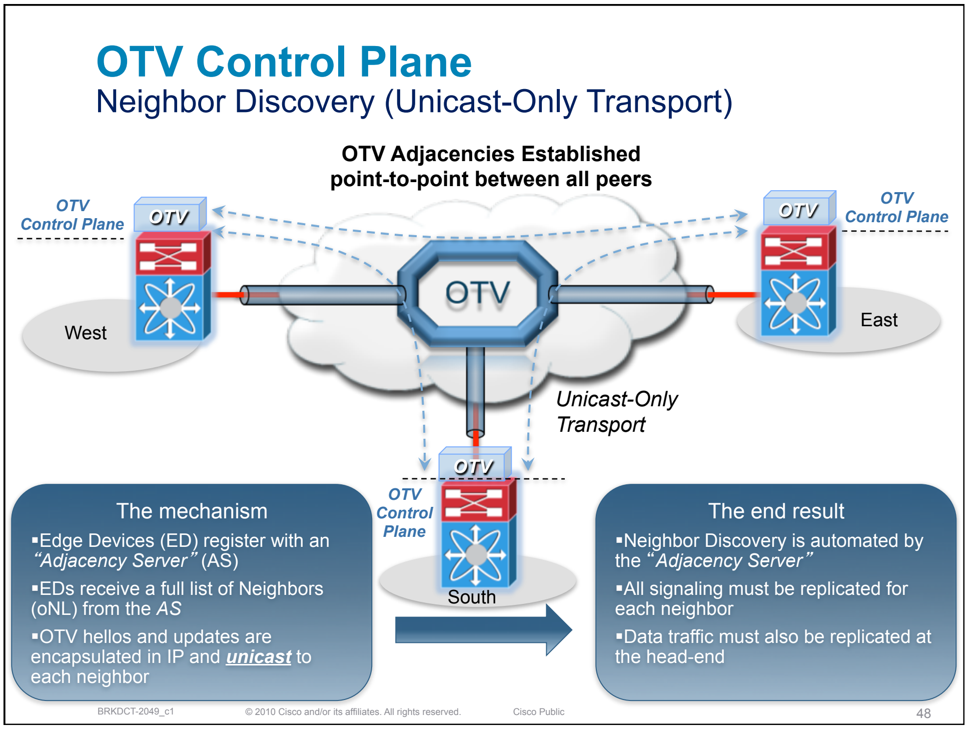
Unicast-Only Transport?

OTV has a solution for it

Adjacency Server Mode

- The use of multicast in the core provides significant benefits:
  - Reduces the amount of hellos and updates OTV must issue
  - Streamlines neighbor discovery, site adds and removes
  - Optimizes the handling of broadcast and multicast data traffic
- However multicast support may not always be available.
- The *OTV Adjacency Server Mode* of operation provides the solution for the unicast-only cores.

Supported in the Next Software Release



OTV Control Plane

Neighbor Discovery (Unicast-Only Transport)

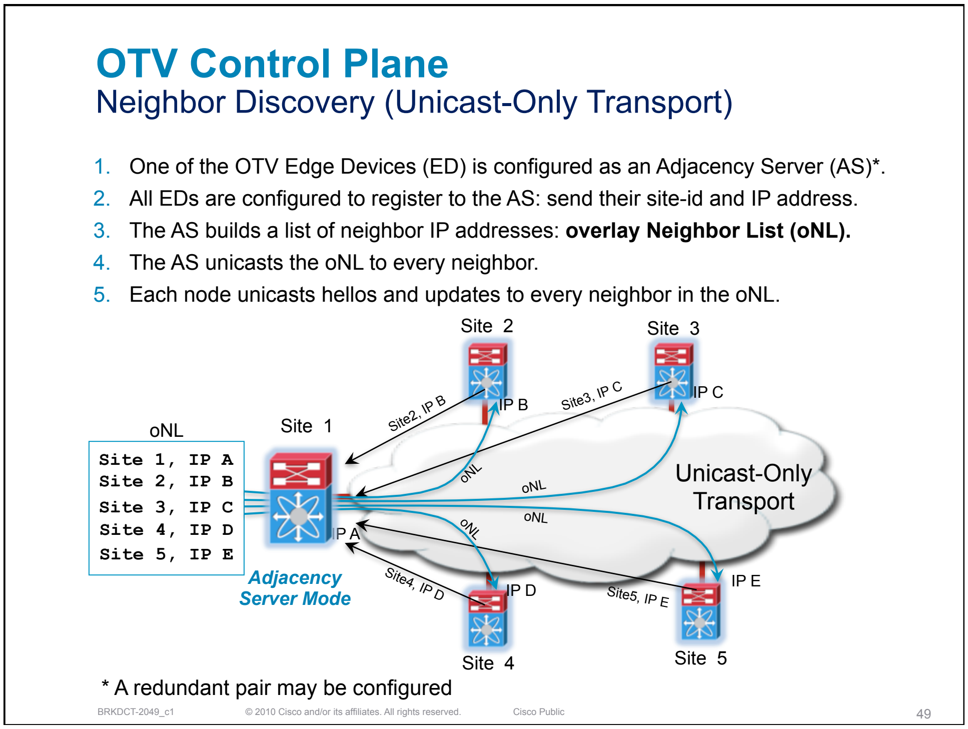
OTV Adjacencies Established point-to-point between all peers

The mechanism

- Edge Devices (ED) register with an *"Adjacency Server"* (AS)
- EDs receive a full list of Neighbors (oNL) from the *AS*
- OTV hellos and updates are encapsulated in IP and ***unicast*** to each neighbor

The end result

- Neighbor Discovery is automated by the *"Adjacency Server"*
- All signaling must be replicated for each neighbor
- Data traffic must also be replicated at the head-end



OTV Control Plane

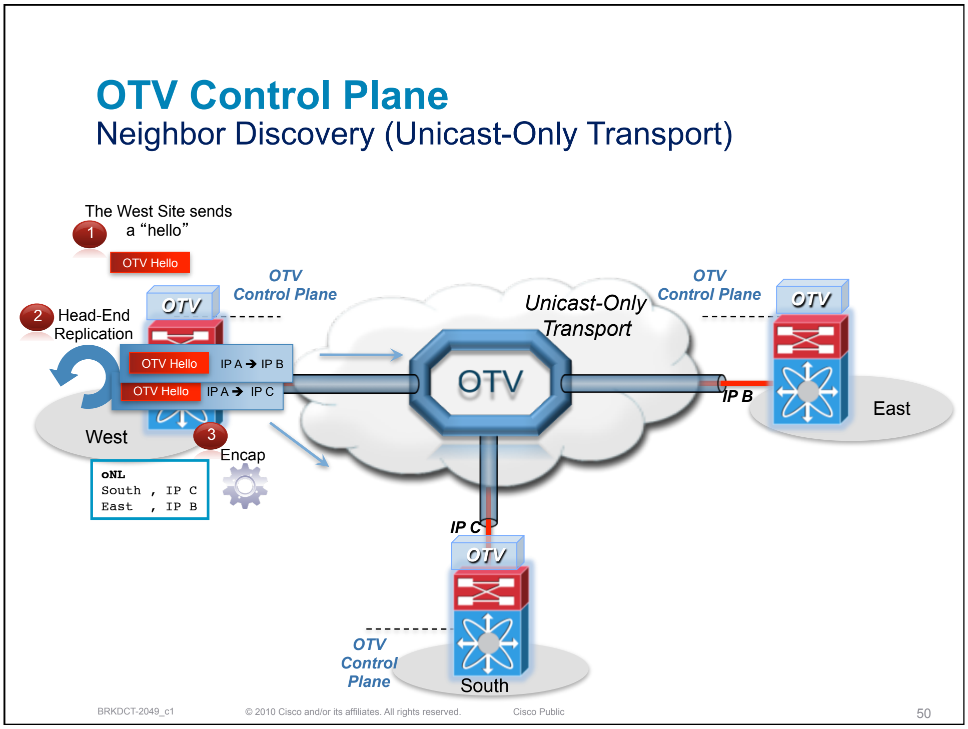
Neighbor Discovery (Unicast-Only Transport)

1. One of the OTV Edge Devices (ED) is configured as an Adjacency Server (AS)*.
2. All EDs are configured to register to the AS: send their site-id and IP address.
3. The AS builds a list of neighbor IP addresses: **overlay Neighbor List (oNL).**
4. The AS unicasts the oNL to every neighbor.
5. Each node unicasts hellos and updates to every neighbor in the oNL.

oNL

Site 1,	IP A
Site 2,	IP B
Site 3,	IP C
Site 4,	IP D
Site 5,	IP E

* A redundant pair may be configured



OTV Control Plane

Neighbor Discovery (Unicast-Only Transport)

1 The West Site sends a "hello"

2 Head-End Replication

3 Encap

OTV Hello	IP A → IP B
OTV Hello	IP A → IP C

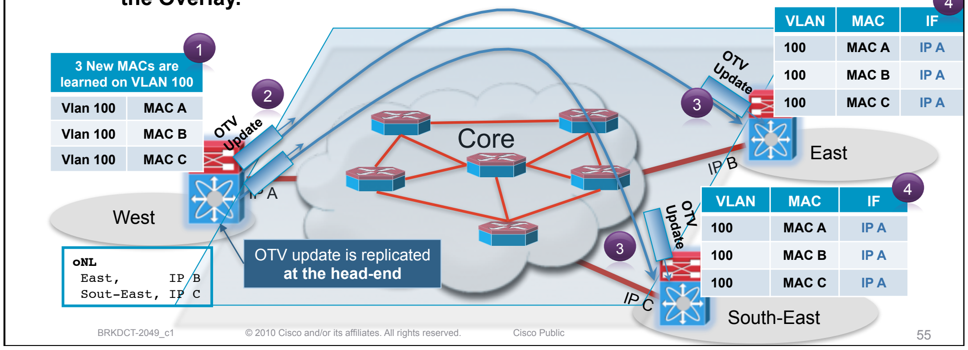
oNL

South ,	IP C
East ,	IP B

OTV Control Plane

MAC Advertisements (Unicast-Only Transport)

- Every time an Edge Device learns a new MAC address, the OTV control plane will advertise it together with its associated VLAN IDs and IP next hop.
- The IP next hops are the addresses of the Edge Devices through which these MACs are reachable in the core.
- A single OTV update can contain multiple MAC addresses for different VLANs.
- **A single update needs to be created for each destination EDs present on the Overlay.**



3 New MACs are learned on VLAN 100

Vlan 100	MAC A
Vlan 100	MAC B
Vlan 100	MAC C

```
oNL
 East,       IP B
 Sout-East,  IP C
```

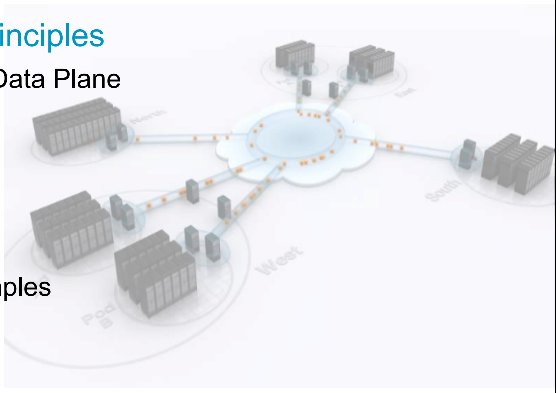
OTV update is replicated **at the head-end**

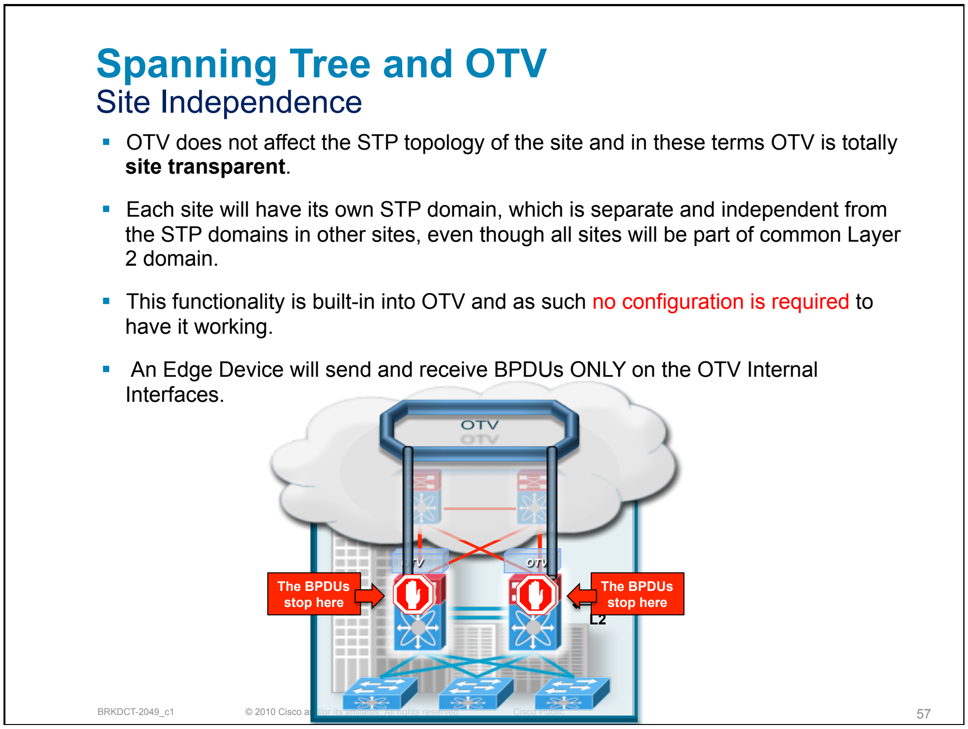
VLAN	MAC	IF
100	MAC A	IP A
100	MAC B	IP A
100	MAC C	IP A

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

Agenda

- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
  - Control Plane and Data Plane
  - Failure Isolation
  - Multi-homing
  - Mobility
  - Path Optimization
  - Configuration Examples
- Use Cases

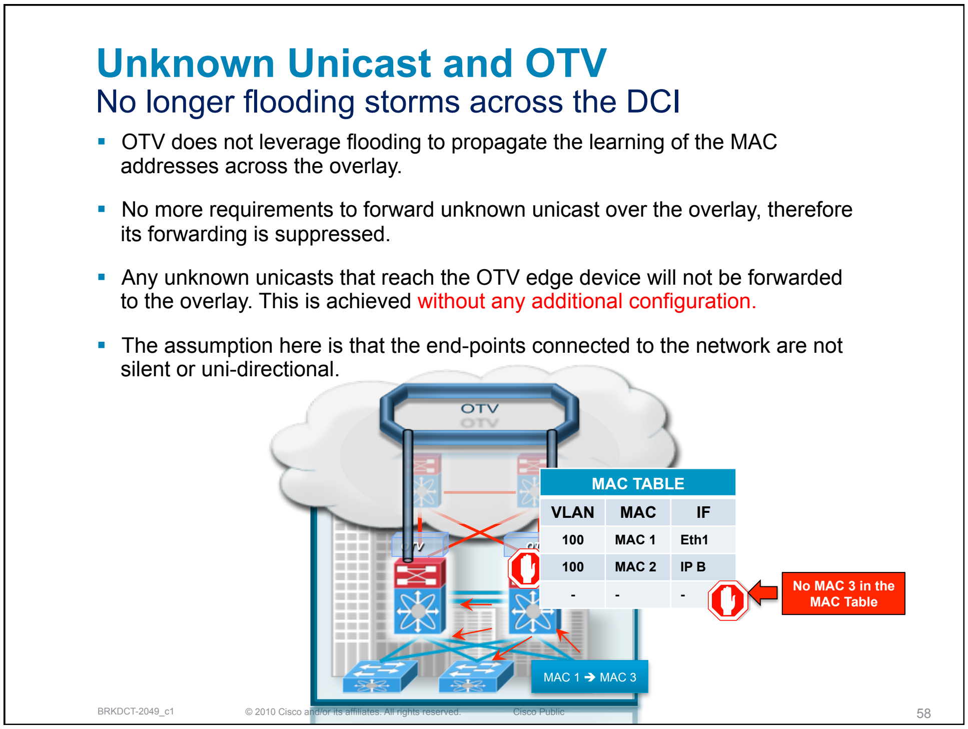




Spanning Tree and OTV
Site Independence

- OTV does not affect the STP topology of the site and in these terms OTV is totally **site transparent**.
- Each site will have its own STP domain, which is separate and independent from the STP domains in other sites, even though all sites will be part of common Layer 2 domain.
- This functionality is built-in into OTV and as such no configuration is required to have it working.
- An Edge Device will send and receive BPDUs ONLY on the OTV Internal Interfaces.

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 57



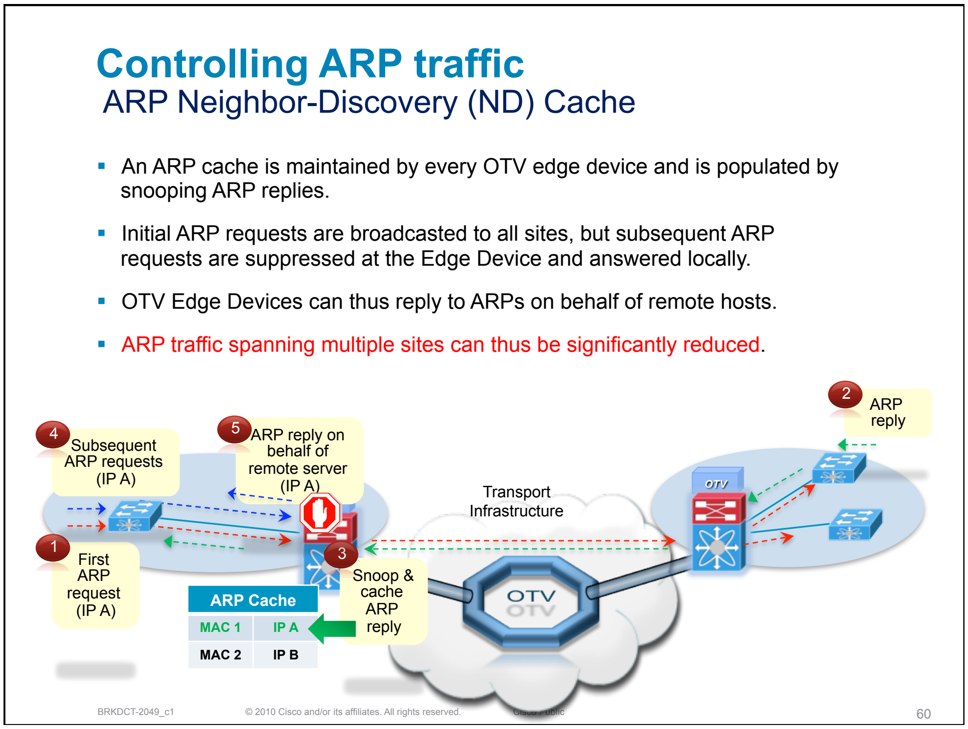
Unknown Unicast and OTV
No longer flooding storms across the DCI

- OTV does not leverage flooding to propagate the learning of the MAC addresses across the overlay.
- No more requirements to forward unknown unicast over the overlay, therefore its forwarding is suppressed.
- Any unknown unicasts that reach the OTV edge device will not be forwarded 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
-	-	-

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 58



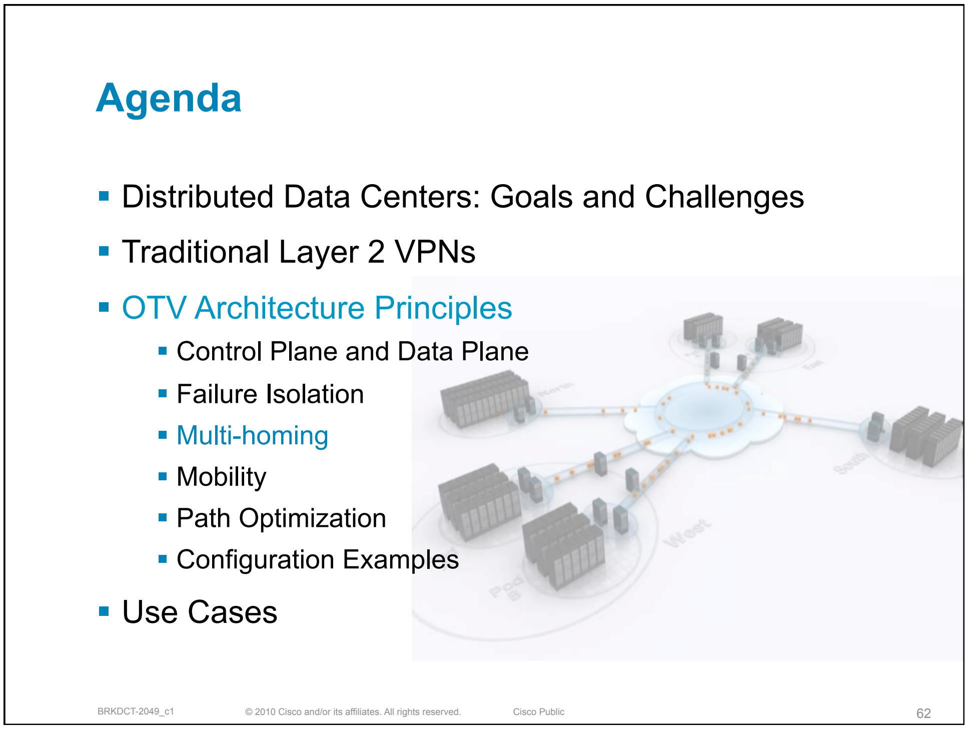
Controlling ARP traffic

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.
- OTV Edge Devices can thus reply to ARPs on behalf of remote hosts.
- ARP traffic spanning multiple sites can thus be significantly reduced.

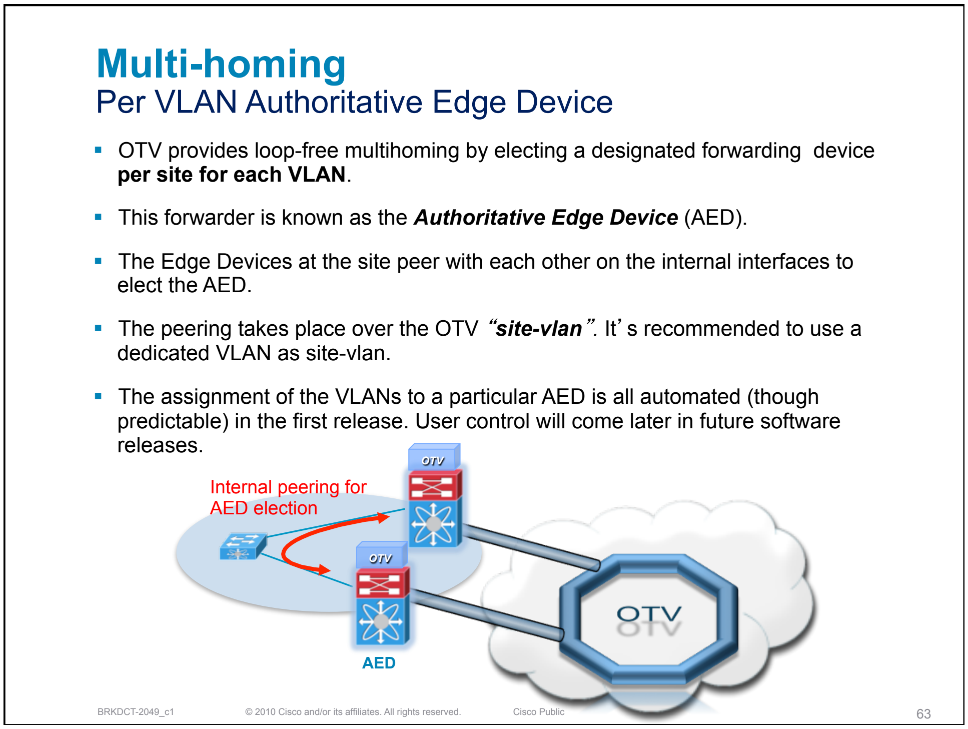
1 First ARP request (IP A)
2 ARP reply
3 Snoop & cache ARP reply
4 Subsequent ARP requests (IP A)
5 ARP reply on behalf of remote server (IP A)
Transport Infrastructure
OTV
OTV

ARP Cache	
MAC 1	IP A
MAC 2	IP B



Agenda

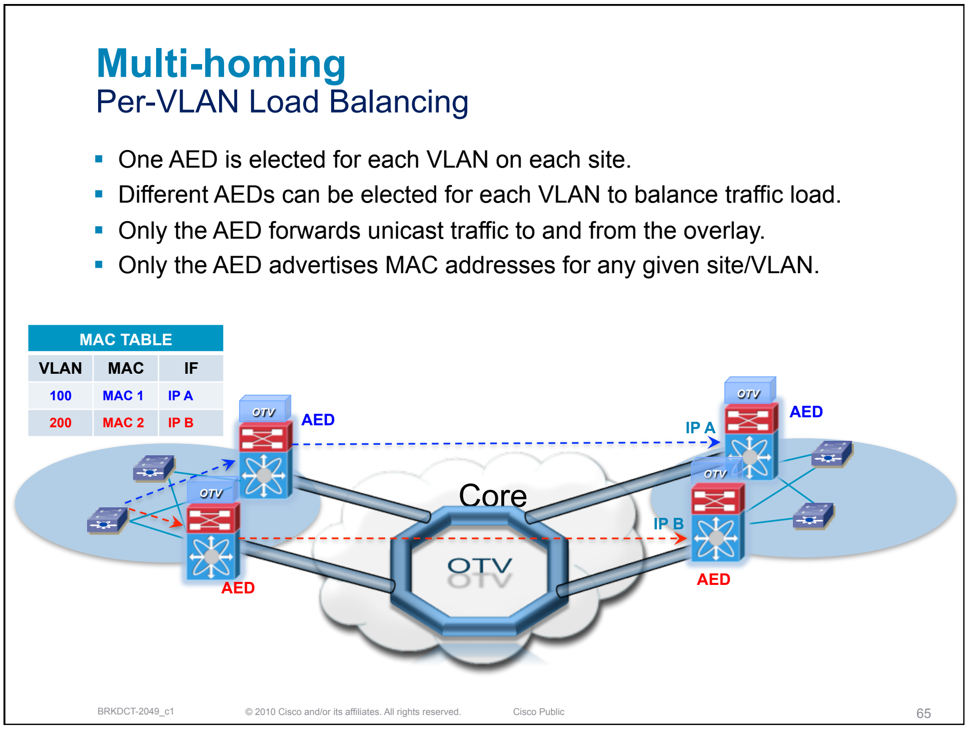
- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
  - Control Plane and Data Plane
  - Failure Isolation
  - Multi-homing
  - Mobility
  - Path Optimization
  - Configuration Examples
- Use Cases



Multi-homing

Per VLAN Authoritative Edge Device

- OTV provides loop-free multihoming by electing a designated forwarding device **per site for each VLAN**.
- This forwarder is known as the ***Authoritative Edge Device*** (AED).
- The Edge Devices at the site peer with each other on the internal interfaces to elect the AED.
- The peering takes place over the OTV *“**site-vlan**”*. It’ s recommended to use a dedicated VLAN as site-vlan.
- The assignment of the VLANs to a particular AED is all automated (though predictable) in the first release. User control will come later in future software releases.



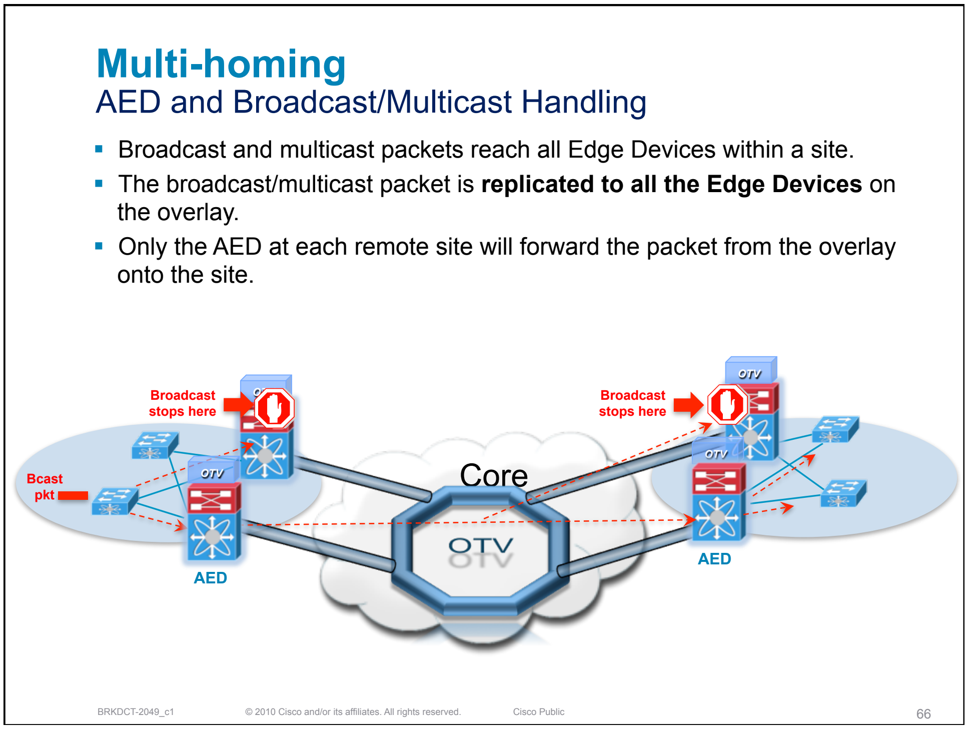
Multi-homing

Per-VLAN Load Balancing

- One AED is elected for each VLAN on each site.
- Different AEDs can be elected for each VLAN to balance traffic load.
- Only the AED forwards unicast traffic to and from the overlay.
- Only the AED advertises MAC addresses for any given site/VLAN.

MAC TABLE

VLAN	MAC	IF
100	MAC 1	IP A
200	MAC 2	IP B

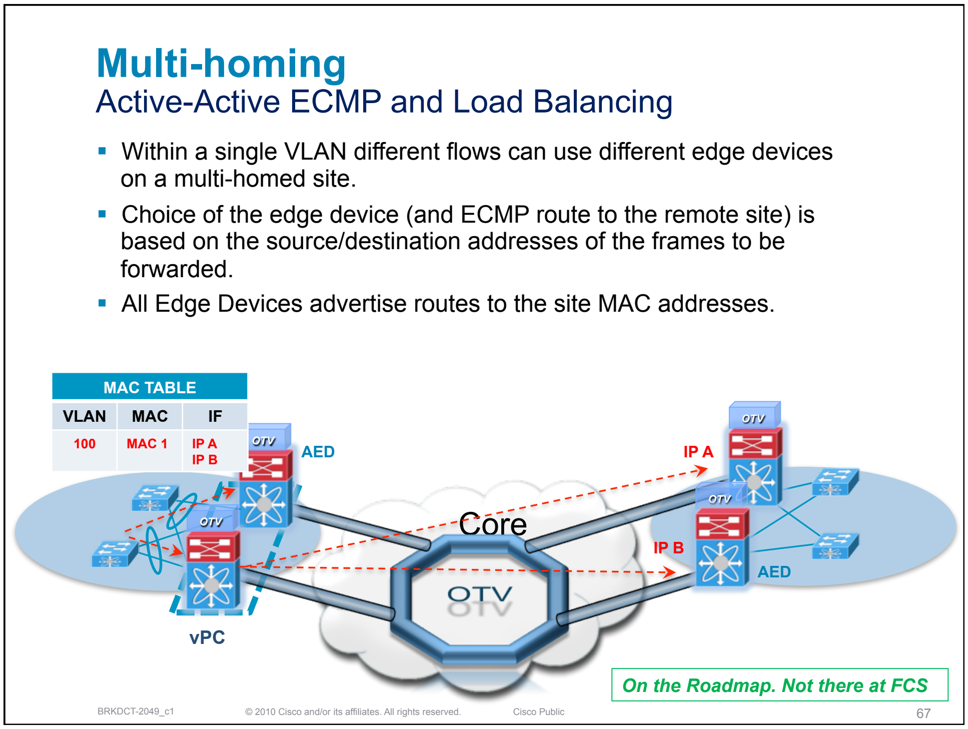


Multi-homing

AED and Broadcast/Multicast Handling

- Broadcast and multicast packets reach all Edge Devices within a site.
- The broadcast/multicast packet is **replicated to all the Edge Devices** on the overlay.
- Only the AED at each remote site will forward the packet from the overlay onto the site.

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 66



Multi-homing

Active-Active ECMP and Load Balancing

- Within a single VLAN different flows can use different edge devices on a multi-homed site.
- Choice of the edge device (and ECMP route to the remote site) is based on the source/destination addresses of the frames to be forwarded.
- All Edge Devices advertise routes to the site MAC addresses.

MAC TABLE

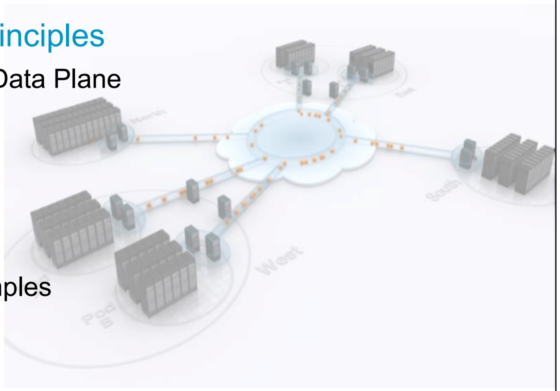
VLAN	MAC	IF
100	MAC 1	IP A IP B

On the Roadmap. Not there at FCS

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 67

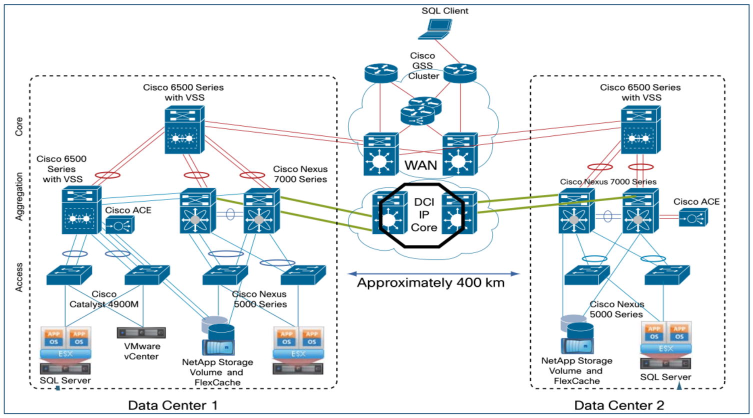
Agenda

- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
  - Control Plane and Data Plane
  - Failure Isolation
  - Multi-homing
  - Mobility
  - Path Optimization
  - Configuration Examples
- Use Cases



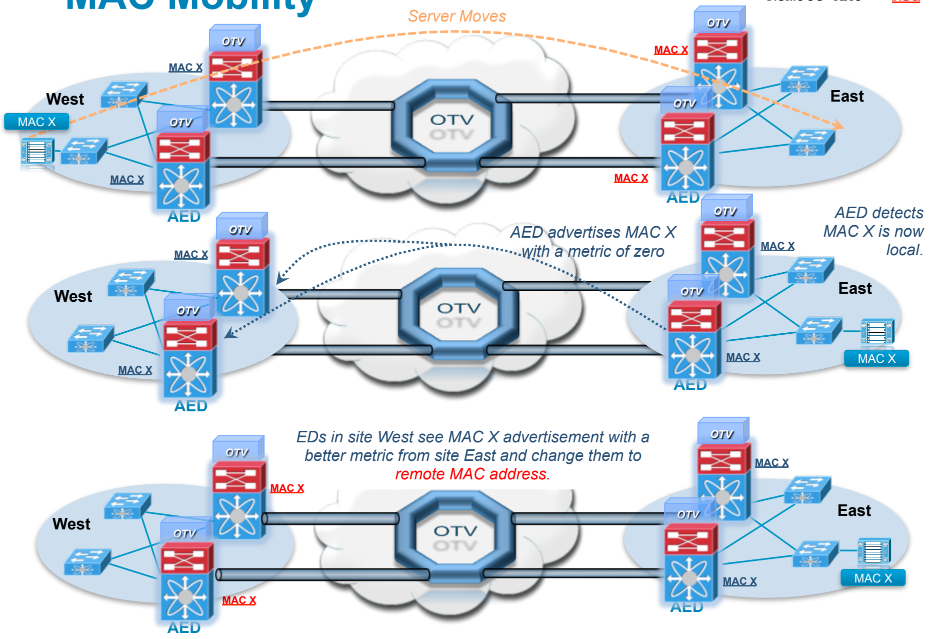
OTV and Long Distance Vmotion

- Cisco, NetApp and VMWare jointly test:
  - Two Data Centers distant ***400 Km*** from each other
  - ***OTV used to extend Layer 2 on 2 pairs of Nexus 7000***



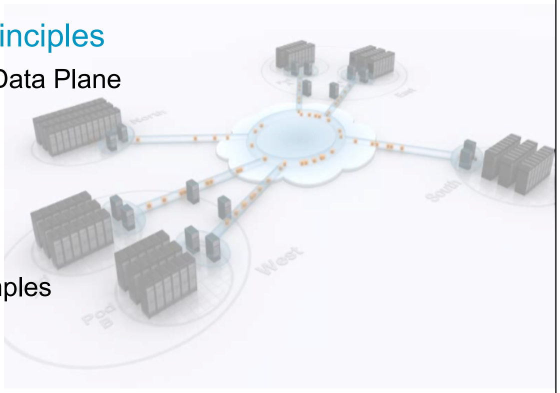
MAC Mobility

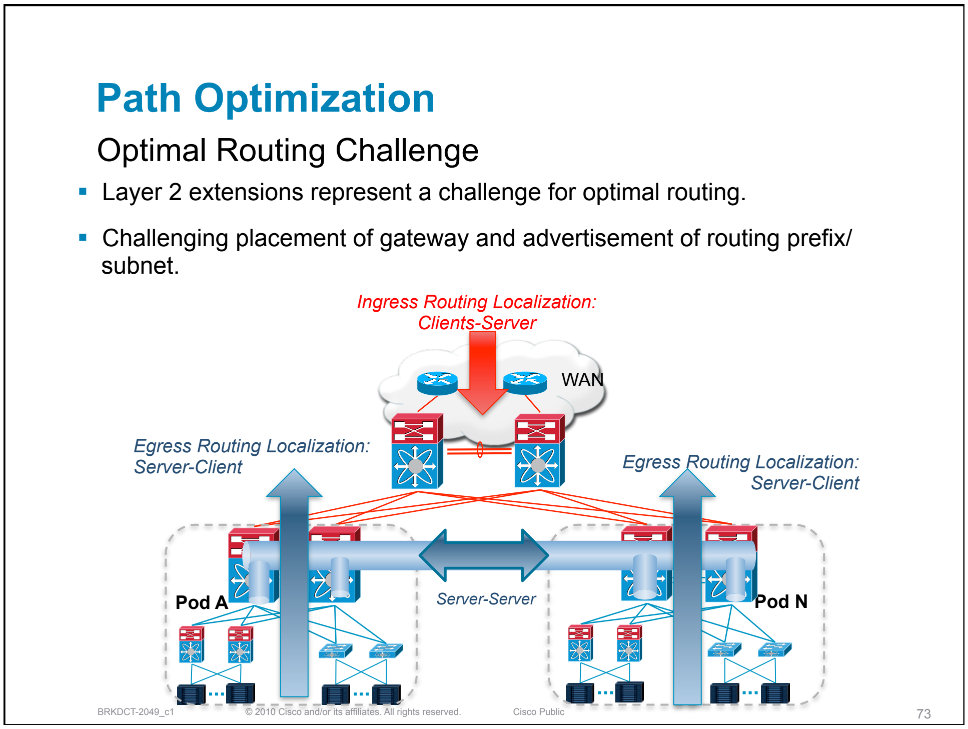
Local MAC = **Blue**
Remote MAC = **Red**


Server Moves
West
East
MAC X
OTV
AED
AED detects MAC X is now local.
AED advertises MAC X with a metric of zero
EDs in site West see MAC X advertisement with a better metric from site East and change them to remote MAC address.

Agenda

- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
  - Control Plane and Data Plane
  - Failure Isolation
  - Multi-homing
  - Mobility
  - Path Optimization
  - Configuration Examples
- Use Cases



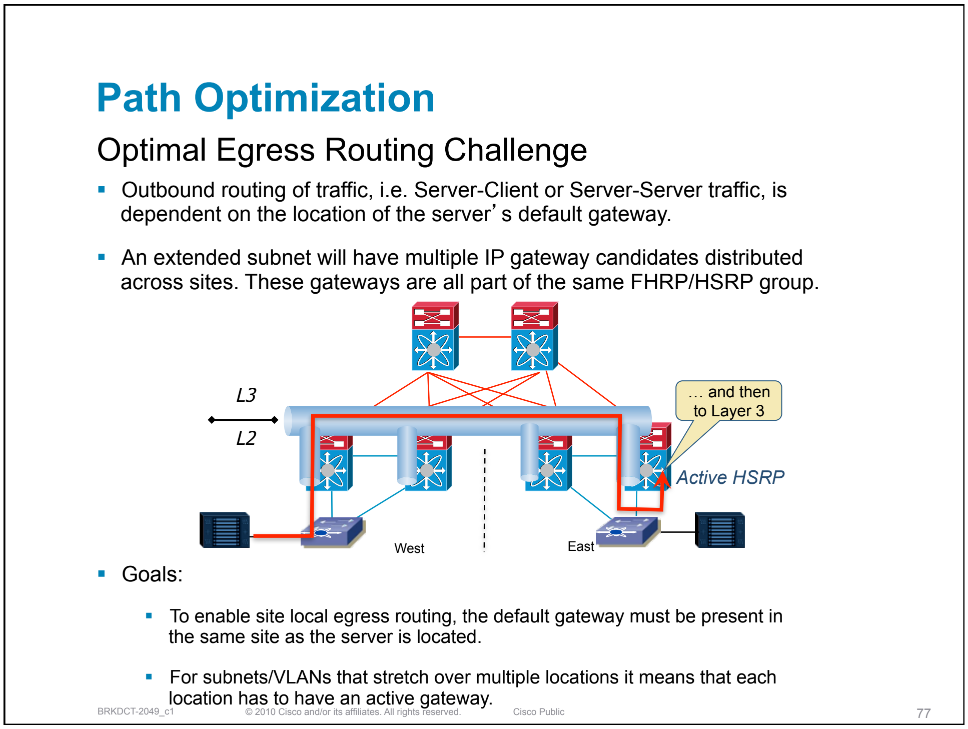


Path Optimization

Optimal Routing Challenge

- Layer 2 extensions represent a challenge for optimal routing.
- Challenging placement of gateway and advertisement of routing prefix/ subnet.

Ingress Routing Localization: Clients-Server
WAN
Egress Routing Localization: Server-Client
Egress Routing Localization: Server-Client
Pod A
Server-Server
Pod N

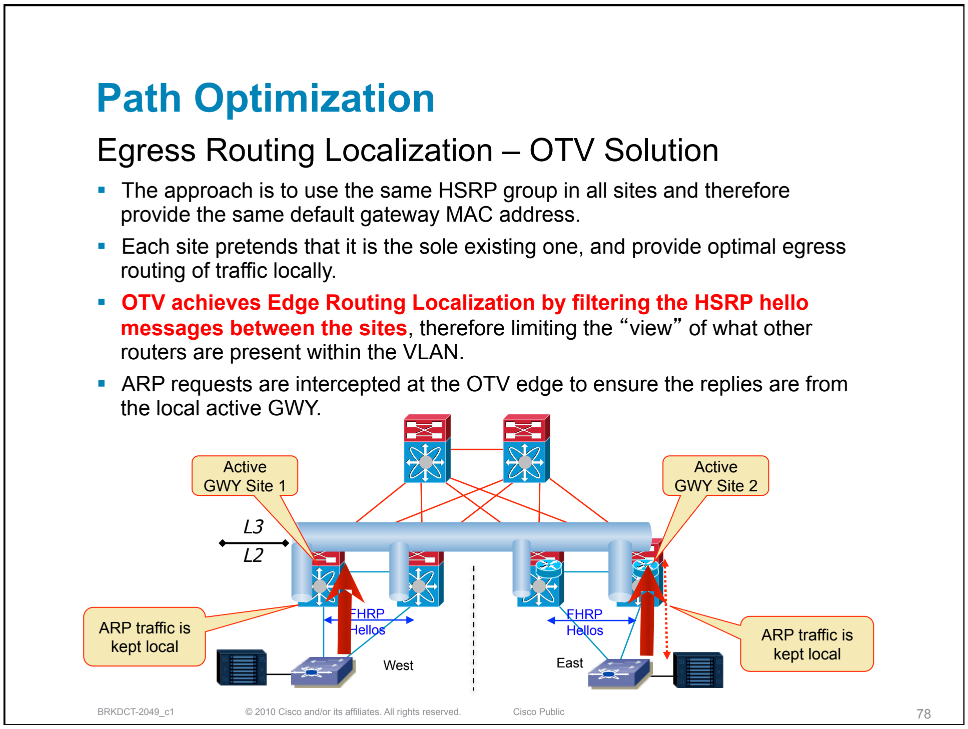


Path Optimization

Optimal Egress Routing Challenge

- Outbound routing of traffic, i.e. Server-Client or Server-Server traffic, is dependent on the location of the server' s default gateway.
- An extended subnet will have multiple IP gateway candidates distributed across sites. These gateways are all part of the same FHRP/HSRP group.

- Goals:
  - To enable site local egress routing, the default gateway must be present in the same site as the server is located.
  - For subnets/VLANs that stretch over multiple locations it means that each location has to have an active gateway.

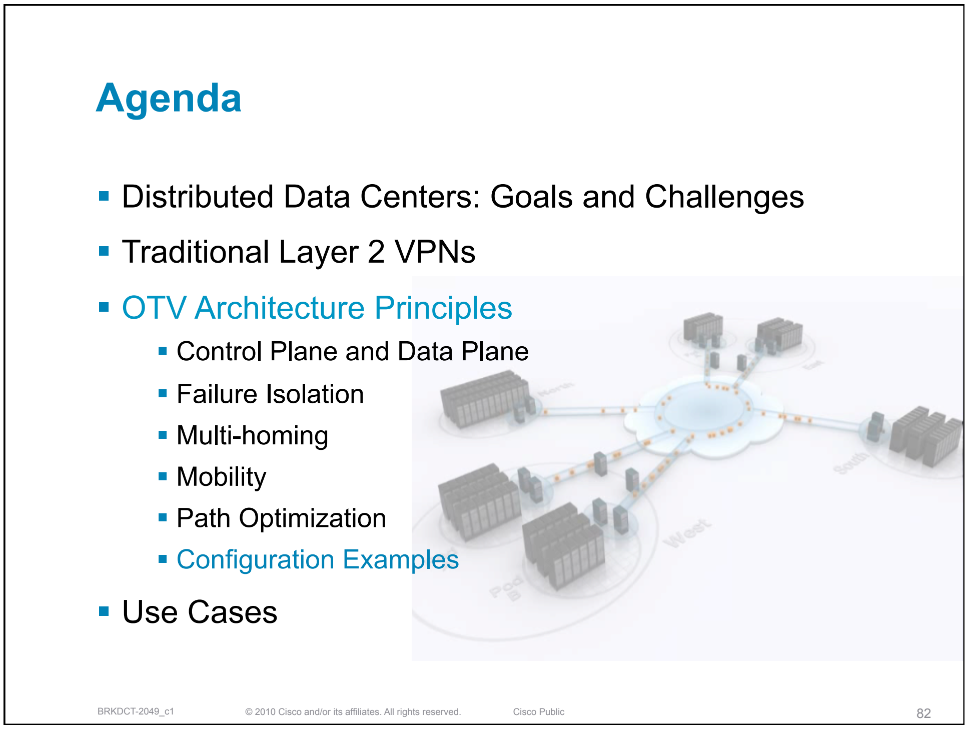


Path Optimization

Egress Routing Localization – OTV Solution

- The approach is to use the same HSRP group in all sites and therefore provide the same default gateway MAC address.
- Each site pretends that it is the sole existing one, and provide optimal egress routing of traffic locally.
- **OTV achieves Edge Routing Localization by filtering the HSRP hello messages between the sites**, therefore limiting the “view” of what other routers are present within the VLAN.
- ARP requests are intercepted at the OTV edge to ensure the replies are from the local active GWY.

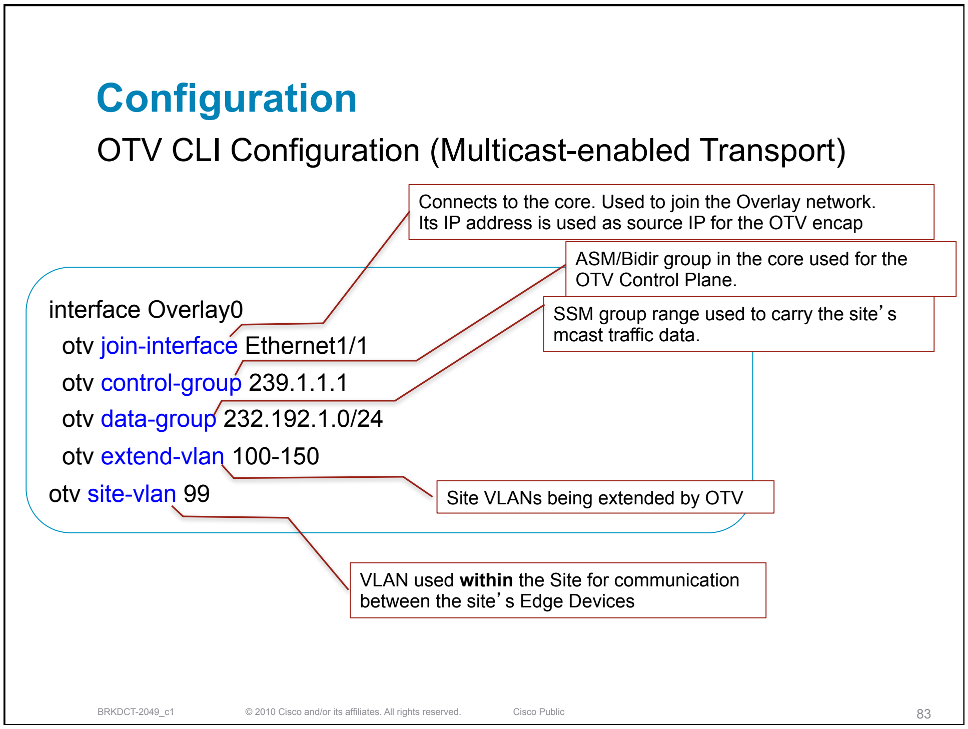
BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 78



Agenda

- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
  - Control Plane and Data Plane
  - Failure Isolation
  - Multi-homing
  - Mobility
  - Path Optimization
  - Configuration Examples
- Use Cases

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 82



Configuration

OTV CLI Configuration (Multicast-enabled Transport)

```
interface Overlay0
  otv join-interface Ethernet1/1
  otv control-group 239.1.1.1
  otv data-group 232.192.1.0/24
  otv extend-vlan 100-150
otv site-vlan 99
```

Connects to the core. Used to join the Overlay network. Its IP address is used as source IP for the OTV encap

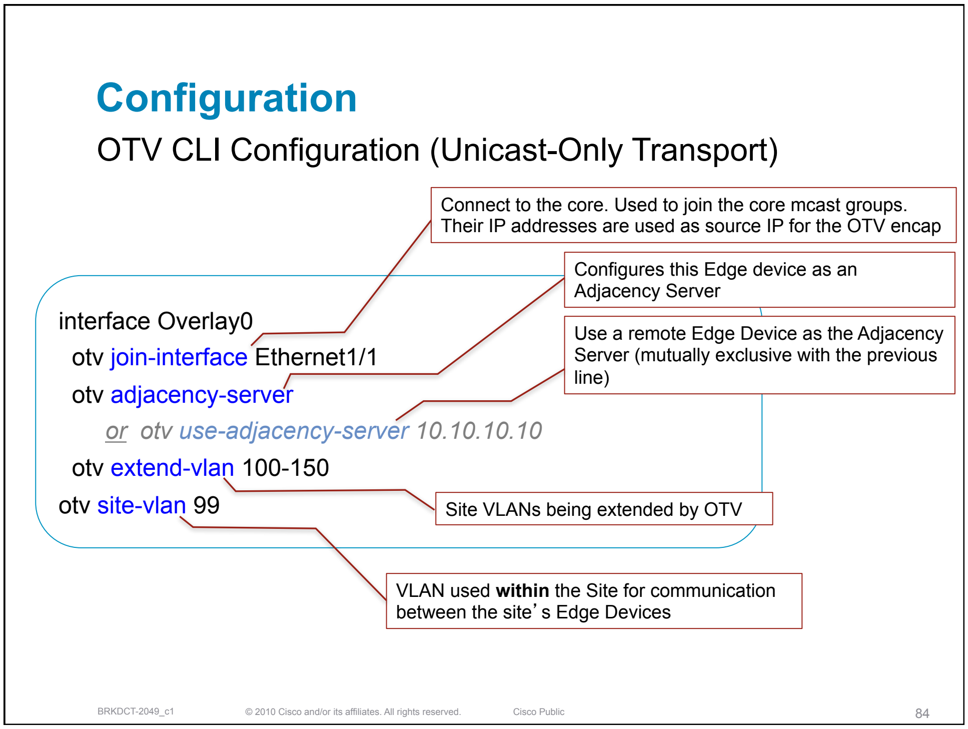
ASM/Bidir group in the core used for the OTV Control Plane.

SSM group range used to carry the site' s mcast traffic data.

Site VLANs being extended by OTV

VLAN used **within** the Site for communication between the site' s Edge Devices

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 83



Configuration

OTV CLI Configuration (Unicast-Only Transport)

```
interface Overlay0
  otv join-interface Ethernet1/1
  otv adjacency-server
     or  otv use-adjacency-server 10.10.10.10
  otv extend-vlan 100-150
otv site-vlan 99
```

Connect to the core. Used to join the core mcast groups. Their IP addresses are used as source IP for the OTV encap

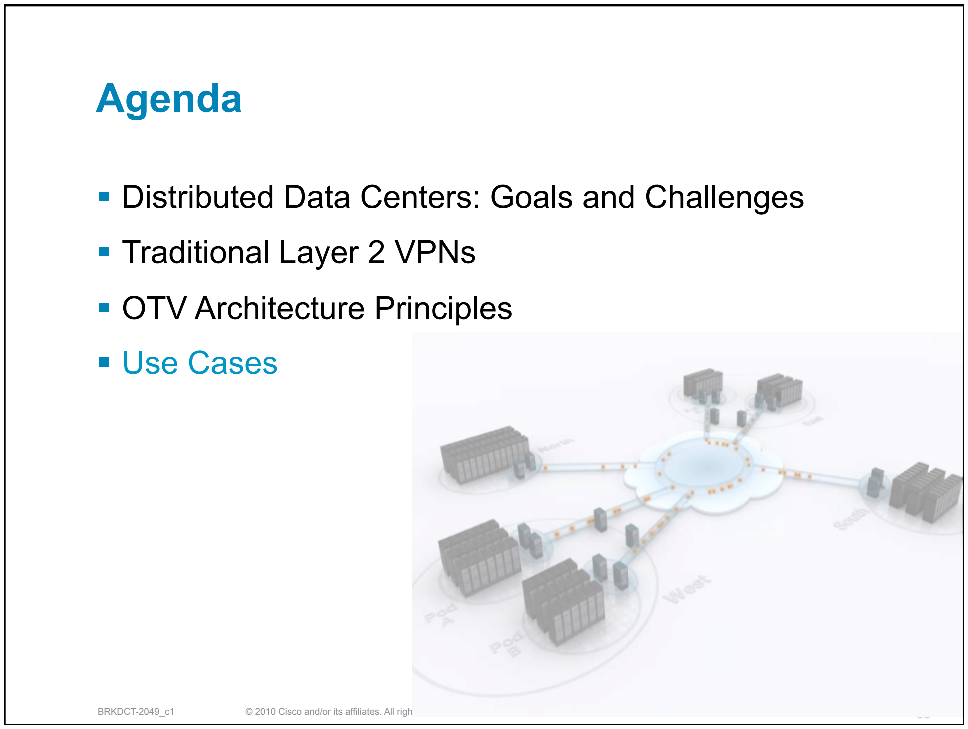
Configures this Edge device as an Adjacency Server

Use a remote Edge Device as the Adjacency Server (mutually exclusive with the previous line)

Site VLANs being extended by OTV

VLAN used **within** the Site for communication between the site' s Edge Devices

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 84

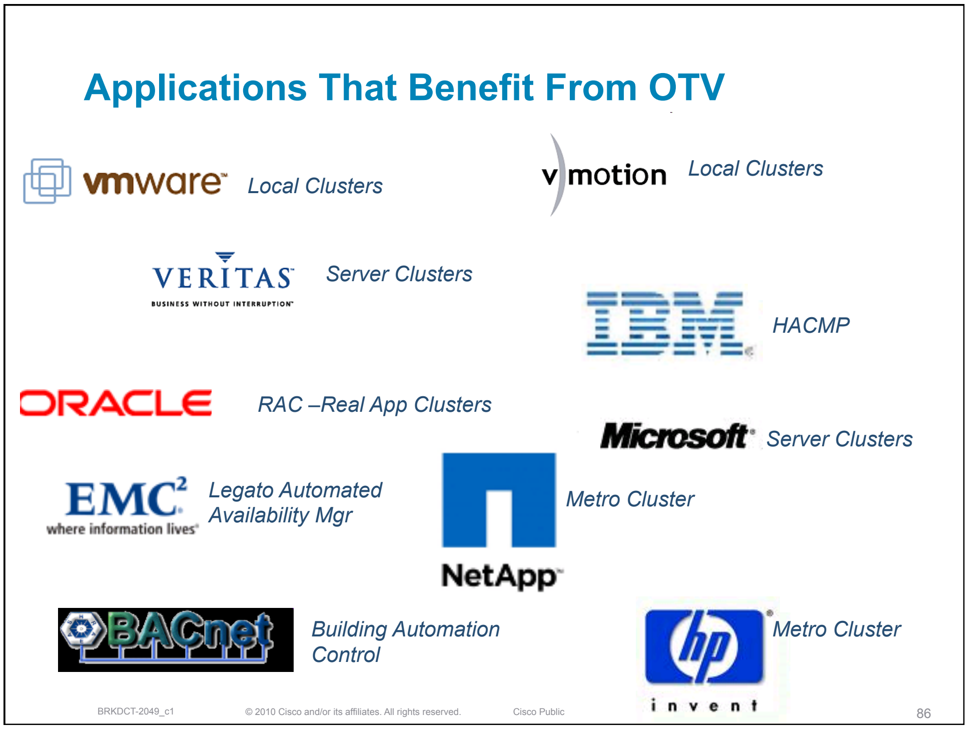


Agenda

- Distributed Data Centers: Goals and Challenges
- Traditional Layer 2 VPNs
- OTV Architecture Principles
- Use Cases

North
East
South
West
Pod A
Pod B

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All righ



Applications That Benefit From OTV

vmware Local Clusters

vmotion Local Clusters

VERITAS BUSINESS WITHOUT INTERRUPTION Server Clusters

IBM HACMP

ORACLE RAC –Real App Clusters

Microsoft Server Clusters

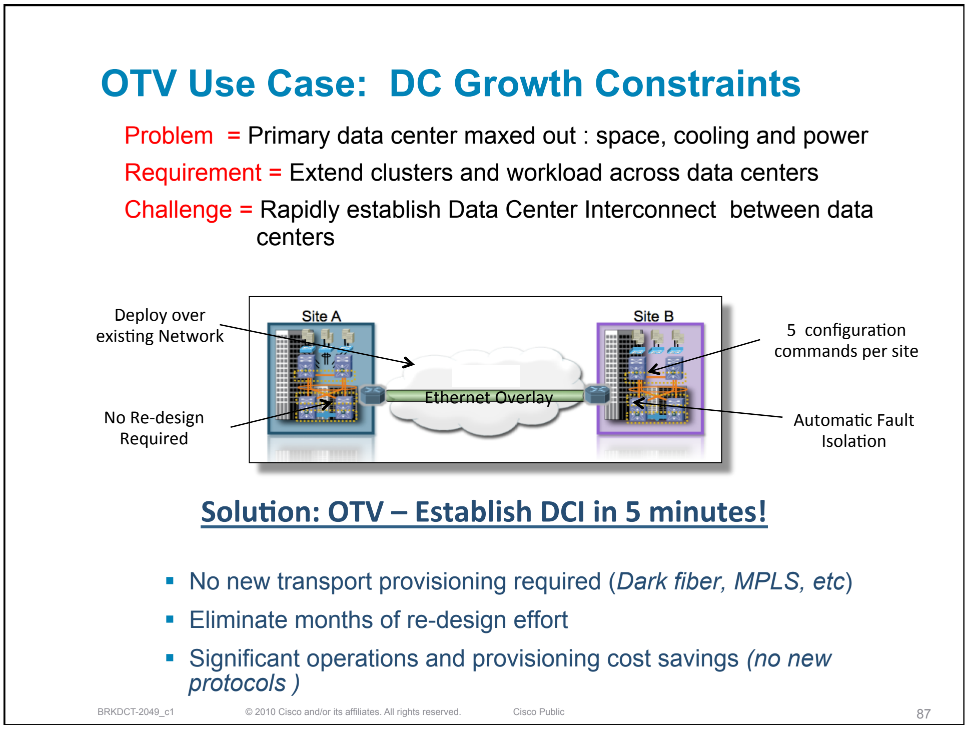
EMC² where information lives Legato Automated Availability Mgr

NetApp Metro Cluster

BACnet Building Automation Control

hp invent Metro Cluster

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 86



OTV Use Case: DC Growth Constraints

Problem = Primary data center maxed out : space, cooling and power

Requirement = Extend clusters and workload across data centers

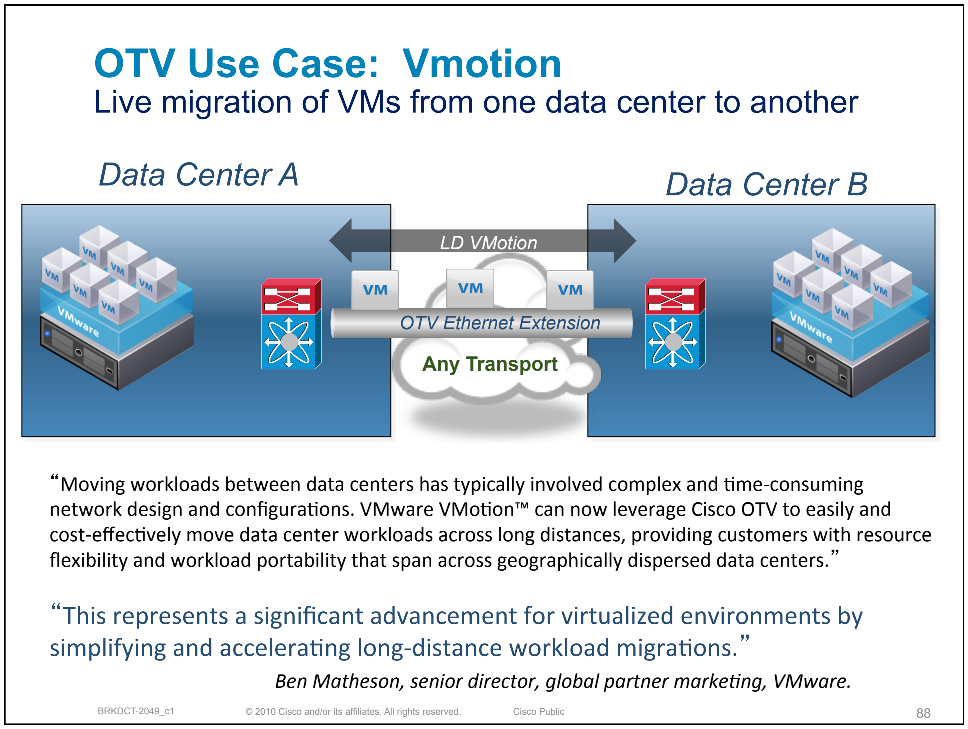
Challenge = Rapidly establish Data Center Interconnect between data centers

Deploy over existing Network
No Re-design Required
Site A
Ethernet Overlay
Site B
5 configuration commands per site
Automatic Fault Isolation

Solution: OTV – Establish DCI in 5 minutes!

- No new transport provisioning required (*Dark fiber, MPLS, etc*)
- Eliminate months of re-design effort
- Significant operations and provisioning cost savings *(no new protocols)*

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 87



OTV Use Case: Vmotion

Live migration of VMs from one data center to another

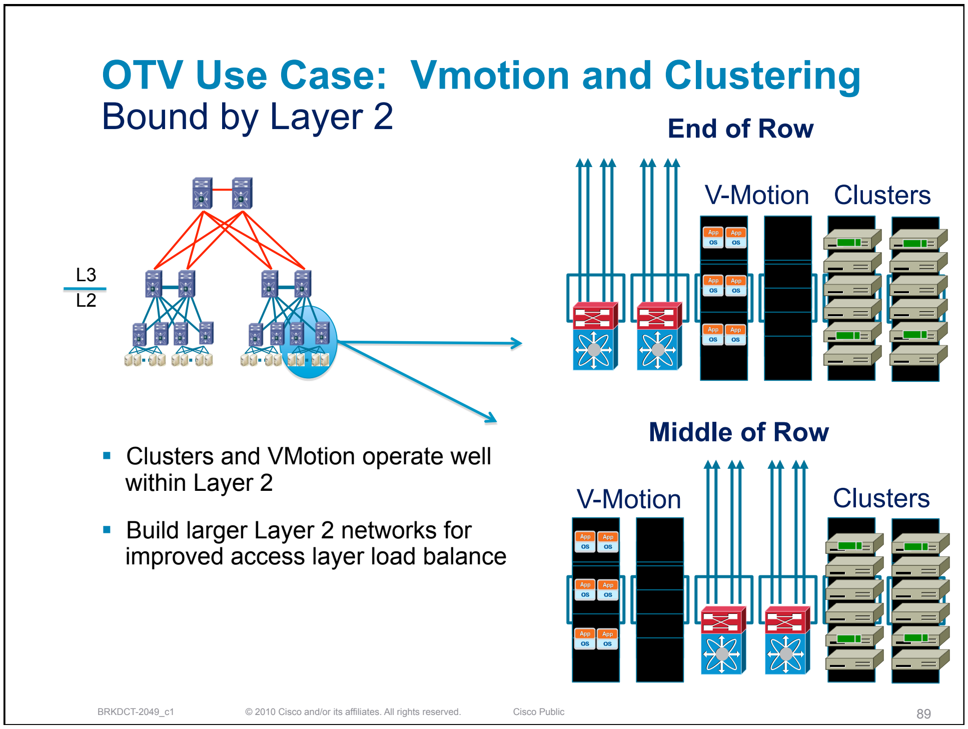
Data Center A
Data Center B
LD VMotion
VM
VM
VM
OTV Ethernet Extension
Any Transport
VMware
VM

"Moving workloads between data centers has typically involved complex and time-consuming network design and configurations. VMware VMotion™ can now leverage Cisco OTV to easily and cost-effectively move data center workloads across long distances, providing customers with resource flexibility and workload portability that span across geographically dispersed data centers."

"This represents a significant advancement for virtualized environments by simplifying and accelerating long-distance workload migrations."

Ben Matheson, senior director, global partner marketing, VMware.

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 88

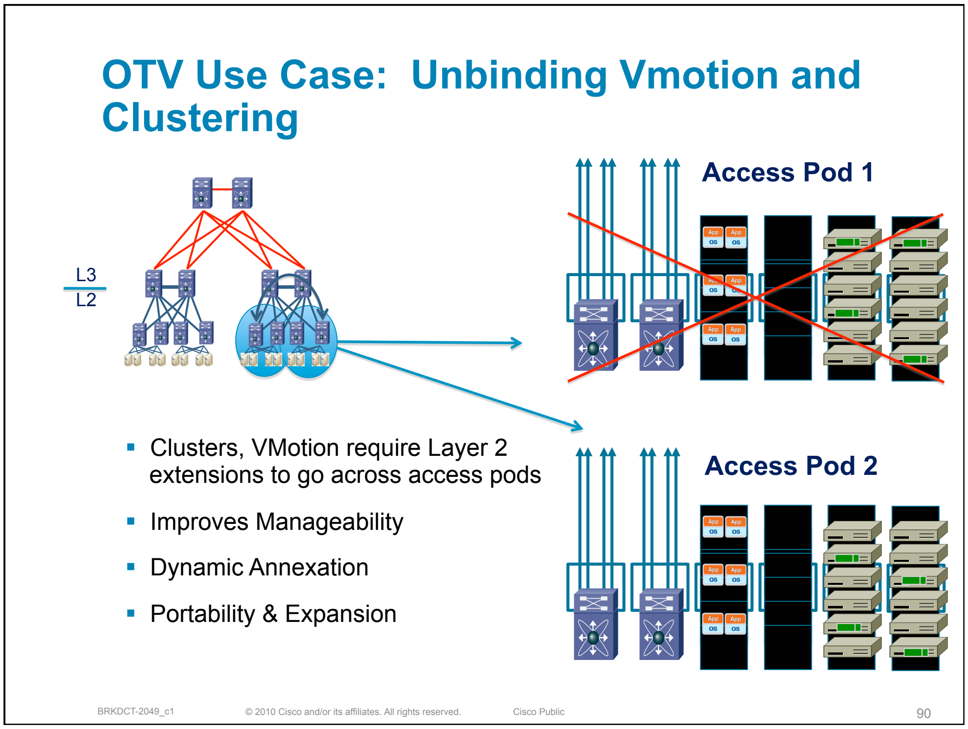


OTV Use Case: Vmotion and Clustering

Bound by Layer 2

- Clusters and VMotion operate well within Layer 2
- Build larger Layer 2 networks for improved access layer load balance

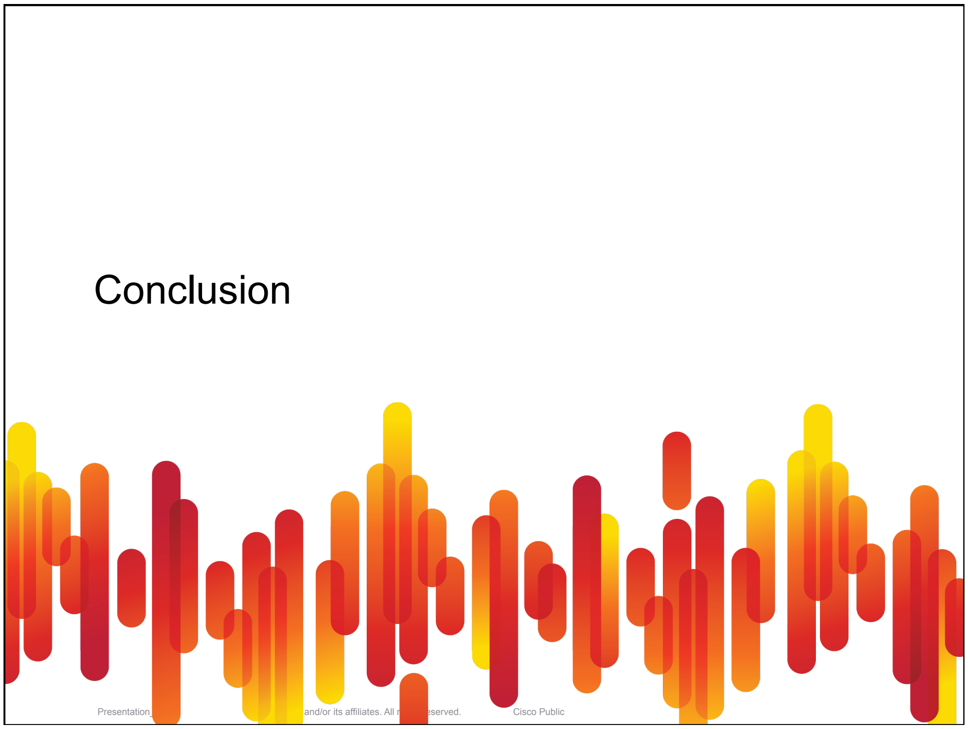
BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 89



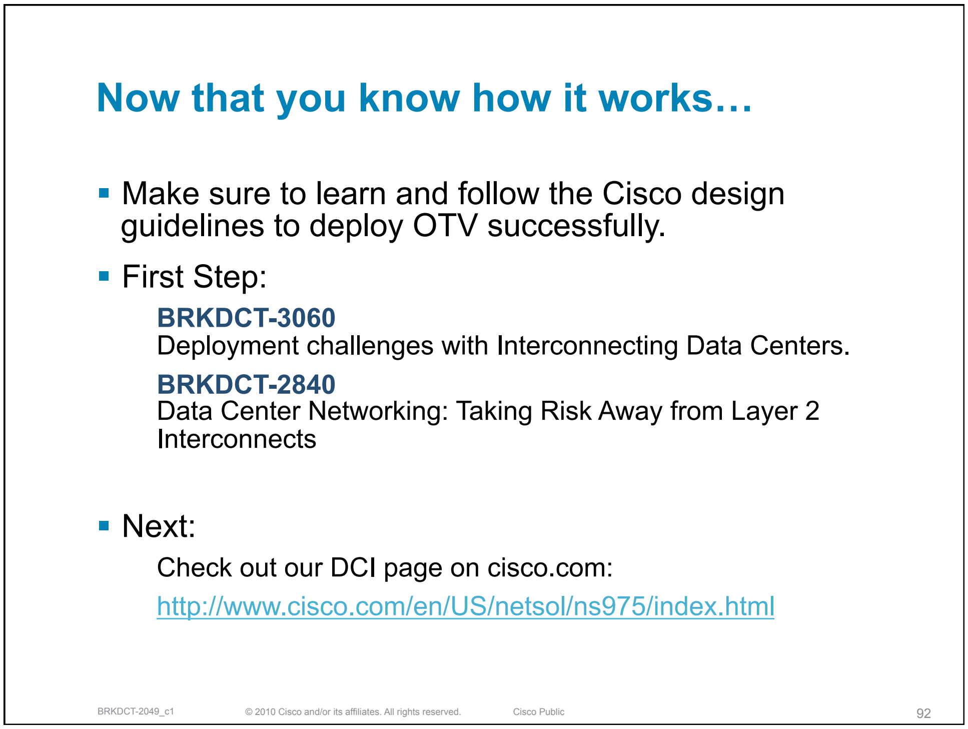
OTV Use Case: Unbinding Vmotion and Clustering

- Clusters, VMotion require Layer 2 extensions to go across access pods
- Improves Manageability
- Dynamic Annexation
- Portability & Expansion

BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 90



Conclusion



Now that you know how it works…

- Make sure to learn and follow the Cisco design guidelines to deploy OTV successfully.
- First Step:

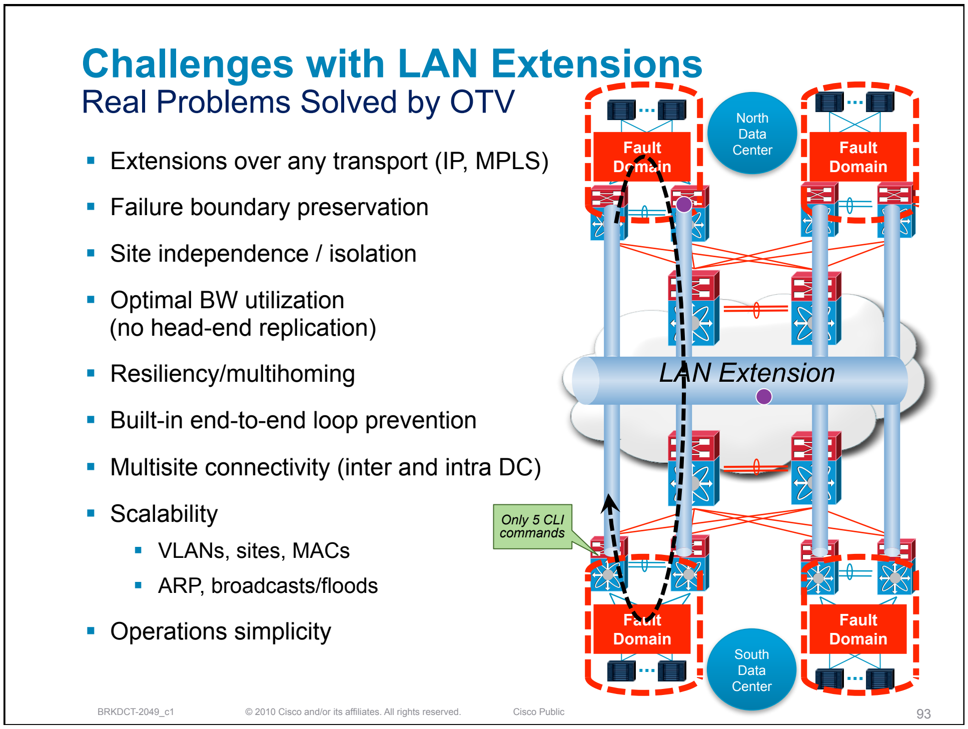
  BRKDCT-3060
  Deployment challenges with Interconnecting Data Centers.

  BRKDCT-2840
  Data Center Networking: Taking Risk Away from Layer 2 Interconnects

- Next:

  Check out our DCI page on cisco.com:

  http://www.cisco.com/en/US/netsol/ns975/index.html

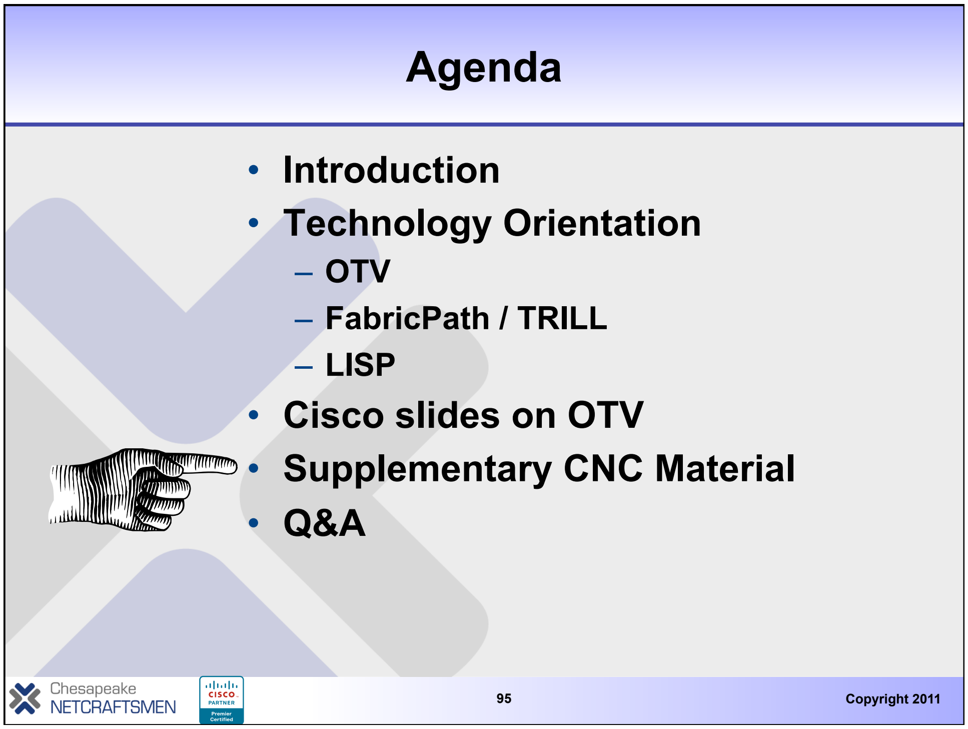


Challenges with LAN Extensions

Real Problems Solved by OTV

- Extensions over any transport (IP, MPLS)
- Failure boundary preservation
- Site independence / isolation
- Optimal BW utilization (no head-end replication)
- Resiliency/multihoming
- Built-in end-to-end loop prevention
- Multisite connectivity (inter and intra DC)
- Scalability
  - VLANs, sites, MACs
  - ARP, broadcasts/floods
- Operations simplicity

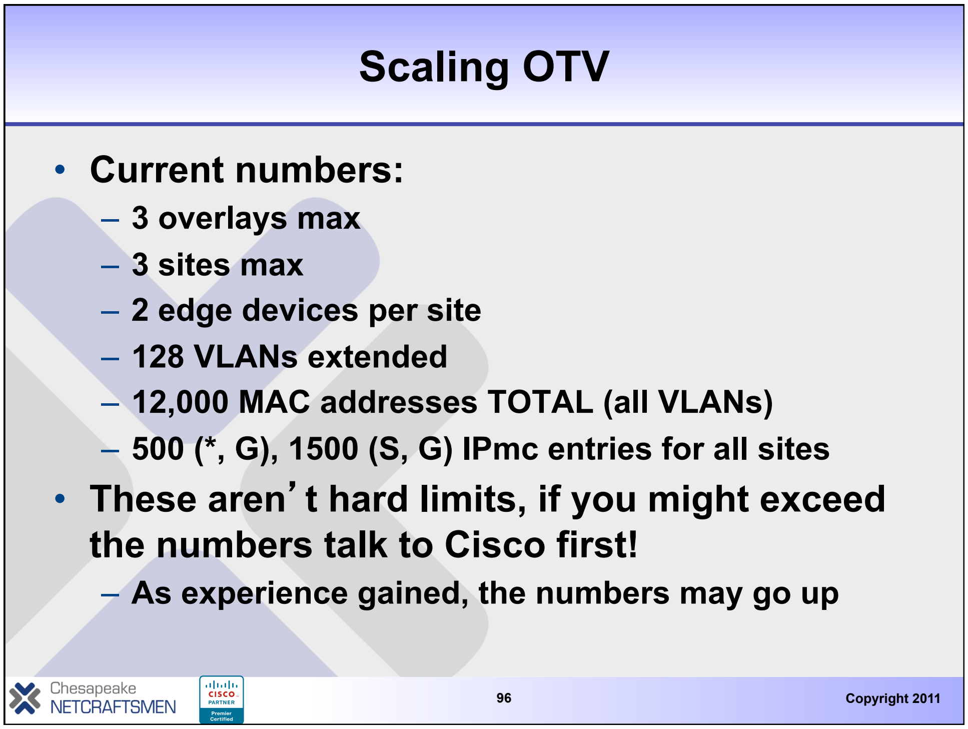
BRKDCT-2049_c1 © 2010 Cisco and/or its affiliates. All rights reserved. Cisco Public 93



Agenda

- Introduction
- Technology Orientation
  - OTV
  - FabricPath / TRILL
  - LISP
- Cisco slides on OTV
- Supplementary CNC Material
- Q&A

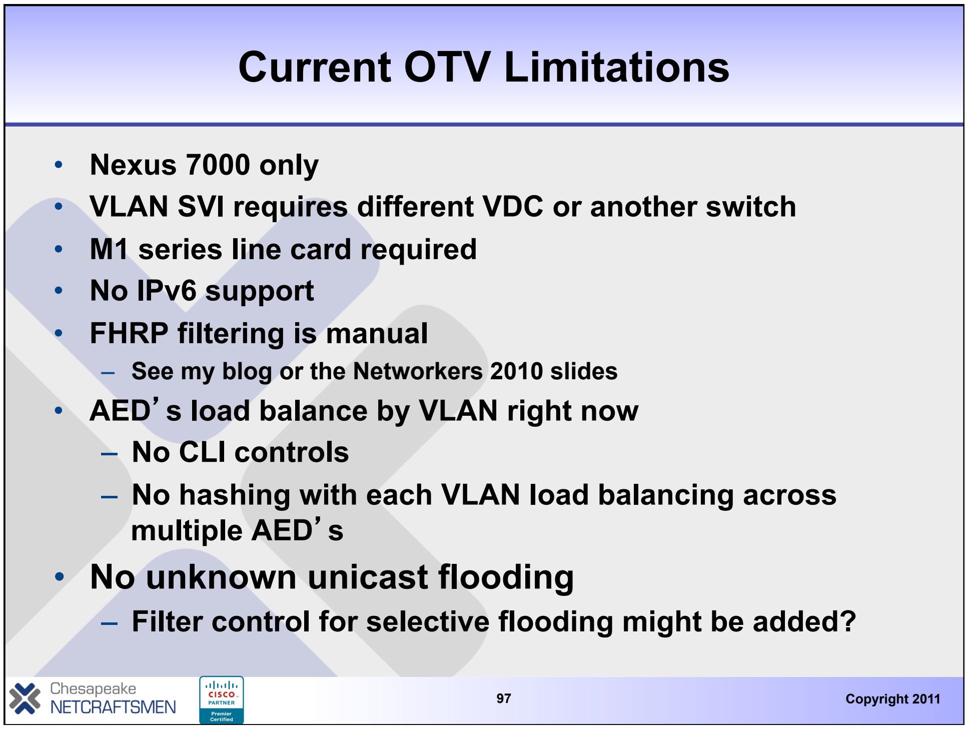
Chesapeake NETCRAFTSMEN 95 Copyright 2011



Scaling OTV

- Current numbers:
  - 3 overlays max
  - 3 sites max
  - 2 edge devices per site
  - 128 VLANs extended
  - 12,000 MAC addresses TOTAL (all VLANs)
  - 500 (*, G), 1500 (S, G) IPmc entries for all sites
- These aren't hard limits, if you might exceed the numbers talk to Cisco first!
  - As experience gained, the numbers may go up

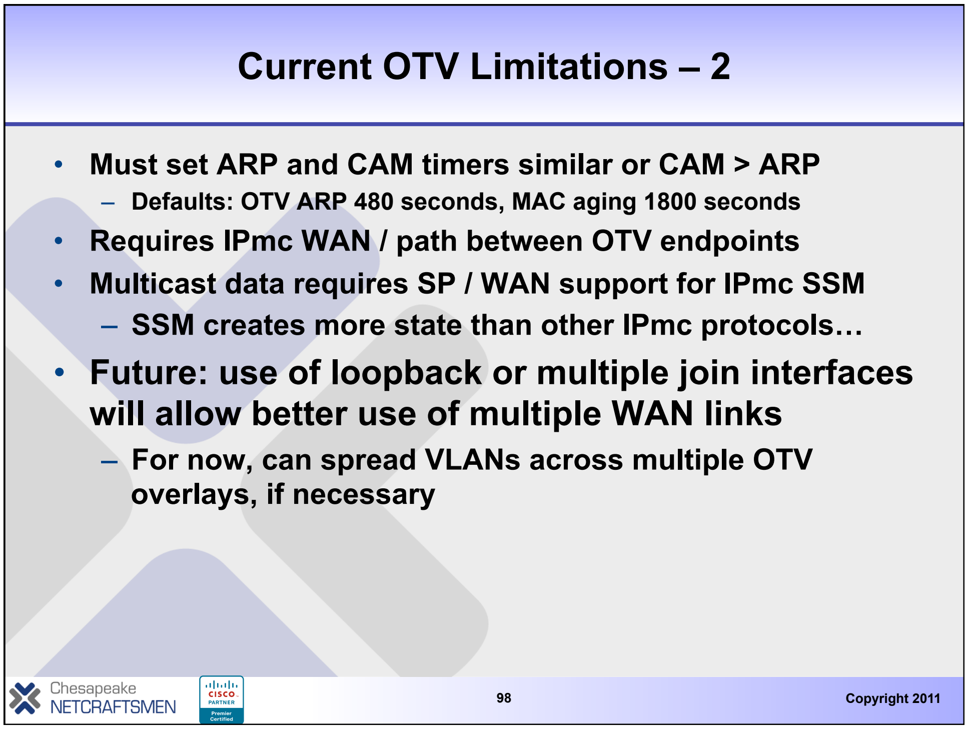
Chesapeake NETCRAFTSMEN 96 Copyright 2011



Current OTV Limitations

- Nexus 7000 only
- VLAN SVI requires different VDC or another switch
- M1 series line card required
- No IPv6 support
- FHRP filtering is manual
  - See my blog or the Networkers 2010 slides
- AED's load balance by VLAN right now
  - No CLI controls
  - No hashing with each VLAN load balancing across multiple AED's
- No unknown unicast flooding
  - Filter control for selective flooding might be added?

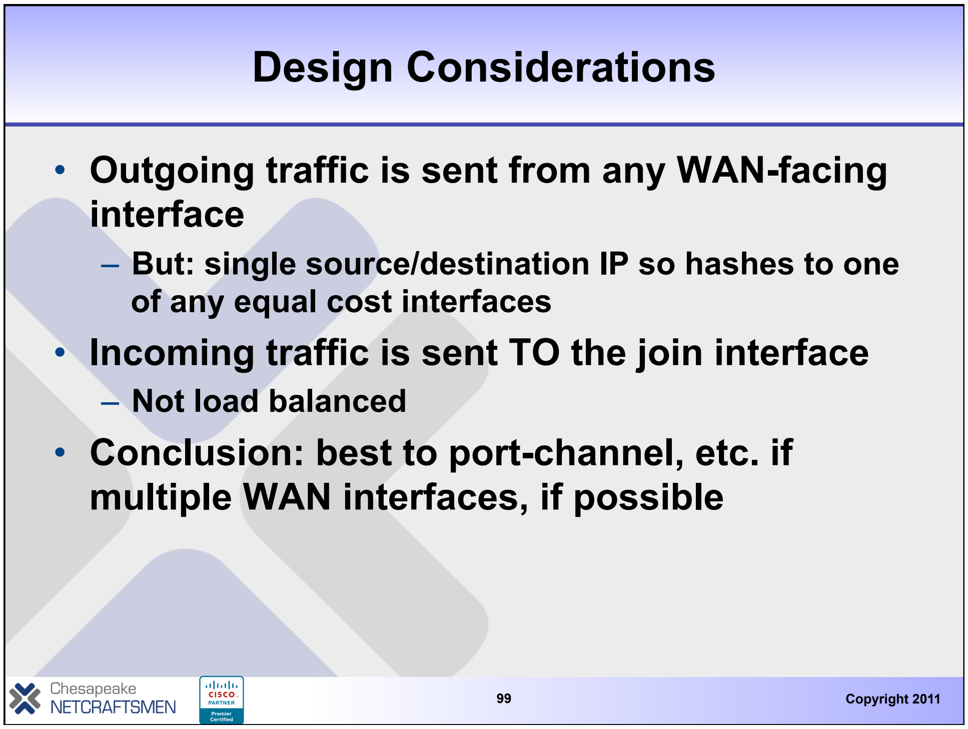
Chesapeake NETCRAFTSMEN 97 Copyright 2011



Current OTV Limitations – 2

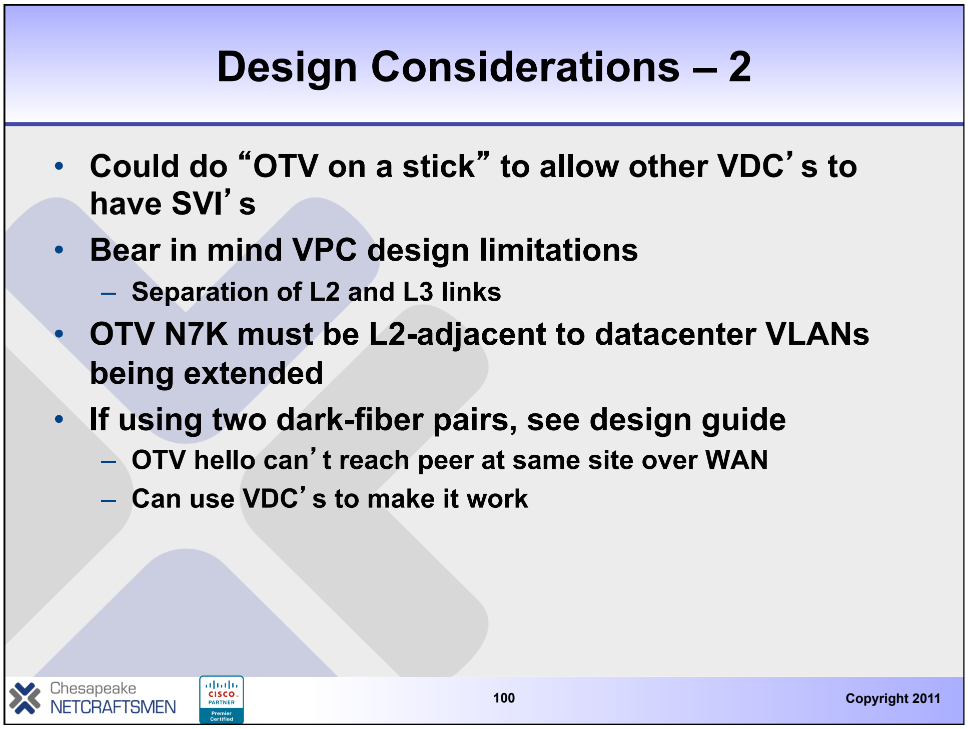
- Must set ARP and CAM timers similar or CAM > ARP
  - Defaults: OTV ARP 480 seconds, MAC aging 1800 seconds
- Requires IPmc WAN / path between OTV endpoints
- Multicast data requires SP / WAN support for IPmc SSM
  - SSM creates more state than other IPmc protocols…
- Future: use of loopback or multiple join interfaces will allow better use of multiple WAN links
  - For now, can spread VLANs across multiple OTV overlays, if necessary

Chesapeake NETCRAFTSMEN 98 Copyright 2011



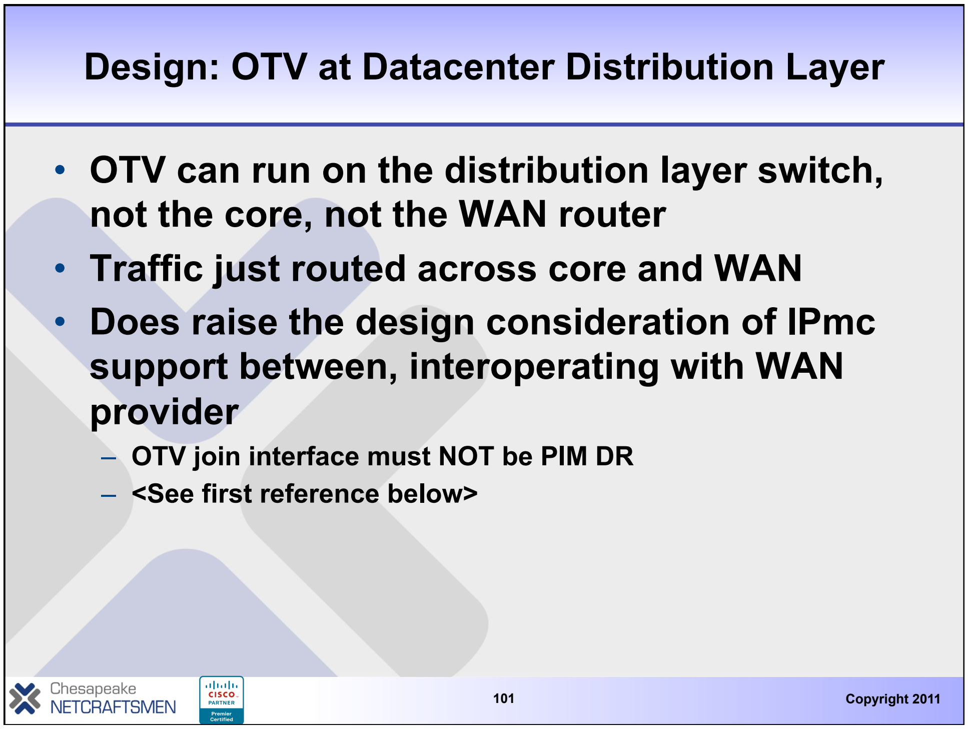
Design Considerations

- Outgoing traffic is sent from any WAN-facing interface
  - But: single source/destination IP so hashes to one of any equal cost interfaces
- Incoming traffic is sent TO the join interface
  - Not load balanced
- Conclusion: best to port-channel, etc. if multiple WAN interfaces, if possible



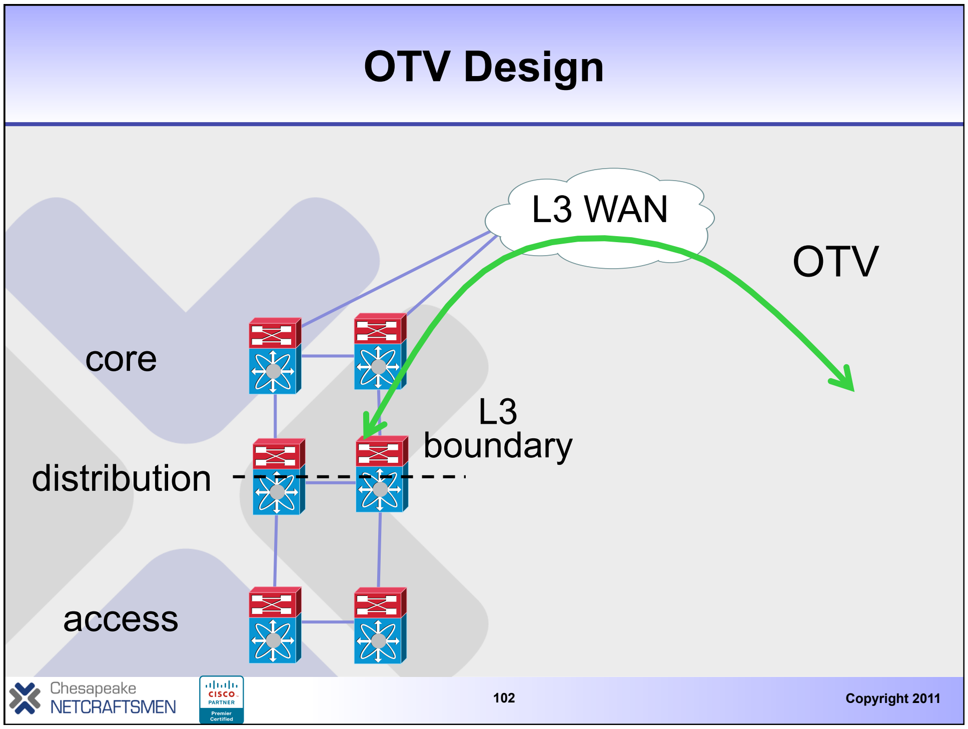
Design Considerations – 2

- Could do “OTV on a stick” to allow other VDC’s to have SVI’s
- Bear in mind VPC design limitations
  - Separation of L2 and L3 links
- OTV N7K must be L2-adjacent to datacenter VLANs being extended
- If using two dark-fiber pairs, see design guide
  - OTV hello can’t reach peer at same site over WAN
  - Can use VDC’s to make it work

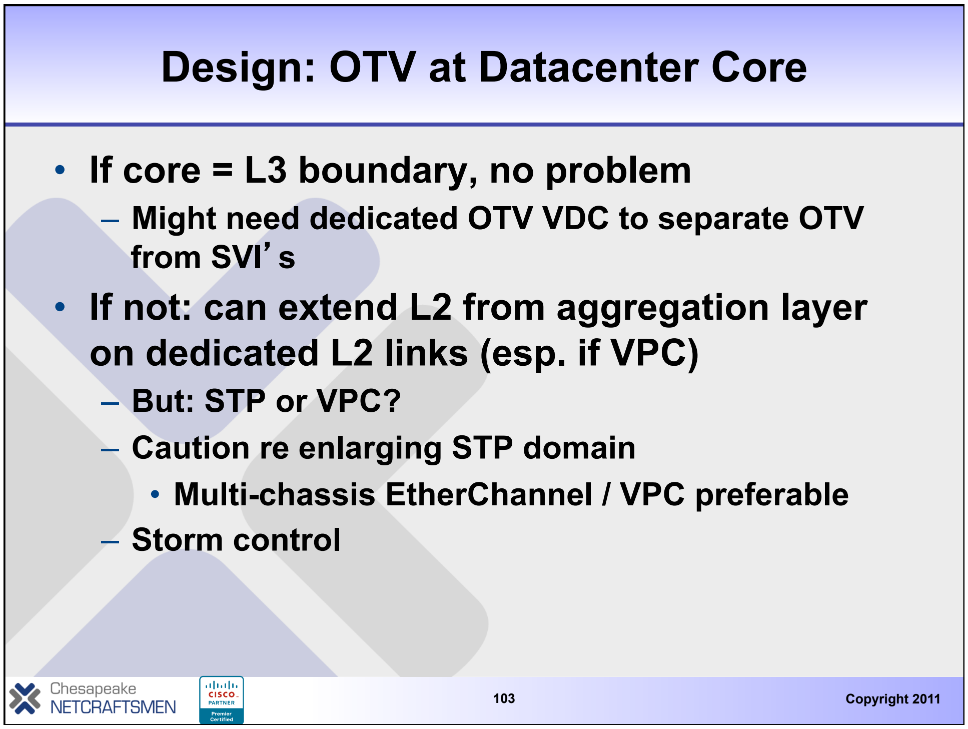


Design: OTV at Datacenter Distribution Layer

- OTV can run on the distribution layer switch, not the core, not the WAN router
- Traffic just routed across core and WAN
- Does raise the design consideration of IPmc support between, interoperating with WAN provider
  - OTV join interface must NOT be PIM DR
  - <See first reference below>



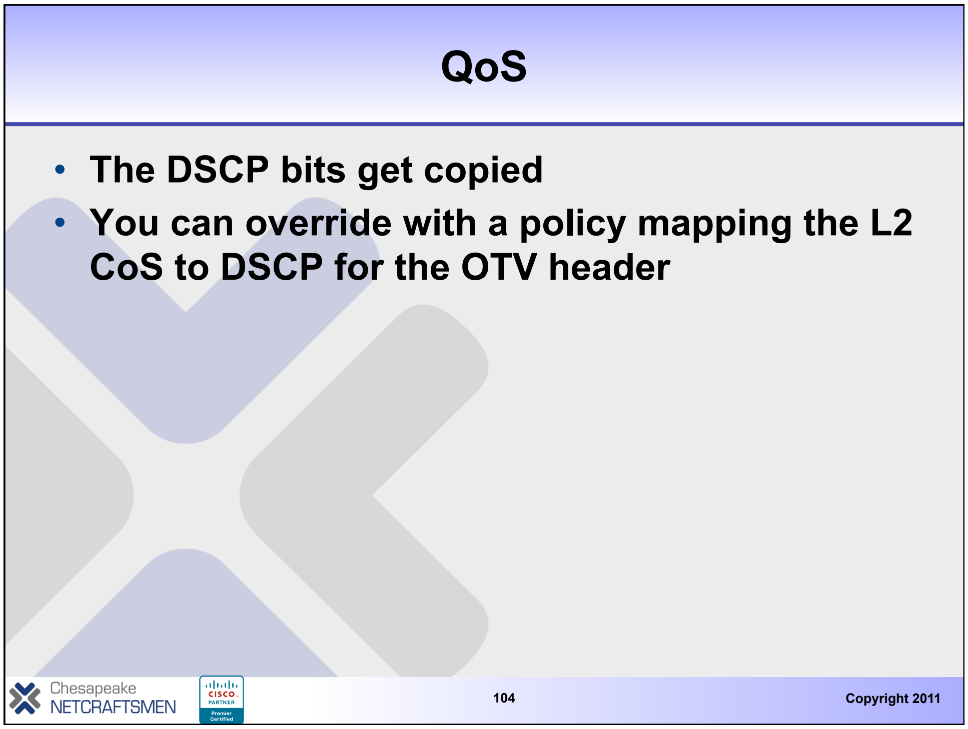
OTV Design



Design: OTV at Datacenter Core

- If core = L3 boundary, no problem
  - Might need dedicated OTV VDC to separate OTV from SVI’ s
- If not: can extend L2 from aggregation layer on dedicated L2 links (esp. if VPC)
  - But: STP or VPC?
  - Caution re enlarging STP domain
    - Multi-chassis EtherChannel / VPC preferable
  - Storm control

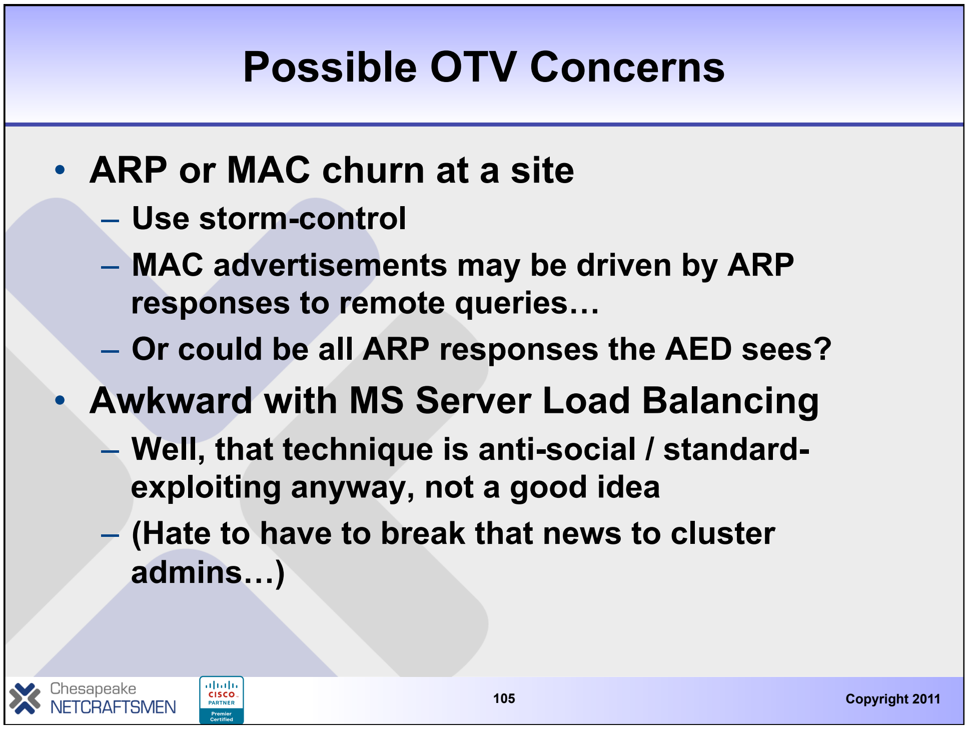
Chesapeake NETCRAFTSMEN 103 Copyright 2011



QoS

- The DSCP bits get copied
- You can override with a policy mapping the L2 CoS to DSCP for the OTV header

Chesapeake NETCRAFTSMEN 104 Copyright 2011



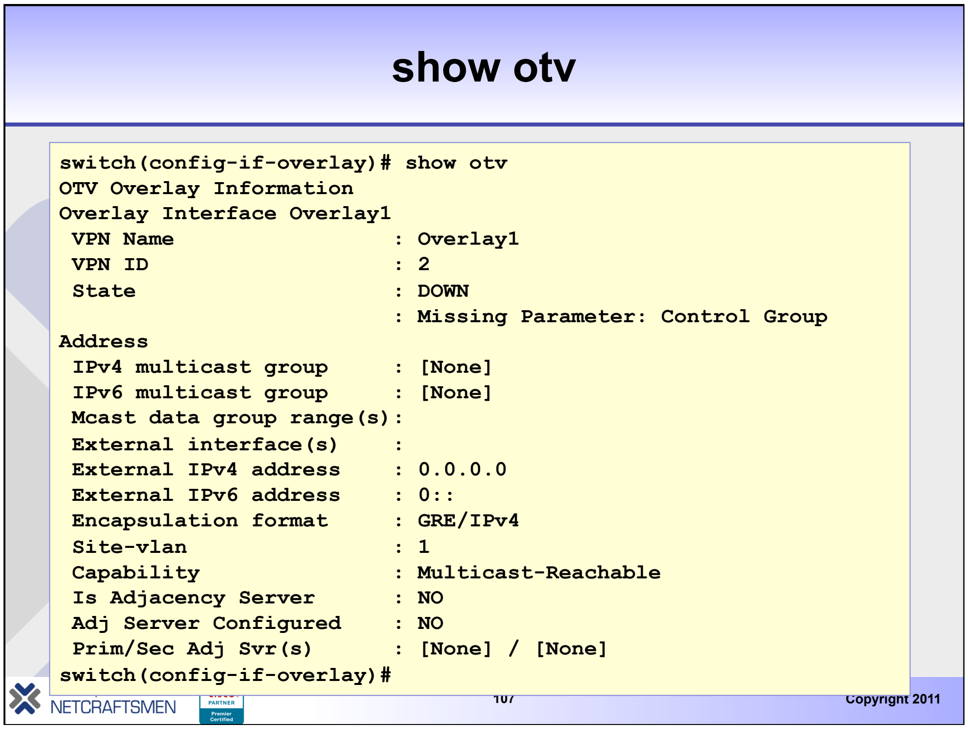
Possible OTV Concerns

- ARP or MAC churn at a site
  - Use storm-control
  - MAC advertisements may be driven by ARP responses to remote queries…
  - Or could be all ARP responses the AED sees?
- Awkward with MS Server Load Balancing
  - Well, that technique is anti-social / standard-exploiting anyway, not a good idea
  - (Hate to have to break that news to cluster admins…)

Chesapeake NETCRAFTSMEN 105 Copyright 2011

SAMPLE SHOW COMMANDS

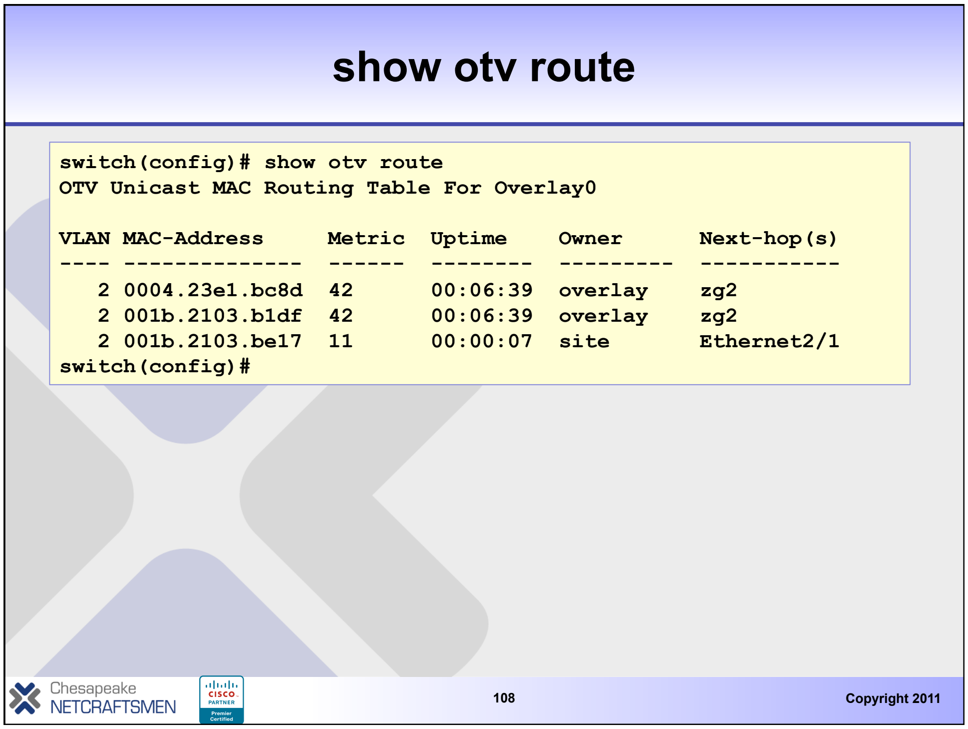
Chesapeake NETCRAFTSMEN 106 Copyright 2011



show otv

```
switch(config-if-overlay)# show otv
OTV Overlay Information
Overlay Interface Overlay1
 VPN Name                  : Overlay1
 VPN ID                    : 2
 State                     : DOWN
                           : Missing Parameter: Control Group
Address
 IPv4 multicast group      : [None]
 IPv6 multicast group      : [None]
 Mcast data group range(s):
 External interface(s)     :
 External IPv4 address     : 0.0.0.0
 External IPv6 address     : 0::
 Encapsulation format      : GRE/IPv4
 Site-vlan                 : 1
 Capability                : Multicast-Reachable
 Is Adjacency Server       : NO
 Adj Server Configured     : NO
 Prim/Sec Adj Svr(s)       : [None] / [None]
switch(config-if-overlay)#
```

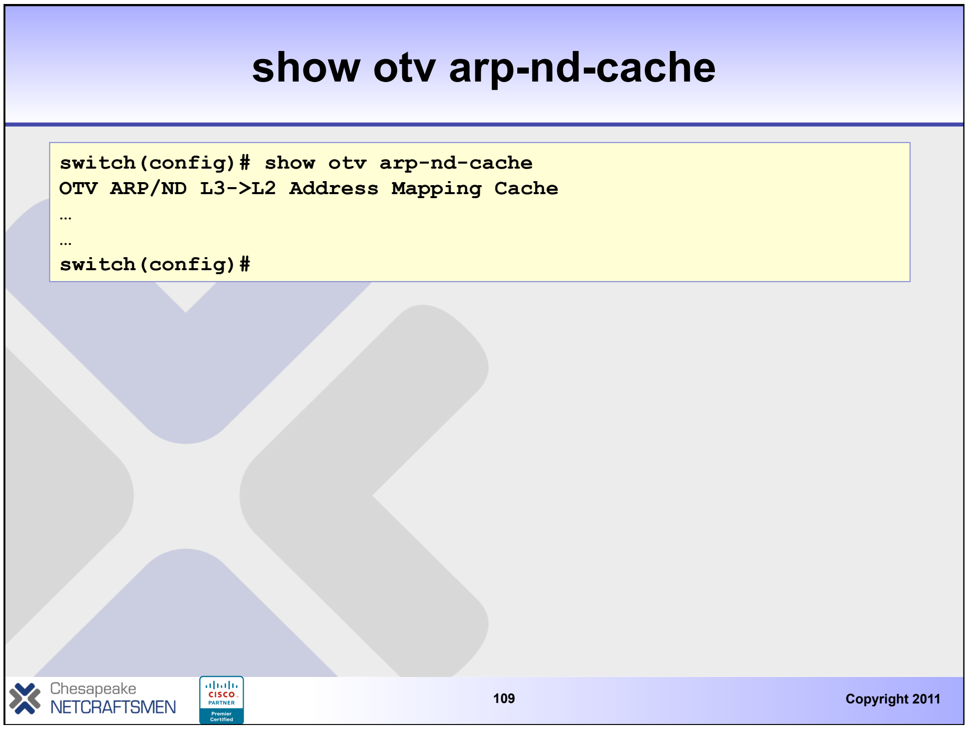
NETCRAFTSMEN 107 Copyright 2011



show otv route

```
switch(config)# show otv route
OTV Unicast MAC Routing Table For Overlay0

VLAN MAC-Address     Metric Uptime    Owner     Next-hop(s)
---- -------------- ------ -------- --------- -----------
   2 0004.23e1.bc8d 42     00:06:39 overlay   zg2
   2 001b.2103.b1df 42     00:06:39 overlay   zg2
   2 001b.2103.be17 11     00:00:07 site      Ethernet2/1
switch(config)#
```

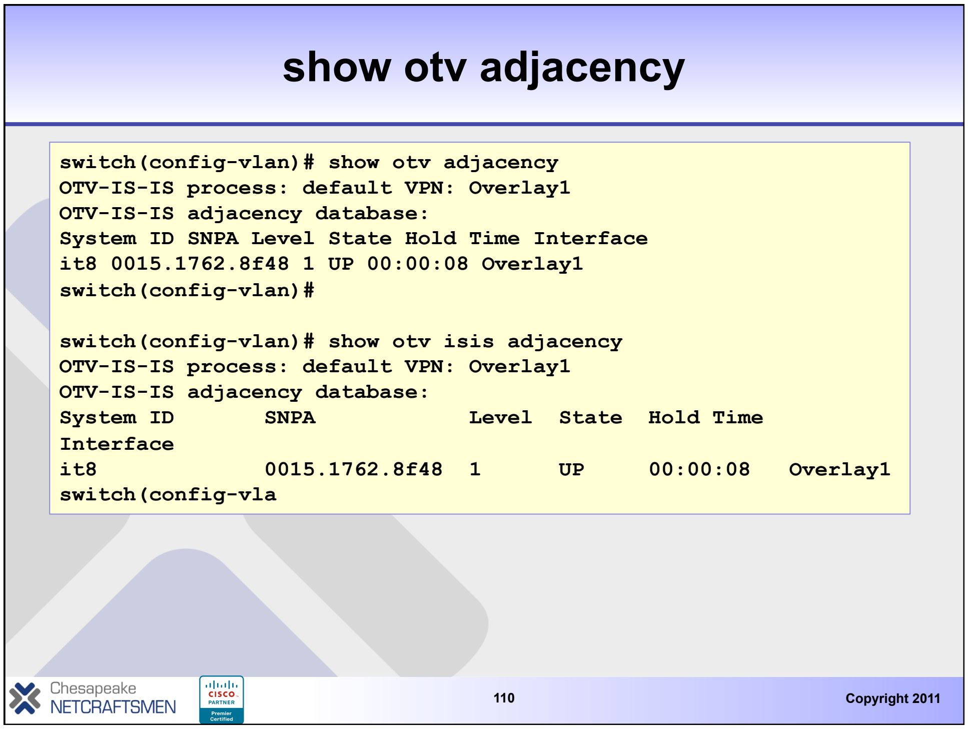
Chesapeake NETCRAFTSMEN 108 Copyright 2011



show otv arp-nd-cache

```
switch(config)# show otv arp-nd-cache
OTV ARP/ND L3->L2 Address Mapping Cache
…
…
switch(config)#
```

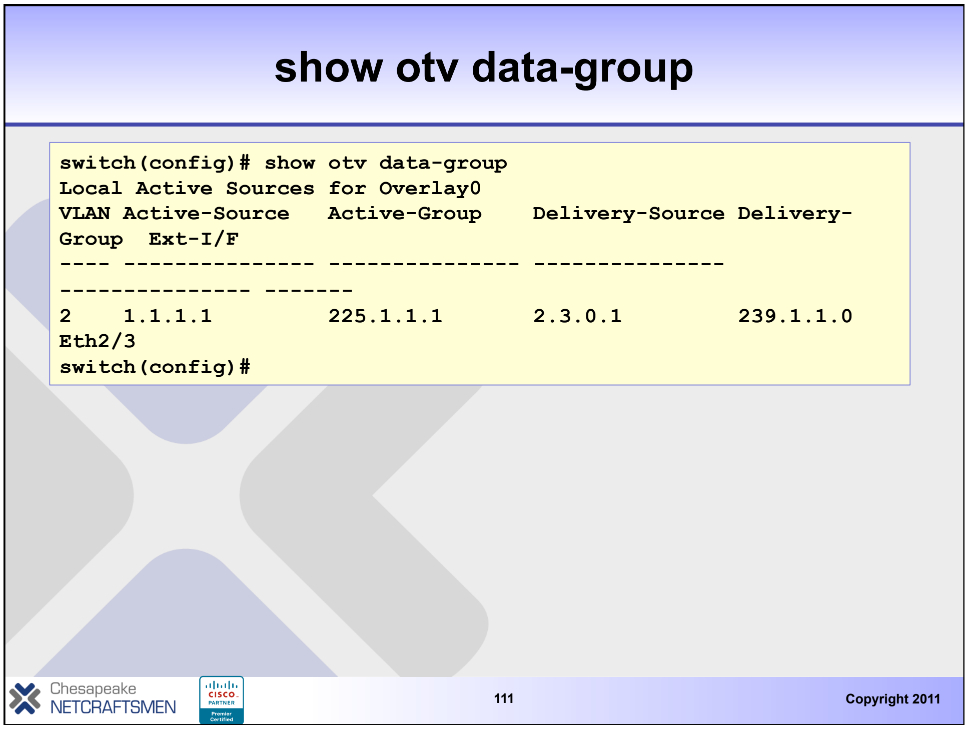
Chesapeake NETCRAFTSMEN 109 Copyright 2011



show otv adjacency

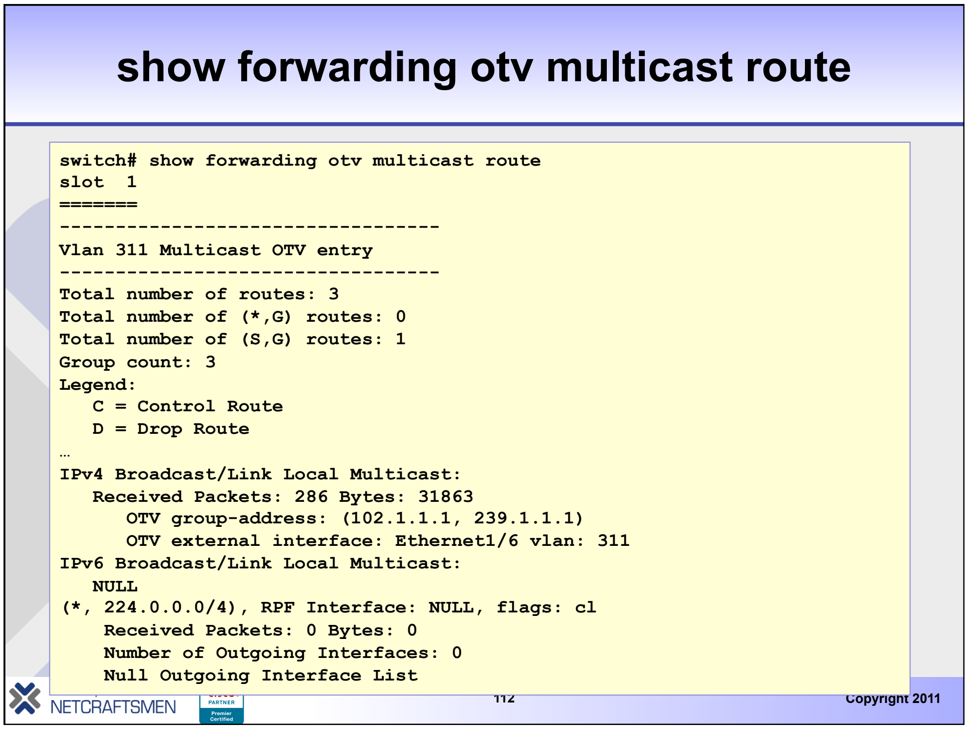
```
switch(config-vlan)# show otv adjacency
OTV-IS-IS process: default VPN: Overlay1
OTV-IS-IS adjacency database:
System ID SNPA Level State Hold Time Interface
it8 0015.1762.8f48 1 UP 00:00:08 Overlay1
switch(config-vlan)#

switch(config-vlan)# show otv isis adjacency
OTV-IS-IS process: default VPN: Overlay1
OTV-IS-IS adjacency database:
System ID       SNPA            Level  State  Hold Time
Interface
it8             0015.1762.8f48  1      UP     00:00:08   Overlay1
switch(config-vla
```

Chesapeake NETCRAFTSMEN 110 Copyright 2011



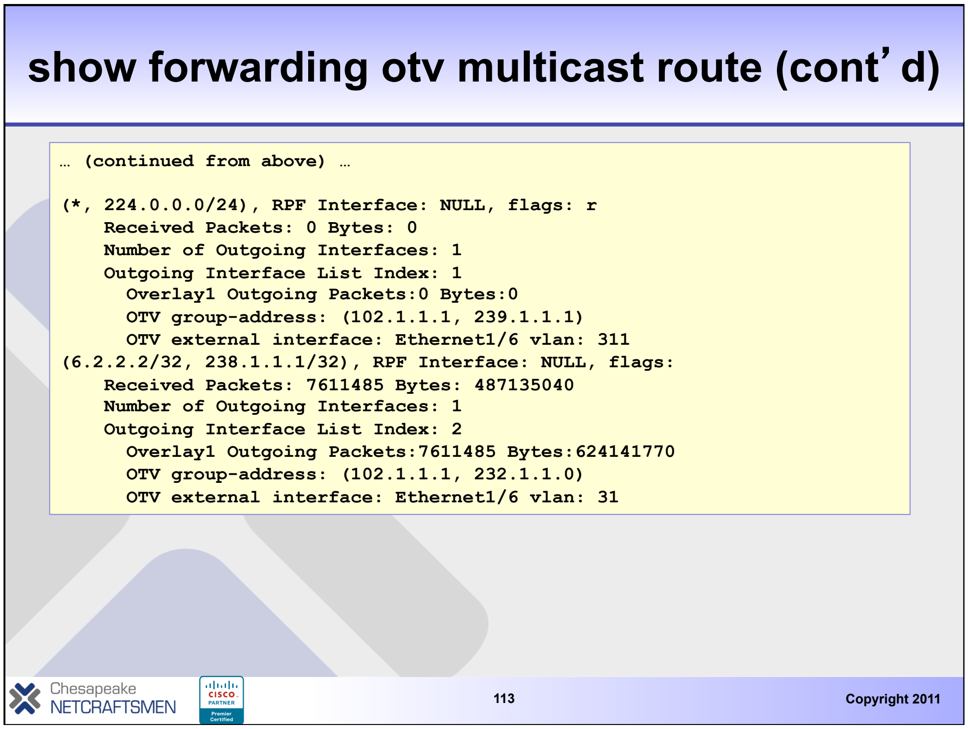
show otv data-group

```
switch(config)# show otv data-group
Local Active Sources for Overlay0
VLAN Active-Source   Active-Group    Delivery-Source Delivery-
Group  Ext-I/F
---- --------------- --------------- ---------------
--------------- -------
2    1.1.1.1         225.1.1.1       2.3.0.1         239.1.1.0
Eth2/3
switch(config)#
```



show forwarding otv multicast route

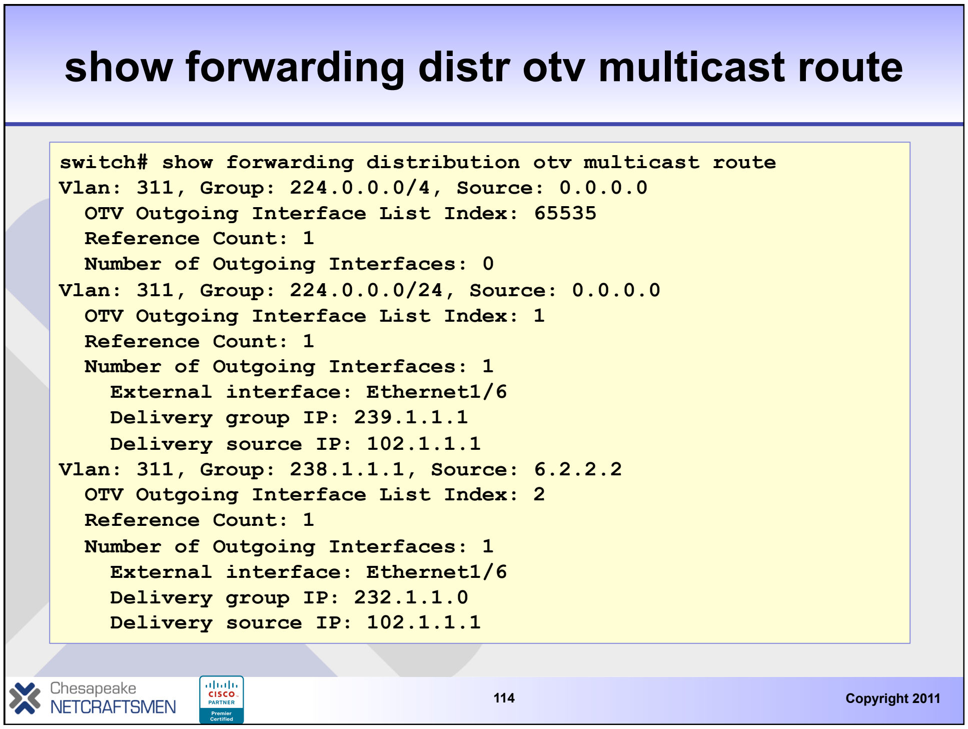
```
switch# show forwarding otv multicast route
slot  1
=======
----------------------------------
Vlan 311 Multicast OTV entry
----------------------------------
Total number of routes: 3
Total number of (*,G) routes: 0
Total number of (S,G) routes: 1
Group count: 3
Legend:
   C = Control Route
   D = Drop Route
…
IPv4 Broadcast/Link Local Multicast:
   Received Packets: 286 Bytes: 31863
      OTV group-address: (102.1.1.1, 239.1.1.1)
      OTV external interface: Ethernet1/6 vlan: 311
IPv6 Broadcast/Link Local Multicast:
   NULL
(*, 224.0.0.0/4), RPF Interface: NULL, flags: cl
    Received Packets: 0 Bytes: 0
    Number of Outgoing Interfaces: 0
    Null Outgoing Interface List
```



show forwarding otv multicast route (cont’d)

```
… (continued from above) …

(*, 224.0.0.0/24), RPF Interface: NULL, flags: r
    Received Packets: 0 Bytes: 0
    Number of Outgoing Interfaces: 1
    Outgoing Interface List Index: 1
      Overlay1 Outgoing Packets:0 Bytes:0
      OTV group-address: (102.1.1.1, 239.1.1.1)
      OTV external interface: Ethernet1/6 vlan: 311
(6.2.2.2/32, 238.1.1.1/32), RPF Interface: NULL, flags:
    Received Packets: 7611485 Bytes: 487135040
    Number of Outgoing Interfaces: 1
    Outgoing Interface List Index: 2
      Overlay1 Outgoing Packets:7611485 Bytes:624141770
      OTV group-address: (102.1.1.1, 232.1.1.0)
      OTV external interface: Ethernet1/6 vlan: 31
```

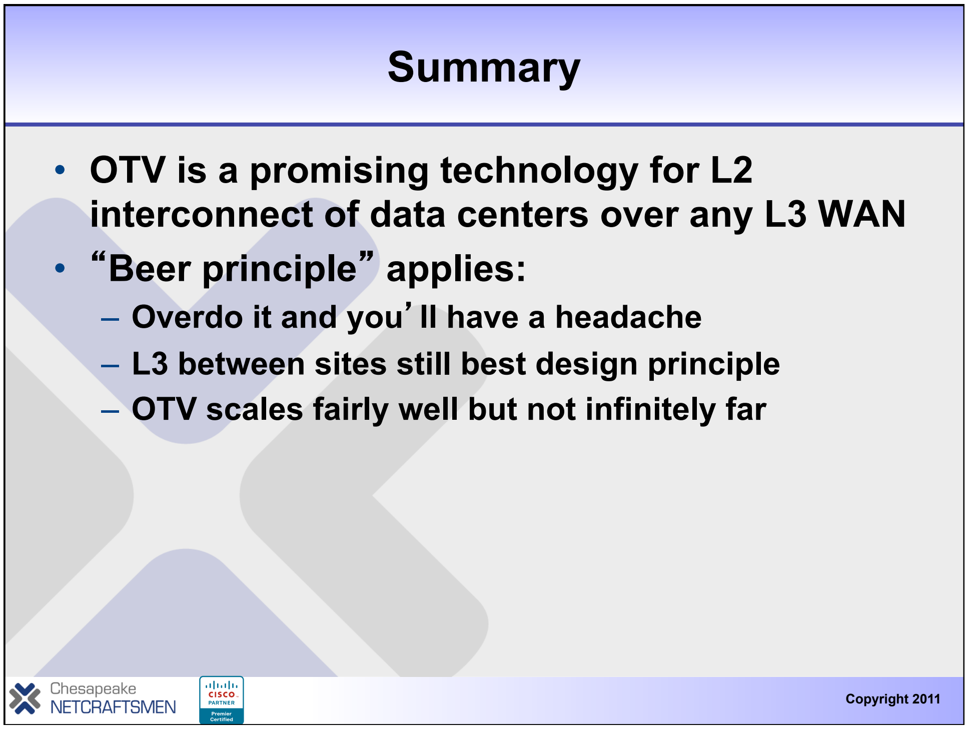
Chesapeake NETCRAFTSMEN 113 Copyright 2011



show forwarding distr otv multicast route

```
switch# show forwarding distribution otv multicast route
Vlan: 311, Group: 224.0.0.0/4, Source: 0.0.0.0
  OTV Outgoing Interface List Index: 65535
  Reference Count: 1
  Number of Outgoing Interfaces: 0
Vlan: 311, Group: 224.0.0.0/24, Source: 0.0.0.0
  OTV Outgoing Interface List Index: 1
  Reference Count: 1
  Number of Outgoing Interfaces: 1
    External interface: Ethernet1/6
    Delivery group IP: 239.1.1.1
    Delivery source IP: 102.1.1.1
Vlan: 311, Group: 238.1.1.1, Source: 6.2.2.2
  OTV Outgoing Interface List Index: 2
  Reference Count: 1
  Number of Outgoing Interfaces: 1
    External interface: Ethernet1/6
    Delivery group IP: 232.1.1.0
    Delivery source IP: 102.1.1.1
```

Chesapeake NETCRAFTSMEN 114 Copyright 2011

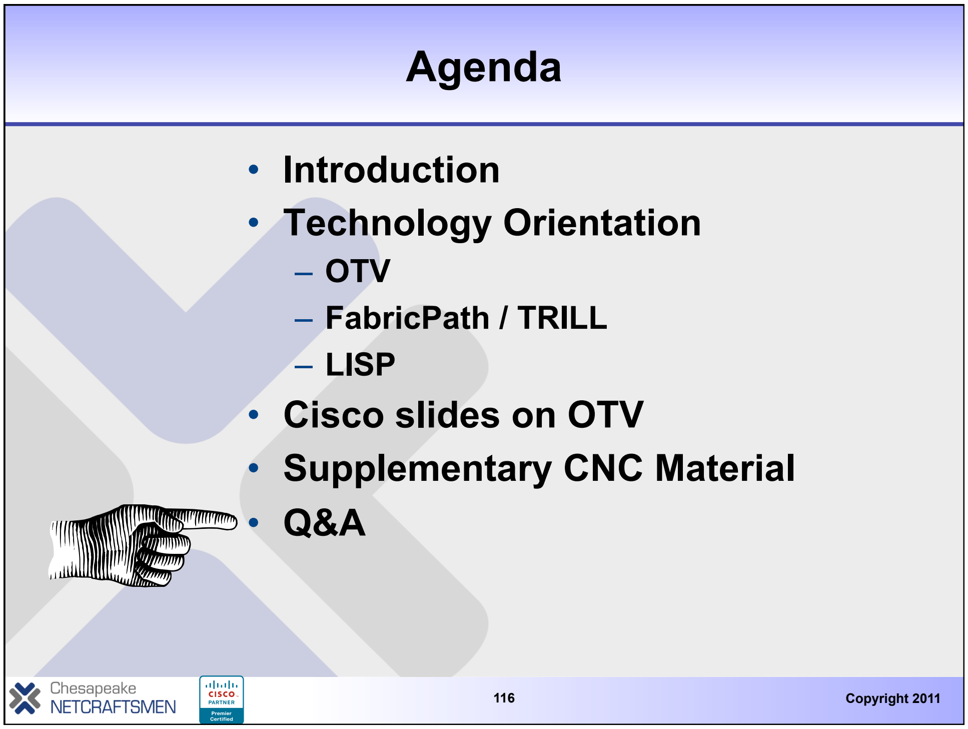


Summary

- OTV is a promising technology for L2 interconnect of data centers over any L3 WAN
- "Beer principle" applies:
  - Overdo it and you'll have a headache
  - L3 between sites still best design principle
  - OTV scales fairly well but not infinitely far

Chesapeake NETCRAFTSMEN

Copyright 2011



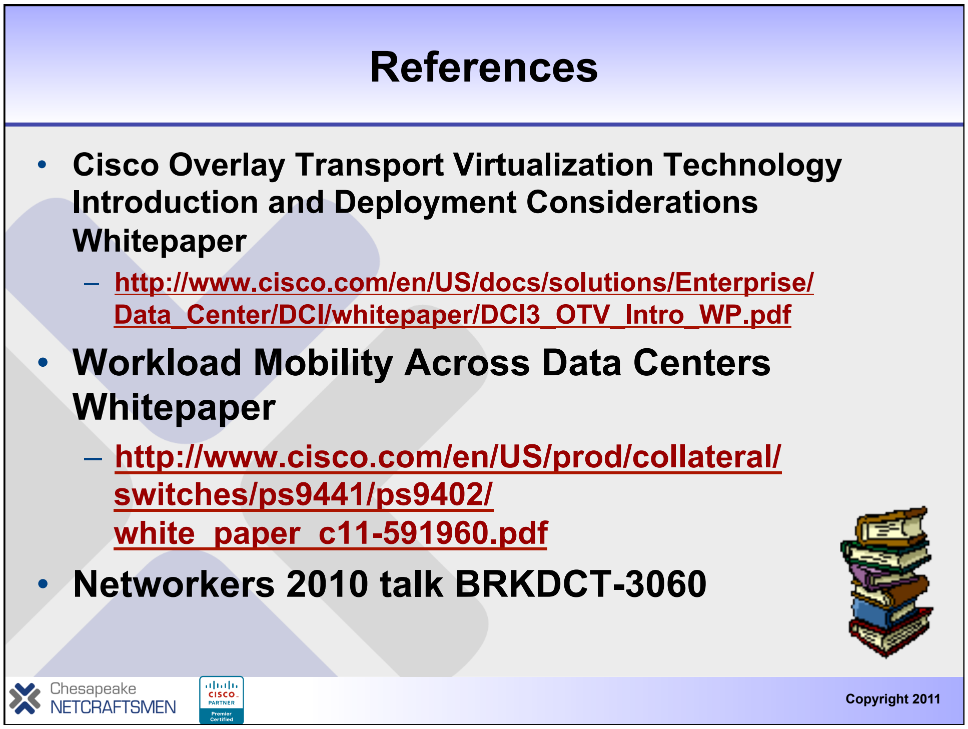
Agenda

- Introduction
- Technology Orientation
  - OTV
  - FabricPath / TRILL
  - LISP
- Cisco slides on OTV
- Supplementary CNC Material
- Q&A

Chesapeake NETCRAFTSMEN

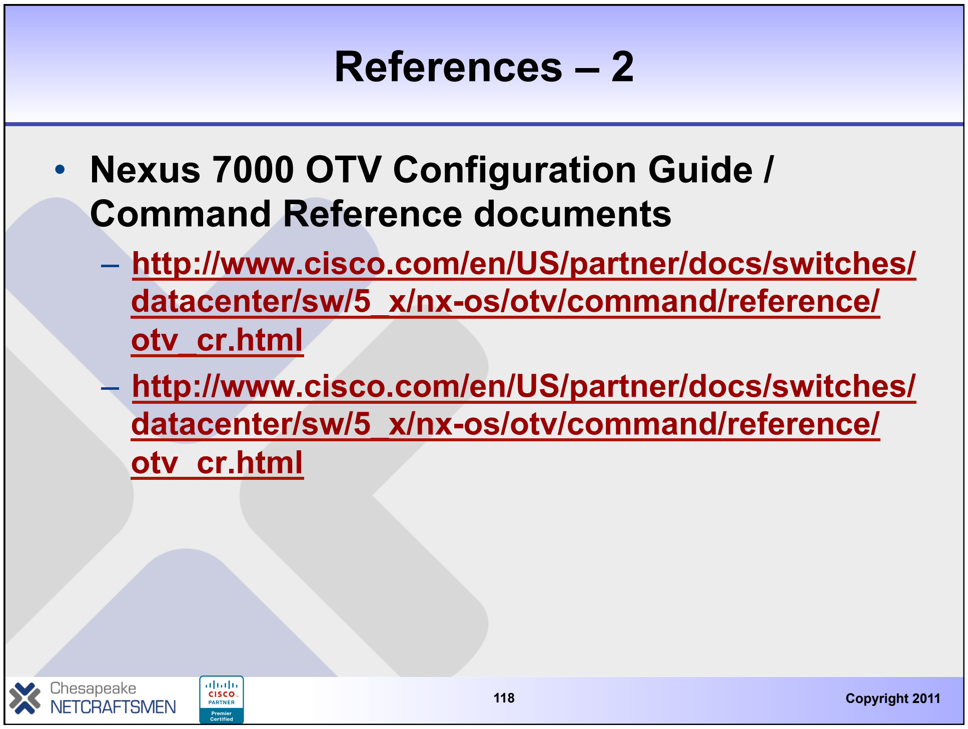
116

Copyright 2011



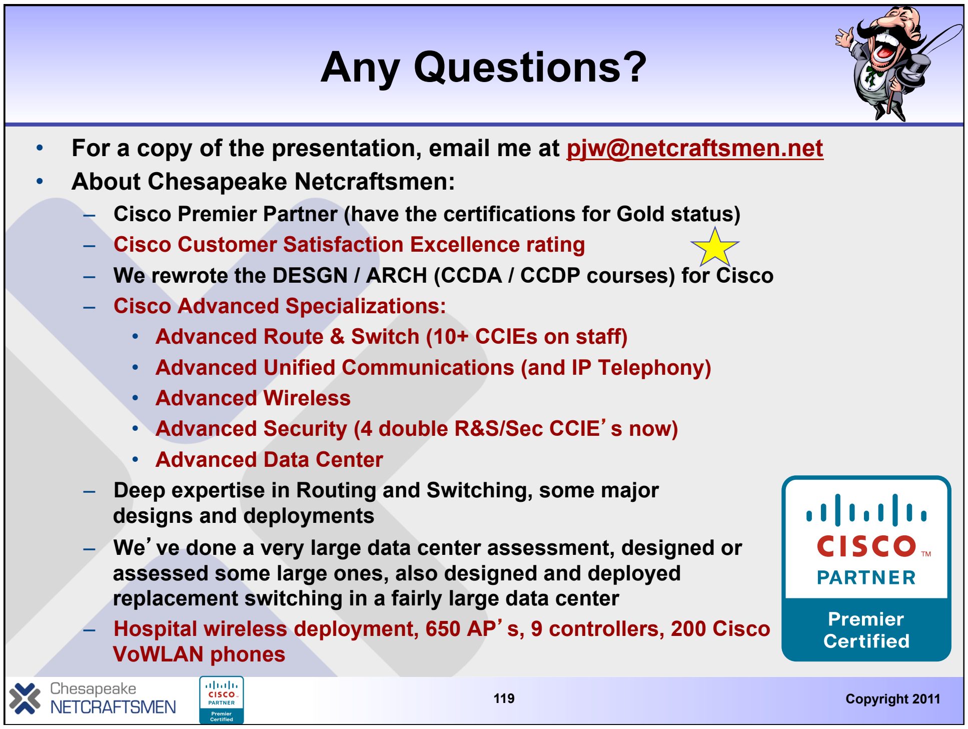
References

- Cisco Overlay Transport Virtualization Technology Introduction and Deployment Considerations Whitepaper
  - http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/whitepaper/DCI3_OTV_Intro_WP.pdf
- Workload Mobility Across Data Centers Whitepaper
  - http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-591960.pdf
- Networkers 2010 talk BRKDCT-3060



References – 2

- Nexus 7000 OTV Configuration Guide / Command Reference documents
  - http://www.cisco.com/en/US/partner/docs/switches/datacenter/sw/5_x/nx-os/otv/command/reference/otv_cr.html
  - http://www.cisco.com/en/US/partner/docs/switches/datacenter/sw/5_x/nx-os/otv/command/reference/otv_cr.html



Any Questions?

- For a copy of the presentation, email me at pjw@netcraftsmen.net
- About Chesapeake Netcraftsmen:
  - Cisco Premier Partner (have the certifications for Gold status)
  - Cisco Customer Satisfaction Excellence rating
  - We rewrote the DESGN / ARCH (CCDA / CCDP courses) for Cisco
  - Cisco Advanced Specializations:
    - Advanced Route & Switch (10+ CCIEs on staff)
    - Advanced Unified Communications (and IP Telephony)
    - Advanced Wireless
    - Advanced Security (4 double R&S/Sec CCIE's now)
    - Advanced Data Center
  - Deep expertise in Routing and Switching, some major designs and deployments
  - We've done a very large data center assessment, designed or assessed some large ones, also designed and deployed replacement switching in a fairly large data center
  - Hospital wireless deployment, 650 AP's, 9 controllers, 200 Cisco VoWLAN phones

CISCO PARTNER Premier Certified

Chesapeake NETCRAFTSMEN 119 Copyright 2011