What You Make Possible
Deployment Considerations with Interconnecting Data Centers

BRKDCT-3060
Session Objectives

The main goals of this session are:

- Highlighting the main business requirements driving Data Center Interconnect (DCI) deployments
- Understand the functional components of the holistic Cisco DCI solutions
- Get a full knowledge of Cisco LAN extension technologies and associated deployment considerations
- Integrate routing aspect induced by the emerging application mobility offered by DCI

This session does not include:

- Storage extension considerations associated to DCI deployments
## Related Cisco Live 2011 Events

**DCI Sessions**

<table>
<thead>
<tr>
<th>Session-ID</th>
<th>Session Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>TECDCT-2181</td>
<td>Deployment Considerations for Interconnecting Distributed Virtual Data Centers</td>
</tr>
<tr>
<td>BRKDCT-2011</td>
<td>Design and Deployment of Data Center Interconnects using Advanced VPLS (A-VPLS)</td>
</tr>
<tr>
<td>BRKDCT-2049</td>
<td>Overlay Transport Virtualization</td>
</tr>
<tr>
<td>BRKDCT-3103</td>
<td>Advanced OTV - Configure, Verify and Troubleshoot OTV in Your Network</td>
</tr>
<tr>
<td>BRKDCT-2081</td>
<td>Cisco FabricPath Technology and Design</td>
</tr>
<tr>
<td>BRKDCT-2131</td>
<td>Mobility and Virtualization in the Data Center with LISP and OTV</td>
</tr>
</tbody>
</table>
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
  - IP Based Solutions
- LISP for DCI Deployments
  - LISP and Path Optimization
  - LISP and Services (FW, SLB) Integration
- Summary and Q&A
## Data Center Interconnect

### Business Drivers

- Data Centers are extending beyond traditional boundaries
- Virtualization applications are driving DCI across PODs (aggregation blocks) and Data Centers

<table>
<thead>
<tr>
<th>Drivers</th>
<th>Business Solution</th>
<th>Constraints</th>
<th>IT Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operation Cost Containment</td>
<td>✓ Data Center Maintenance / Migration / Consolidation</td>
<td>✓ Host Mobility</td>
<td>✓ Distributed Virtual Data Center</td>
</tr>
<tr>
<td>Business Resource Optimization</td>
<td>✓ Disaster Avoidance ✓ Workload Mobility</td>
<td>✓ VLAN Extension ✓ Statefulness ✓ Bandwidth &amp; Latency</td>
<td>✓ VM Mobility</td>
</tr>
<tr>
<td>Cloud Services</td>
<td>✓ Inter-Cloud Networking ✓ XaaS</td>
<td>✓ Flexibility ✓ Application mobility</td>
<td>✓ VM Mobility ✓ Automation</td>
</tr>
</tbody>
</table>
Data Center Interconnect
LAN Extension Model

Path Optimization
Any type of links
STP Domain isolation + Storm-control
Dual-Homing

Storage extension
Data Center Interconnect
Host Mobility Scenarios

Moves Without LAN Extension
- IP Mobility Across Subnets
  - Disaster Recovery
  - Cloud Bursting
  - Migration
- Application Members in One Location

Moves With LAN Extension
- Routing for Extended Subnets
  - Active-Active Data Centers
  - Distributed Clusters
  - Workload mobility
- Application Members Distributed
  (Broadcasts across sites)
LAN Extension for DCI

VLAN Types

Type T0
Limited to a single access layer

Type T1
Extended inside an aggregation block (POD)

Type T2
Extended between PODs part of the same DC site

Type T3
Extended between twin DC sites connected via dedicated dark fiber links

Type T4
Extended between twin DC sites using non 5*9 connection

Type T5
Extended between remote DC sites
LAN Extension for DCI
Technology Selection Criteria

- Over dark fiber or protected D-WDM
  - VSS & vPC
    - Dual site interconnection
  - FabricPath (TRILL)

- MPLS Transport
  - EoMPLS
    - Transparent point to point
  - A-VPLS
    - Enterprise style MPLS
  - H-VPLS
    - Large scale & Multi-tenants

- IP Transport
  - OTV
    - Enterprise style Inter-site MAC Routing
  - VXLAN
    - Intra-site MAC bridging in total virtualized context
LAN Extension for DCI
Technology Selection Criteria

- **Transport**
  - Fiber
  - LOS report / Protected DWDM
  - L2 SP offer (HA=99.7+)
  - IP

- **Scale**
  - Site
  - VLAN (10^2 or 10^3 or 10^4)
  - MAC (10^3 or 10^4 or 10^5)

- **Multi-tenants**
  - Tagging (VLAN / 2Q / VRF)
  - Overlapping / Translation

- **Multi-point or point to point**

- **Greenfield vs. Brownfield**
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
    - VSS, vPC and FabricPath
  - MPLS Based Solutions
  - IP Based Solutions
- LISP for DCI Deployments
  - LISP and Path Optimization
  - LISP and Services (FW, SLB) Integration
- Summary and Q&A
Dual Site Interconnection
Leveraging EtherChannel between Sites

- Link utilization with Multi-Chassis EtherChannel
- DCI port-channel
  - 2 or 4 links
- Requires protected DWDM or Direct fibers

On DCI Etherchannel:
- STP Isolation (BPDU Filtering)
- Broadcast Storm Control
- FHRP Isolation

```
interface port-channel10
 desc DCI point to point connection
 switchport
 switchport mode trunk
 vpc 10
 switchport trunk allowed vlan 100-600
 spanning-tree port type edge trunk
 spanning-tree bpdufilter enable
 storm-control broadcast level 1
 storm-control multicast level x
```

- vPC does not support L3 peering:
  Use dedicated L3 Links for Inter-DC routing!
- Validated design:
  200 Layer 2 VLANs + 100 VLAN SVIs
  1000 VLAN + 1000 SVI (static routing)
Ingress FabricPath switch determines destination Switch ID and imposes FabricPath header.

Destination Switch ID used to make routing decisions through FabricPath core.

No MAC learning or lookups required inside core.

Egress FabricPath switch removes FabricPath header and forwards to CE.
**FabricPath**

Conversational MAC Learning

<table>
<thead>
<tr>
<th>MAC</th>
<th>IF/SID</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>e1/1 (local)</td>
</tr>
<tr>
<td>B</td>
<td>S200 (remote)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAC</th>
<th>IF/SID</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>S200 (remote)</td>
</tr>
<tr>
<td>C</td>
<td>e7/10 (local)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAC</th>
<th>IF/SID</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>S100 (remote)</td>
</tr>
<tr>
<td>B</td>
<td>e12/1 (local)</td>
</tr>
<tr>
<td>C</td>
<td>S300 (remote)</td>
</tr>
</tbody>
</table>
FabricPath for DCI
Partial-Meshed Topology for different models of DC

- Conversational Mac Learning
- Offer a full HA DCI solution with Native STP Isolation
- Provides easy integration with Brownfield DC
- Optimized using vPC+

- F1/F2 End to End for optimal design
- Required point to point connections
- Relies on Flooding for Unknown Unicast traffic
- No current Broadcast suppression
- L2 Multipath only for equal cost path can be leveraged (i.e. A⇔B or C⇔D)
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
    - EoMPLS
    - VPLS
    - H-VPLS
  - IP Based Solutions
- LISP for DCI Deployments
  - LISP and Path Optimization
  - LISP and Services (FW, SLB) Integration
- Summary and Q&A
interface g1/1
description EoMPLS port mode connection
no switchport
no ip address
xconnect 2.2.2.2 vcid 1 encapsulation mpls
EoMPLS Usage for DCI
End-to-End Loop Avoidance using Edge to Edge LACP

On DCI Etherchannel:
- STP Isolation (BPDU Filtering)
- Broadcast Storm Control
- FHRP Isolation

Encryption Services with 802.1AE
- Requires a full meshed vPC ➔ 4 PW
EoMPLS Usage for DCI
Over IP Core

crypto ipsec profile MyProfile
set transform-set MyTransSet

interface Tunnel100
ip address 10.11.11.11 255.255.255.0
ip mtu 9216
mpls ip
tunnel source Loopback100
tunnel destination 12.11.11.21
tunnel protection ipsec profile MyProfile
Dealing with PseudoWire (PW) Failures
Remote Ethernet Port Shutdown

PE receives the PW down notification and shutdown its transmit signal toward aggregation

ASR1000 feature configuration:

```plaintext
interface GigabitEthernet1/0/0
xconnect 1.1.1.1 1 pw-class eompls
remote link failure notification ! (default)
```

<table>
<thead>
<tr>
<th>Bridged traffic</th>
<th>Failover (msec)</th>
<th>Fallback (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>281</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>453</td>
<td>300</td>
</tr>
</tbody>
</table>
EoMPLS Deployment on VSS
Point to Point EoMPLS with Port-Channel xconnect

- Instead of xconnecting physical port, xconnect port-channel
- LACP is kept local, no more extended over EoMPLS
- PW is virtual on both VSS members
  SSO protection in 12.2(33)SXJ
- Requires VSS or Nexus as DC device
  Limited support of L3 routing with vPC
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
    - EoMPLS
    - VPLS
    - H-VPLS
  - IP Based Solutions
- LISP for DCI Deployments
  - LISP and Path Optimization
  - LISP and Services (FW, SLB) Integration
- Summary and Q&A
Multi-Point Topologies
What is VPLS?

One extended bridge-domain built using:
- VFI = Virtual Forwarding Instance
  (VSI = Virtual Switch Instance)
- PW = Pseudo-Wire
- SVI = Switch Virtual Interface
- xconnect

Mac address table population ➔ is pure Learning-Bridge
VPLS Cluster Solutions

- Using clustering mechanism
  - Two devices in fusion as one
    - VSS Sup720
    - VSS Sup2T
    - ASR9K nV virtual cluster
  - One control-plane / two data-planes
- Dual node is acting as one only device
  - Native redundancy (SSO cross chassis)
  - Native load balancing
  - Capability to use port-channel as attachment circuit
VPLS Redundancy
Making Usage of Clustering

- LDP session protection & Loopback usage allows PW state to be unaffected
- LDP + IGP convergence in sub-second
  Fast failure detection on Carrier-delay / BFD
- Immediate local fast protection
  Traffic exit directly from egress VSS node

<table>
<thead>
<tr>
<th>VSS</th>
<th>Fallover (msec)</th>
<th>Fallback (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bridged traffic</td>
<td>258</td>
<td>218</td>
</tr>
<tr>
<td></td>
<td>162</td>
<td>174</td>
</tr>
</tbody>
</table>
VPLS Redundancy
Making Usage of Clustering

- If failing slave node: PW state is unaffected
- If failing master node:
  - PW forwarding is ensured via SSO
  - PW state is maintained on the other side using Graceful restart
- Edge Ether-channel convergence in sub-second
- Traffic is directly going to working VSS node
- Traffic exits directly from egress VSS node
- Quad sup SSO for SUP2T in 1QCY13

<table>
<thead>
<tr>
<th>VSS</th>
<th>Failover (msec)</th>
<th>Fallback (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bridged traffic</td>
<td>224</td>
<td>412</td>
</tr>
<tr>
<td></td>
<td>326</td>
<td>316</td>
</tr>
</tbody>
</table>
Remote VSS are having two un-equal cost path to others, so one only route is put in RIB

- Stops forwarding traffic for 2mn when primary route is removed
  (there is no control-plane to insert backup route)

Build a symmetric core with two ECMP paths between each VSS
Remark: ASR9K is better supporting asymmetric core
VSS - A-VPLS CLI
Available Q4CY12 for SUP2T

Rem: One PW per VLAN per destination

#sh mpls l2 vc

<table>
<thead>
<tr>
<th>Local intf</th>
<th>Local circuit</th>
<th>Dest address</th>
<th>VC ID</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>VFI VFI_610_ VFI</td>
<td>10.100.2.2</td>
<td>610</td>
<td>UP</td>
<td></td>
</tr>
<tr>
<td>VFI VFI_610_ VFI</td>
<td>10.100.3.3</td>
<td>610</td>
<td>UP</td>
<td></td>
</tr>
<tr>
<td>VFI VFI_611_ VFI</td>
<td>10.100.2.2</td>
<td>611</td>
<td>UP</td>
<td></td>
</tr>
<tr>
<td>VFI VFI_611_ VFI</td>
<td>10.100.3.3</td>
<td>611</td>
<td>UP</td>
<td></td>
</tr>
</tbody>
</table>

interface Virtual-Ethernet1
switchport
switchport mode trunk
switchport trunk allowed vlan 610-619
neighbor 10.100.2.2 pw-class Core
neighbor 10.100.3.3 pw-class Core
pseudowire-class Core
encapsulation mpls

- Any card type facing edge
- SUP720 + SIP-400 facing core (5Gbps) or
- SUP720 + ES-40 (40Gbps) support with 12.2(33)SXJ
- SUP2T Q4CY12
ASR9K VPLS Set-up

```
12vpn
  router-id 10.0.1.1
  bridge group BG
    bridge-domain BD
      interface TenGigE0/0/0/4
      interface TenGigE0/0/0/5
!
  vfi VFI
    vpn-id 4003
    neighbor 10.0.1.2 pw-id 4003
```
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
    - EoMPLS
    - VPLS
    - H-VPLS
  - IP Based Solutions
- LISP for DCI Deployments
  - LISP and Path Optimization
  - LISP and Services (FW, SLB) Integration
- Summary and Q&A
Multi-Tenant Data Center

ASR 9000
Wan Connection
Intra-DC Inter-POD Routing
DCI

POD 1 = 1 x Row = 10 x Rack
POD N = M x Rows = M x 10 x Rack
## DC Access Multi-Homing

### Solution Summary

<table>
<thead>
<tr>
<th>Highlights</th>
</tr>
</thead>
</table>
| **Node clustering** | VSS (Catalyst / Cisco 7600-Sup2T)  
Nv cluster (ASR9k)  
One control-plane for two chassis  
Easiness, Active/Active |
| **Multi-chassis LAG** | Simple solution for spoke-and-hub topology, works for both bridging and non-bridging access device  
Standard based solution by using 802.3ad  
Sub-second convergence  
Phase 1 implement is active/standby mode.  
Phase 2 is per VLAN load balancing  
Ring topology support is under investigation |
| **REP /REP access gateway** | Sub 200msec convergence  
Good access ring isolation  
Now standard based → G.8032 (XR4.1 release)  
Spoke-and-hub and ring topology, not works well for mesh network |
| **MST/PVST access gateway** | Standard based solution as long as access network support MST/PVST  
Works for any access network topology  
Good access domain isolation  
Work with 802.1ah PBB  
Convergence time depends on access network STP |
DC Access Multi-Homing
Inter Chassis Communication Protocol - ICCP

- ICCP synchronizes event/states between multiple chassis in a redundancy group
- ICCP runs over reliable LDP / TCP
- ICCP relies on BFD/IP route-watch as keepalive
- ICCP message to synch state
  Ex: LACP, IGMP query …

Terminology:
- mLACP: Multi-Chassis Link Aggregation Control Protocol
- MC-LAG: Multi-Chassis Link Aggregation Group
- DHD: Dual Homed Device (Customer Edge)
- DHN: Dual Homed Network (Customer Edge)
- POA: Point of Attachment (Provider Edge)
DC Access Multi-Homing
Inter Chassis Communication Protocol - ICCP

Multi-Chassis LACP synchronization:
- LACP BPDUs (01:80:C2:00:00:00) are exchanged on each Link
- System Attributes: Priority + bundle MAC Address
- Port Attributes: Key + Priority + Number + State

Terminology:
- mLACP : Multi-Chassis Link Aggregation Control Protocol
- MC-LAG : Multi-Chassis Link Aggregation Group
- DHD : Dual Homed Device (Customer Edge)
- DHN : Dual Homed Network (Customer Edge)
- POA : Point of Attachment (Provider Edge)
MC-LAG to VPLS Testing


Only error 2/3/4 are leading to ICCP convergence
Rem: 2 & 4 are dual errors

- 500 VLAN Unicast: Link error sub-1s & Node error sub-2s
- 1200 VLAN unicast: Link error sub-2s & Node error sub-4s
Flexible VLAN Handling
Ethernet Virtual Circuit - EVC

1. Selective Trunk Support
   Group multiple VLAN in one only core bridge domain
   - QinQ model
   - VLAN overlapping

2. VLAN translation 121 / 222 / …
   Inter-DC VLAN numbering independency

3. Scale to 4000 * 4000 VLAN
   Scale above 4000 VLAN

4. Routing for multi-TAG
   Multi-tenant default gateway
   IRB - IP routing / VRF routing for QinQ tagged frames
E-VPN (aka Routed VPLS)
Main Principles

- Control-Plane Distribution of Customer MAC-Addresses using BGP
  - PE continues to learn C-MAC over AC
  - When multiple PEs announce the same C-MAC, hash to pick one PE
- MP2MP/P2MP LSPs for Multicast Traffic Distribution
- MP2P (like L3VPN) LSPs for Unicast Distribution
- Full-Mesh of PW no longer required !!
MPLS DCI Conclusion
A Mature Solution

- **EoMPLS** DCI is an easy point to point solution
- **VPLS** DCI is having two flavors:
  1. **Cluster**
     - Simplicity
     - Very fast convergence
     - Available
       - Cat 6K : SUP720 / SUP2T
       - 7600: SUP2T (Q3CY2012)
       - ASR9K: Nv with multi-tenant features
  2. **Dual node** based on mLACP attachment circuit
     - High-end devices (7600 / ASR9K, …)
     - Multi-tenant features / VLAN Translation
     - High scale
     - High SLA features
     - Standard based
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
  - IP Based Solutions
    - OTV Technology Overview
    - OTV Deployment Considerations
- LISP for DCI Deployments
  - LISP and Path Optimization
  - LISP and Services (FW, SLB) Integration
- Summary and Q&A
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

Protocol Learning
- Nexus 7000
  First platform to support OTV (since 5.0 NXOS Release)
- ASR 1000
  Now also supporting OTV (since 3.5 XE Release)

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

OTV Control Plane

- **Edge Device (ED):** connects the site to the (WAN/MAN) core and responsible for performing all the OTV functions
- **Internal Interfaces:** L2 interfaces (usually 802.1q trunks) of the ED that face the site
- **Join Interface:** L3 interface of the ED that faces the core
- **Overlay Interface:** logical multi-access multicast-capable interface. It encapsulates Layer 2 frames in IP unicast or multicast headers
OTV Data Plane
Inter-Site Packet Flow
Overlay Transport Virtualization

OTV Control Plane

- Neighbor discovery and adjacency over
  Multicast (Nexus 7000 and ASR 1000)
  Unicast (Adjacency Server Mode currently available with Nexus 7000 from 5.2 release)
- OTV proactively **advertises/withdraws** MAC reachability (control-plane learning)
- IS-IS is the OTV Control Protocol - No specific configuration required

<table>
<thead>
<tr>
<th>VLAN</th>
<th>MAC</th>
<th>IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>MAC A</td>
<td>IP A</td>
</tr>
<tr>
<td>100</td>
<td>MAC B</td>
<td>IP A</td>
</tr>
<tr>
<td>100</td>
<td>MAC C</td>
<td>IP A</td>
</tr>
</tbody>
</table>

3 New MACs are learned on VLAN 100

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

OTV updates exchanged via the L3 core

OTV Update 1
OTV Update 2
OTV Update 3
OTV Update 4
OTV Failure Domain Isolation
Spanning-Tree Site Independence

- Site transparency: no changes to the STP topology
- Total isolation of the STP domain
- **Default behavior: no configuration is required**
- BPDUs sent and received ONLY on Internal Interfaces
OTV Failure Domain Isolation
Preventing Unknown Unicast Storms

- No requirements to forward unknown unicast frames
- Assumption: end-host are not silent or uni-directional
- Default behavior: no configuration is required
# OTV Multi-homing

VLANs Split Across AEDs

- Automated and deterministic algorithm (not configurable)
- In a dual-homed site:
  - Lower IS-IS System-ID (Ordinal 0) = EVEN VLANs
  - Higher IS-IS System-ID (Ordinal 1) = ODD VLANs
- Future functionality will allow to tune the behavior

---

**OTV-a**

```
# show otv vlan
```

<table>
<thead>
<tr>
<th>VLAN</th>
<th>Auth. Edge Device</th>
<th>Vlan State</th>
<th>Overlay</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>East-b</td>
<td>inactive(Non AED)</td>
<td>Overlay100</td>
</tr>
<tr>
<td>101*</td>
<td>East-a</td>
<td>active</td>
<td>Overlay100</td>
</tr>
<tr>
<td>102</td>
<td>East-b</td>
<td>inactive(Non AED)</td>
<td>Overlay100</td>
</tr>
</tbody>
</table>

---

**OTV-b**

```
# show otv vlan
```

<table>
<thead>
<tr>
<th>VLAN</th>
<th>Auth. Edge Device</th>
<th>Vlan State</th>
<th>Overlay</th>
</tr>
</thead>
<tbody>
<tr>
<td>100*</td>
<td>East-b</td>
<td>active</td>
<td>Overlay100</td>
</tr>
<tr>
<td>101</td>
<td>East-a</td>
<td>inactive(Non AED)</td>
<td>Overlay100</td>
</tr>
<tr>
<td>102*</td>
<td>East-b</td>
<td>active</td>
<td>Overlay100</td>
</tr>
</tbody>
</table>

---

**Remote OTV Device MAC Table**

<table>
<thead>
<tr>
<th>VLAN</th>
<th>MAC</th>
<th>IF</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>MAC 1</td>
<td>IP A</td>
</tr>
<tr>
<td>101</td>
<td>MAC 2</td>
<td>IP B</td>
</tr>
</tbody>
</table>

---

*Supported from 5.2 NX-OS release*
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
  - IP Based Solutions
    - OTV Technology Overview
    - OTV Deployment Considerations
- LISP for DCI Deployments
  - LISP and Path Optimization
  - LISP and Services (FW, SLB) Integration
- Summary and Q&A
Placement of the OTV Edge Device

Option 1 - OTV in the DC Core with L3 Boundary at Aggregation

- Easy deployment for Brownfield
- L2-L3 boundary remains at aggregation
- DC Core devices performs L3 and OTV functionalities
  - May use a pair of dedicated Nexus 7000
- VLANs extended from aggregation layer
  - L2 “Octopus” design
  - Recommended to use separate physical links for L2 & L3 traffic
  - STP and L2 broadcast domains not isolated between PODs (Aggregation Blocks)
Placement of the OTV Edge Device
Option 2 - OTV at the Aggregation with L2-L3 Boundary on External Firewalls

- The Firewalls host the Default Gateway
- No SVIs at the Aggregation Layer
- Requires at least a routed link between Aggregation and Core (OTV Join Interface)
  No SVI supported as Join Interface
OTV and SVI Routing
Introducing the OTV VDC

- **Guideline:** The current OTV implementation on the Nexus 7000 enforces the separation between SVI routing and OTV encapsulation for any extended VLAN.

- This separation 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.
### Placement of the OTV Edge Device

**Option 3—OTV in the DC Aggregation**

- L2-L3 boundary at aggregation
- DC Core performs only L3 role
- STP and L2 broadcast Domains isolated between PODs
- Intra-DC and Inter-DCs LAN extension provided by OTV
  - Requires the deployment of dedicated OTV VDCs
- Ideal for single aggregation block topologies
- Recommended for Green Field deployments
  - Nexus 7000 required in aggregation
Single Homed OTV VDC

Simple Model

- May use a single physical link for Join and Internal interfaces
  - Minimizes the number of ports required to interconnect the VDCs
- Single link or physical node (or VDC) failures lead to AED re-election
  - 50% of the extended VLANs affected
- Failure of the routed link to the core is not OTV related
  - Recovery is based on IP convergence
Dual Homed OTV VDC

Improving the Design Resiliency

- Logical Port-channels used for the Join and the Internal interfaces
  Increases the number of physical interfaces required to interconnect the VDCs
- Traffic recovery after single link failure event based on port-channel re-hashing
  No need for AED re-election
- Physical node (or VDC) failure still requires AED re-election
  In the current implementation may cause few seconds of outage (for 50% of the extended VLANs)
OTV in the DC Aggregation
Site Based Per-VLAN Load Balancing

- AED role negotiated between the two OTV VDCs (on a per VLAN basis)
  - Internal IS-IS peering on the site VLAN
    - Recommended to carry the site VLAN on vPC links and vPC peer-link
  - For a given VLAN all traffic must be carried to the AED Device
    - Part of the flows carried across the vPC peer-link
    - Optimized traffic flows is achieved in the most resilient model leveraging Port-Channels as Internal Interfaces

- The AED encapsulates the original L2 frame into an IP packet and send it back to the aggregation layer device
- The aggregation layer device routes the IP packet toward the DC Core/WAN edge
- L3 routed traffic bypasses the OTV VDC
OTV in the DC Aggregation
Per-Device Load Balancing

- Unicast traffic directed to (received from) the same remote site (AED) will always use the same physical link
  - OTV encapsulated packets characterized by the same <Src-IP, Dst-IP> information

- In multipoint deployments unicast traffic **may** leverage multiple equal cost paths
  - <Dts-IP> value changes with the remote OTV Edge Device (AED1, AED2)

- Next generation HW (CY12) would allow to achieve flow based load-balancing
  - OTV traffic encapsulated into UDP with variable source port #
OTV in the DC Aggregation
Using F-Series Linecards

- F1 and F2 linecards do not support OTV natively.
- As of today, the OTV VDC must use only M-series ports for both Internal and Join Interfaces.
  - Recommendation is to allocate M1 only interfaces to the OTV VDC.
- Native OTV support on F-series is targeted for 6.2 release (Q1CY12).
**Could use static default route or ospf stub**

---

**OTV in the DC Aggregation**

**Configuration (Multicast Transport)**

OTV VDC

Routing VDC

N7K-Agg1

N7K-Agg2

PIM enabled interfaces

Establish L3 peering on a dedicated VLAN

---

**hostname routing-vdc**

```
! interface Ethernet1/1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 100,600-700
!
interface Ethernet2/1
  ip address 3.3.3.1/24
  ip router ospf 1 area 0.0.0.0
  ip ospf passive-interface
  ip pim sparse-mode
  ip igmp version 3
!
interface Overlay100
  otv join-interface Ethernet2/2
  otv control-group 239.1.1.2
  otv data-group 232.1.1.0/24
  otv extend-vlan 600-700
!
ip route 0.0.0.0 0.0.0.0 3.3.3.1
```

---

**hostname otv-vdc**

```
feature otv
!
otv site-vlan 100
!
interface Ethernet1/2
  description Internal Interface
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 100,600-700
!
interface Ethernet2/2
  description Join Interface
  ip address 3.3.3.2/24
  ip igmp version 3
!
interface Overlay100
  otv join-interface Ethernet2/2
  otv control-group 239.1.1.2
  otv data-group 232.1.1.0/24
  otv extend-vlan 600-700
!
ip route 0.0.0.0 0.0.0.0 3.3.3.1
```
OTV in the DC Aggregation
Configuration (Unicast Transport)

```
hostname routing-vdc
!
interface Ethernet1/1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 100,600-700
!
interface Ethernet2/1
  ip address 3.3.3.1/24
  ip router ospf 1 area 0.0.0.0
  ip ospf passive-interface

hostname otv-vdc
  feature otv
    otv site-vlan 100
    !
interface Ethernet1/2
  description Internal Interface
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 100,600-700
!
interface Ethernet2/2
  description Join Interface
  ip address 3.3.3.2/24
  !
interface Overlay100
  otv join-interface Ethernet2/2
  otv adjacency-server*
  otv use-adjacency-server 10.1.1.1 11.1.1.1
  otv extend-vlan 600-700
  !
  ip route 0.0.0.0 0.0.0.0 3.3.3.1
```

* Needed only on the Adjacency Server

Release 5.2 and above

Establish L3 peering on a dedicated VLAN
OTV in the DC Aggregation

HSRP Isolation Configuration

1. Create and apply the policies to filter out HSRP messages (both v1 and v2 in this example)

   ip access-list ALL_IPs
   10 permit ip any any

   mac access-list ALL_MACs
   10 permit any any

   !
   ip access-list HSRP_IP
   10 permit udp any 224.0.0.2/32 eq 1985
   20 permit udp any 224.0.0.102/32 eq 1985

   !
   mac access-list HSRP_VMAC
   10 permit 0000.0c07.ac00 0000.0000.00ff any
   20 permit 0000.0c9f.f000 0000.0000.0fff any

   !
   vlan access-map HSRP_Localization 10
   match mac address HSRP_VMAC
   match ip address HSRP_IP
   action drop

   !
   vlan access-map HSRP_Localization 20
   match mac address ALL_MACs
   match ip address ALL_IPs
   action forward

   !
   vlan filter HSRP_Localization vlan-list 600-1000

2. Configure ARP filtering to ensure ARP replies (or Gratuitous ARP) are not received from the remote site

   arp access-list HSRP_VMAC_ARP
   10 deny ip any mac 0000.0c07.ac00 ffff.ffff.ff00
   20 deny ip any mac 0000.0c9f.f000 ffff.ffff.f000
   30 permit ip any mac any

   !
   feature dhcp

   ip arp inspection filter HSRP_VMAC_ARP 600-1000

3. Apply a route-map on the OTV control plane to avoid communicating vMAC info to remote OTV edge devices

   mac-list HSRP-vmac-deny seq 5 deny 0000.0c07.ac00 ffff.ffff.ff00
   mac-list HSRP-vmac-deny seq 10 deny 0000.0c9f.f000 ffff.ffff.f000
   mac-list HSRP-vmac-deny seq 20 permit 0000.0000.0000 0000.0000.0000

   !
   route-map stop-HSRP permit 10
   match mac-list HSRP-vmac-deny

   otv-isis default
   vpn Overlay1

   redistribute filter route-map stop-HSRP

---

**Network Diagram**

- OTV VDCs
- Default VDCs
- NTK-Agg1
- NTK-Agg2

**Traffic Flow**

- OTV Traffic

---
OTV in the DC Aggregation

Verification

DC1-Agg1# sh hsrp
Vlan600 - Group 1 (HSRP-V1) (IPv4)
  Local state is Active, priority 130 (Cfged 130), may preempt
  Forwarding threshold (for vPC), lower: 1 upper: 130
  Hellotime 3 sec, holdtime 10 sec
  Next hello sent in 2.999000 sec(s)
  Virtual IP address is 10.15.0.1 (Cfged)
  Active router is local
  Standby router is 10.15.0.2, priority 120 expires in 4.96 sec(s)
  Authentication text "cisco"
  Virtual mac address is 0000.0c07.ac01 (Default MAC)
  2 state changes, last state change 3d19h
  IP redundancy name is hsrp-Vlan600-1 (default)

DC1-Agg2# sh hsrp
Vlan600 - Group 1 (HSRP-V1) (IPv4)
  Local state is Standby, priority 120 (Cfged 120)
  Forwarding threshold (for vPC), lower: 1 upper: 120
  Hellotime 3 sec, holdtime 10 sec
  Next hello sent in 1.515000 sec(s)
  Virtual IP address is 10.15.0.1 (Cfged)
  Active router is 10.15.0.3, priority 130 expires in 5.89 sec(s)
  Standby router is local
  Authentication text "cisco"
  Virtual mac address is 0000.0c07.ac01 (Default MAC)
  1 state changes, last state change 00:02:46
  IP redundancy name is hsrp-Vlan600-1 (default)

DC2-Agg1# sh hsrp
Vlan600 - Group 1 (HSRP-V1) (IPv4)
  Local state is Active, priority 110 (Cfged 110), may preempt
  Forwarding threshold (for vPC), lower: 1 upper: 110
  Hellotime 3 sec, holdtime 10 sec
  Next hello sent in 2.2000 sec(s)
  Virtual IP address is 10.15.0.1 (Cfged)
  Active router is local
  Standby router is 10.15.0.4, priority 100 expires in 3.96 sec(s)
  Authentication text "cisco"
  Virtual mac address is 0000.0c07.ac01 (Default MAC)
  2 state changes, last state change 2d2h
  IP redundancy name is hsrp-Vlan600-1 (default)

DC2-Agg2# sh hsrp
Vlan600 - Group 1 (HSRP-V1) (IPv4)
  Local state is Standby, priority 100 (Cfged 100), may preempt
  Forwarding threshold (for vPC), lower: 1 upper: 100
  Hellotime 3 sec, holdtime 10 sec
  Next hello sent in 1.2000 sec(s)
  Virtual IP address is 10.15.0.1 (Cfged)
  Active router is 10.15.0.5, priority 110 expires in 4.29 sec(s)
  Standby router is local
  Authentication text "cisco"
  Virtual mac address is 0000.0c07.ac01 (Default MAC)
  2 state changes, last state change 1d2h
  IP redundancy name is hsrp-Vlan600-1 (default)
Placement of the OTV Edge Device
Connecting Brownfield and Greenfield Data Centers

- Leverage OTV capabilities on Nexus 7000 (Greenfield) and ASR 1000 (Brownfield)
- Build on top of the traditional DC L3 switching model (L2-L3 boundary in Agg, Core is pure L3)
- Possible integration with the FabricPath/TRILL model
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
  - IP Based Solutions
- LISP for DCI Deployments
  - LISP and Path Optimization
  - LISP and Services (FW, SLB) Integration
- Summary and Q&A
Path Optimization and DCI
Avoid Suboptimal Traffic Path After Workload Motion

- Move the whole application tier
- Optimize the whole path:
  - Client to Server
  - Server to Server
  - Server to Client

144.254.100.0/25 & 144.254.100.128/25
EEM or RHI can be used to get very granular

Layer 3 Core
Ingress Path Optimization: Clients-Server

Server-Server Path Optimization

VLAN A

Egress Path Optimization: Server-Client
Outbound Path Optimization

FHRP Filtering

- Filter FHRP with combination of VACL or PACL
- Result: Still have one HSRP group with one VIP, but now have active router at each site for optimal first-hop routing
Inbound Path Optimization
Extending Subnets Creates a Routing Challenge

- A subnet usually implies location
- Yet we use LAN extensions to stretch subnets across locations
  - Location semantics of subnets are lost
- Traditional routing relies on the location semantics of the subnet
  - Can’t tell if a server is at the East or West location of the subnet
- More granular (host level) information is required
  - LISP provides host level location semantics
Inbound Path Optimization

LISP Host Mobility

1. **DNS Entry:**
   - D.abc.com A 10.2.0.1

2. **Mapping Cache Entry:**
   - 10.1.0.0/24
   - 10.1.0.1 -> 10.2.0.1

3. **Locator-set:**
   - 10.2.0.1/32
   - EID-prefix: 10.2.0.1/32
   - Locator-set:
     - 2.1.1.1, priority: 1, weight: 50 (D1)
     - 2.1.2.1, priority: 1, weight: 50 (D2)

4. **Mapping DB:**
   - 10.1.0.1 -> 2.1.1.1
   - 10.1.0.1 -> 2.1.2.1

5. **Inbound Path Optimization:**
   - 10.1.0.1 -> 10.2.0.1
Inbound Path Optimization
LISP Host Mobility

DNS Entry: D.abc.com A 10.2.0.1
LISP Host-Mobility with Extended Configuration

```plaintext
ip lisp itr-etr
ip lisp database-mapping 10.2.0.0/16 <RLOC-A> p 1 w 10
ip lisp database-mapping 10.2.0.0/16 <RLOC-B> p 1 w 10

lisp dynamic-eid roamer
database-mapping 10.2.0.0/24 <RLOC-A> p 1 w 10
database-mapping 10.2.0.0/24 <RLOC-B> p 1 w 10
map-server 1.1.1.1 key abcd
map-notify-group 239.10.10.10

interface vlan 100
ip address 10.2.0.3 /24
lisp mobility roamer
lisp extended-subnet-mode
hsrp 101
ip 10.2.0.1
```

```
ip lisp itr-etr
ip lisp database-mapping 10.3.0.0/16 <RLOC-C> p 1 w 10
ip lisp database-mapping 1032.0.0/16 <RLOC-D> p 1 w 10

lisp dynamic-eid roamer
database-mapping 10.2.0.0/24 <RLOC-C> p 1 w 10
database-mapping 10.2.0.0/24 <RLOC-D> p 1 w 10
map-server 1.1.1.1 key abcd
map-notify-group 239.10.10.10

interface vlan 100
ip address 10.2.0.5 /24
lisp mobility roamer
lisp extended-subnet-mode
hsrp 101
ip 10.2.0.1
```
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
  - IP Based Solutions
- LISP for DCI Deployments
  - L3 Host Mobility using LISP
  - LISP and Path Optimization
  - Services (FW, SLB) Integration
- Summary and Conclusions
- Q&A
LISP and FW Integration
Deployment Considerations

- FW must currently be positioned “south” of the LISP device
  No inspection possible for LISP encapsulated traffic, only stateless ACLs are possible when deploying the FW “north” of the xTR
- Option 1: FW in routed mode positioned between the default gateway and the LISP xTR
  Recommended with LISP Multi-hop Mobility enhancements (Q1CY13 for Nexus 7000)
- Option 2: FW in transparent mode or Virtual Services Gateway (VSG) → simple scenario since LISP xTR remains co-located on default gateway device
LISP and FW Deployment
Active/Standby Units Deployed in Each Site

- FWs in separated sites work independently
  Stateless Active/Active scenario
  Limit sub-optimal traffic through DCI core
- FW in different sites are not sync’d
  Policies have to be replicated between sites
  No state information maintained between sites
- May drop previously established sessions after workload vMotion
  Not an issue in cold migration scenarios (like Disaster Recovery for example)
LISP and FW Deployment

ASA Cluster Deployment Model*

- FW Clustered spread across multiple DC locations
- “Stateless” clustering approach
- Every flow is active only on one cluster node
- Intra-cluster redirection used if traffic is received by a node that does not have state information → needs extended VLAN for that
- Allows maintaining established sessions after a live workload mobility event
  - Sub-optimal path for already established sessions
  - Optimized path for new sessions

* Availability Q3CY12
LISP and SLB Integration
Deployment Considerations

- SLB VIP is active at one location at a time and represents the LISP EID
  - VIP activity is detected by the LISP Host Mobility logic
  - Migration of workloads belonging to the load-balanced server farm does not necessarily trigger a VIP move
- VIP location is updated in LISP only when moved between site
  - Server farm migration use case
  - Desire is to have an automated way to move the VIP between locations
Agenda

- DCI Business Drivers and Solutions Overview
- LAN Extension Deployment Scenarios
  - Ethernet Based Solutions
  - MPLS Based Solutions
  - IP Based Solutions
- LISP for DCI Deployments
  - L3 Host Mobility using LISP
  - LISP and Path Optimization
- Summary and Q&A
Connecting Virtualized Data Centers

L2 Domain Elasticity
- STP Isolation is the key element
- Multipoint
- Loop avoidance + Storm-Control
- Unknown Unicast & Broadcast control
- Link sturdiness
- Scale & Convergence

Storage Elasticity
- Sync or Async replication modes are driven by the applications, hence the distance/latency is a key component to select the choice
- Localization of Active Storage is key
  ➔ Distance can be improved using IO accelerator or caching
  ➔ Virtual LUN is allowing Active/Active

Path Optimization
Considerations
- Network and Security services deployment
- Server-Client Flows
- Server-Server Flows
Path Optimization Options
- Egress
  ➔ Addressed by FHRP Filtering
- Ingress:
  ➔ Addressed by LISP

- STP Isolation
- Multipoint
- Loop avoidance + Storm-Control
- Unknown Unicast & Broadcast control
- Link sturdiness
- Scale & Convergence

VM-Mobility
LAN Extensions
OTV
OTV
OTV
OTV

Sync or Async replication modes are driven by the applications, hence the distance/latency is a key component to select the choice
- Localization of Active Storage is key
  ➔ Distance can be improved using IO accelerator or caching
  ➔ Virtual LUN is allowing Active/Active

Cisco Live!
Data Center Interconnect
Where to Go for More Information

Extend Layer 2 Applications
Overlay Transport Protocol (OTV) technology supports transparent workload mobility within and across dispersed data centers. (3:54 min)

Get Instant Workload Mobility Among Data Centers
Cisco Data Center Interconnect (DCI) solutions can help your IT organization meet business continuity and corporate compliance objectives.

- Reduce the business impact of disaster events and help ensure business continuity
- Improve productivity through enhanced application and data availability
- Meet corporate and regulatory compliance needs and improve data security

These solutions transparently extend LAN and SAN connectivity and provide accelerated, highly secure data replication, server clustering, and workload mobility between geographically dispersed data centers. This enhances business resilience, and helps enable application and data mobility between data centers, while maintaining operational consistency.

Featured Products

- **Cisco Nexus 7000 Series LAN Extension**
  Simplify Layer 2 applications across distributed data centers.

- **Cisco Catalyst 6500 Series LAN Extension**
  Deliver high-performance, scalable Layer 2 extension with sub-second convergence.

- **Cisco MDS 9500 Series SAN Extension Over IP**
  Gain an integrated, cost-effective, reliable business continuity solution.

http://www.cisco.com/go/dci
Complete Your Online Session Evaluation

- Give us your feedback and you could win fabulous prizes. Winners announced daily.
- Receive 20 Passport points for each session evaluation you complete.
- Complete your session evaluation online now (open a browser through our wireless network to access our portal) or visit one of the Internet stations throughout the Convention Center.

Don’t forget to activate your Cisco Live Virtual account for access to all session material, communities, and on-demand and live activities throughout the year. Activate your account at the Cisco booth in the World of Solutions or visit www.ciscolive.com
Final Thoughts

- Get hands-on experience with the Walk-in Labs located in World of Solutions, booth 1042
- Come see demos of many key solutions and products in the main Cisco booth 2924
- Visit www.ciscoLive365.com after the event for updated PDFs, on-demand session videos, networking, and more!
- Follow Cisco Live! using social media:
  - Facebook: https://www.facebook.com/ciscoliveus
  - Twitter: https://twitter.com/#!/CiscoLive
  - LinkedIn Group: http://linkd.in/CiscoLI