Cisco Expo 2012

Data Center Fabric Evolution

Tomáš Michaeli

Oct-2012
Agenda

DC Fabric Evolution
Spanning Tree -> Virtual Port Channel -> FabricPath

Simplicity in DC operations
How Unified Fabric + FEX + Port Profile + DCNM influence TCO

Automation, virtualization and high-availability
Efficiency, plus scale and agile IT service delivery
Cisco Unified Fabric

Inovate together

CONVERGENCE

SCALE

INTELLIGENCE

ISSU
Minimize Downtime

FabricPath
Architectural Flexibility / Scale

OTV
Workload Mobility

FCoE
Unified fabric – wire once

FEX
Simplified Management w/ Scale

VM-FEX
VM-Aware Networking

DCB/FCoE
Consolidated I/O

vPC
Active-Active Uplinks

VDC
Virtualizes the Switch
DC Fabric Evolution
Why Layer 2 in the Data Center?

Because customers request it!

- Application requirement – Clusters, vMotion
- Plug and play
- No addressing
- Easy server provisioning
- Virtual machine mobility
- Inspiration from SAN fabric
Possible Solution for End-to-End L2

Just extend Spanning-Tree to the whole network
Some STP Layer 2 Limitations

- Local problems have network-wide impact, troubleshooting is difficult
- Blocked ports = limited bandwidth
- Tree topology introduces sub-optimal paths
- STP convergence is disruptive
- MAC address tables don’t scale
- Host flooding impacts the whole network

Only 1 link active between any two switches
Nexus Architecture Flexibility

Spanning Tree
- Single
- Up to 10 Tbps

Virtual Port Channel / VSS
- Dual
- Up to 20 Tbps

FabricPath
- 16 Way
- Up to 160 Tbps

Layer 2 Scalability
Infrastructure Virtualization and Capacity
vPC Terminology

- **vPC Domain** - A pair of vPC switches
- **vPC Peer** - A vPC switch, one of a pair
- **vPC port** - The port channel between the vPC peer and the downstream device
- **Orphan Port** – An orphan port is an interface which connects to an orphan device
- **vPC peer-link** - Link used to synchronize state between vPC peer devices, must be 10GE
- **vPC peer-keepalive link** - The keepalive link between vPC peer devices
**vPC Feature Overview**

- vPC allows a single device to use a port channel across two neighbor switches (vPC peers)
- Eliminate STP blocked ports & Provide fast convergence upon link/device failure
- Supports back-to-back connection of different VPC domains
- Layer 2 only port channel
- Available on Nexus 7000, 5000 and 3000

```
! Enable vpc on the switch
dc11-5020-1(config)# feature vpc

! Check the feature status
dc11-5020-1(config)# show feature | include vpc
vpc 1 enabled
```
vPC Member Port

Definition:
Port-channel member of a vPC

Requirements
Configuration needs to match other vPC peer member
In case of inconsistency a VLAN or the entire port-channel may be suspended (e.g. MTU mismatch)
Up to 16 active ports between both vPC peers with M series LC
Up to 32 active ports between both vPC peers with F series LC

NX7K-1:
interface port-channel201
  switchport mode trunk
  switchport trunk native vlan 100
  switchport trunk allowed vlan 100-105
  vpc 201

NX7K-2:
interface port-channel201
  switchport mode trunk
  switchport trunk native vlan 100
  switchport trunk allowed vlan 100-105
  vpc 201
vPC Peer-Keepalive Link up & Peer-Link down

vPC peer-link failure (link loss):
- Check active status of the remote vPC peer via vPC peer-keepalive link (UDP heartbeat)
- If both peers are active, then Secondary vPC peer will disable all vPCs to avoid Dual-Active
- Data will automatically forward down remaining active port channel ports
- Orphan devices connected to secondary peer will be isolated

```bash
dc11-5020-3(config-vpc-domain)# role priority ?
<1-65535> Specify priority value

dc11-5020-3# sh vpc
<snip>
vPC role          : secondary, operational primary
```
Data-Plane Loop Avoidance with vPC

Data-Plane vs. Control-Plane Loop control
- vPC peers can forward all traffic locally
- Peer-link does not typically forward data packets (control plane extension)
- Traffic on the Peer-link is marked and not allowed to egress on a vPC
- Exception for single-sided vPC failures
- Exception for Orphan ports
- Implications for Dynamic L3 & Mcast
- VSS merging control-plane

<table>
<thead>
<tr>
<th></th>
<th>VSS</th>
<th>VPC 5k</th>
<th>VPC 7k</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3 FHRP</td>
<td>ok</td>
<td>ok</td>
<td>ok</td>
</tr>
<tr>
<td>L3 Dyn.</td>
<td>ok</td>
<td>ok</td>
<td>Q1Y13</td>
</tr>
<tr>
<td>L3 Mcast</td>
<td>ok</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>ok</td>
<td>ok</td>
<td>ok</td>
</tr>
</tbody>
</table>
Edge Best Practices with Enhanced vPC

- Ethernet traffic gets equally distributed on FEX uplinks to both switches
- FCoE traffic is pinned only to uplink to one switch (for fabric isolation)
- Uplink carrying FCoE end up carrying more traffic (50% Eth + FCoE) than other uplink (50% Eth)
Architecture of vPC and FabricPath with vPC+

- Provides active/active HSRP for both sides FP and CE
- Configuration similar to standard VPC
- Key differences are addition of Virtual Switch ID and Peer Link is a FP Core Port
IRF - Intelligent Resilient Framework.

- No STP. Can support fixed switch stacking from 2 to 9 switches. (Broadcom) (FlexStack - 2960 equivalent)
- More switches in stack deliver unacceptable failover times.
- If a redundant Sup. fails, you lose 50% switching capacity.

<table>
<thead>
<tr>
<th>Failover Results Summary</th>
<th>vPC</th>
<th>IRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quad Supervisor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unicast N→S Failover Time (s)</td>
<td>0.284</td>
<td>436.1</td>
</tr>
<tr>
<td>Unicast S→N Failover Time (s)</td>
<td>0.056</td>
<td>441.4</td>
</tr>
<tr>
<td>Multicast Failover Time (s)</td>
<td>0.083</td>
<td>504.2</td>
</tr>
</tbody>
</table>

7 minutes
7 minutes
8 minutes
Cisco FabricPath externally looks like a single switch

Internally, a protocol adds Fabric-wide intelligence

- Eliminate Spanning tree limitations
- MultiPathing across all links, high cross-sectional bandwidth
- High resiliency, faster network re-convergence
- Any VLAN, any where in the fabric eliminate VLAN Scoping
• Shortest path Any to Any switch in fabric
• 16-Way Equal Cost MultiPathing (ECMP) at Layer 2
• Conversational MAC Learning – port MAC address table, unlimited hosts connected
• Interoperability with existing classic Ethernet networks – vPC+ active/active device attached
Cisco FabricPath Feature Set – II.

- 2-way simultaneous Active Default Gateways with VPC+
- Multiple Simultaneously Active Default Gateways with Anycast FHRP
- Unlimited VLAN extension, no risk of loops
- Multi-Topology – providing traffic engineering capabilities
New Control Plane

- IS-IS assigns addresses to all FabricPath switches automatically
- Compute shortest, pair-wise paths
- Support equal-cost paths between any FabricPath switch pairs
New Data Plane

- The association MAC address/Switch ID is maintained at the edge
- Traffic is encapsulated across the Fabric
FabricPath Interfaces Definition

**FabricPath Port**
- Interfaces connected to another FabricPath P.
- Send/Receive traffic with FabricPath header
- No Spanning-Tree!!!
- No ‘MAC Learning’
- Exchange topology info via L2 ISIS Adjacency
- Forwarding based on ‘Switch Table’

**Classic Ethernet (CE) Port**
- Interfaces connected to all existing NICs and Network Devices
- Send/Receive traffic in 802.3 Ethernet frames format
- Participated in STP domain
- Forwarding based on MAC Table
FabricPath Encapsulation

Classical Ethernet Frame

- **Switch ID** – Unique number identifying each FabricPath switch
- **Sub-Switch ID** – Identifies devices hôsts connected via VPC+
- **LID** – Local ID, identifies the destination or source interface
- **Ftag** (Forwarding tag) – Unique number identifying topology and/or distribution tree
- **TTL** – Decrementated at each switch hop to prevent frames looping infinitely
Transparent Interconnection of Lots of Links (TRILL)

• IETF standard for Layer 2 multipathing
• Driven by multiple vendors, including Cisco
• TRILL is now officially a Proposed Standard in IETF
• Nexus 7000 and 5000 will provide a TRILL mode with a software upgrade
• No vPC+ A/A capability in TRILL, only A/S
• No multiple topologies in TRILL

http://datatracker.ietf.org/wg/trill/
FabricPath vs. TRILL: Encapsulation
Juniper Q-Fabric / Competitive?

- Qfabric limits just 100 meter radius
- Multiple isolated partitions require dedicated director, one QF-Nore can be in one partition
- No ISSU, QF-Nodes reboots simultaneously
- 6-8s failover time
- QF-Director becomes a control plane bottleneck for L2, ARP, IGMP, PIM
- Qfabric – the $64,000 question

Juniper Networks Inc. (JNPR) is eliminating 500 jobs, or about 5 percent of its workforce, in an effort to trim expenses amid accelerating competition from networking-equipment makers led by Cisco (CSCO) Systems Inc., people with knowledge of the matter said.

The jobs being cut include some engineers involved with a set of products called QFabric, said the people, who asked not to be identified because the company hasn’t said which operations are affected.
Simplicity in DC operations
Consolidation with Virtual Device Context / VDC

- Consolidation of Core and Aggregation while maintaining network hierarchy
- No reduction in port count or links but fewer physical switches
  Copper Twinax cables (CX-1) provide a low cost 10G interconnect option
Management and CAPEX savings

- Using VDCs for OTV VDC or dedicated MPLS PE for PODs
- VDCs can provide a migration strategy to new hardware and line cards
- VDCs provide consolidation and separation that makes storage administrators comfortable – virtual MDS

<table>
<thead>
<tr>
<th></th>
<th>Traditional</th>
<th>Consolidated with VDC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4x C6509E-S2T</td>
<td>2x C7009-S2E</td>
</tr>
<tr>
<td></td>
<td>10x X6816-10G</td>
<td>2x M224XP</td>
</tr>
<tr>
<td></td>
<td>104x 10G-SR</td>
<td>2x F248XP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>104x 10G-SR</td>
</tr>
<tr>
<td>CAPEX</td>
<td>$ 920.000</td>
<td>- 20%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$ 750.000</td>
</tr>
<tr>
<td>OPEX 5Y</td>
<td>$ 205.000</td>
<td>- 15%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$ 176.000</td>
</tr>
<tr>
<td>ENERGY 5Y</td>
<td>$ 55.200</td>
<td>- 56%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$ 24.570</td>
</tr>
<tr>
<td>ROI 5Y</td>
<td></td>
<td>25%</td>
</tr>
</tbody>
</table>
How Many VDCs Can I have?

- Supervisor 1 – VDCs (4+1)
  May require 8GB of RAM
- Supervisor 2 – VDCs (4+1)
- Supervisor 2E – VDCs (8+1)
  VDCs beyond 4 require additional license
  N7K-VDC1K9 (increments VDCs +4)
- Storage FCoE require dedicated VDC
- M1 + M2 + F1 + F2e in one VDC
- F2 in one VDC
  - PCI Compliant, FIPS 140-2, CCEVS #10349
CPU Shares

- Enables per-VDC CPU Access and Prioritization
- Provides more control and protection per VDC for users
- CPU share is controlled by VDC priority
- CPU is shared equally among VDCs
- The more VDCs configured, the lower the overall percentage per VDC
- Comes into use when CPU utilization increases (contention)
- CPU share = (VDC priority * 1000)
- Available on SUP2/2E and NX-OS 6.1

VDC Shares:
- VDC1 Shares = 2
- VDC2 Shares = 4
- VDC3 Shares = 1
- VDC4 Shares = 8
- VDC5 Shares = 10
- VDC6 Shares = 5
Huawei CloudEngine Series/ Competitive?

**Huawei S9700** or S9300 Series
- Do not match VDC like capability

**Huawei 12800** Series
- Virtual System (VS)
- Maximum of 1:8 VS on one physical chassis (means 8 virtual instances)
- Presented at Interop in Las Vegas May of 2012 – demo version
- Product will be available at the end of 2012
- TRILL, IEEE DCB, FCoE, CSC Clustering

Must have seen it before, maybe on previous slide? Bingo…
Unified Ports and FCoE

- Allows repurposing of any switch port as either Fibre Channel or Ethernet
- Fewer cables and adapters
- Wire Server Once
- Software Provisioning of I/O
- Rapid and seamless transition from FC to FCoE
- Simplifies the purchasing decision
- Allows IT to respond faster to changing business needs
FCoE Protocol Fundamentals

FCoE does not change the explicit port level relationships between devices

- Servers (VN_Ports) connect to Switches (VF_Ports)
- Switches connect to Switches via Expansion Ports (VE_Ports) or (VNP_Ports)
Converged Access

- Shared Physical, Separate Logical at Access Layer
- Separate Physical and Logical at Aggregation Layer
- Storage VDC on Nexus 7000
- Unified Ports on Nexus 5500
- Blades with B22, Nexus 4000, Nexus 2000 or competitive option
- **Edge-Core** Topology
- Savings on HBAs, cables, access switches 30-40% depend on the number of servers connected
- Unchanged operational mode Storage / LAN
Converged Network Fabrics w/ Dedicated Links

- Access and Aggregation share physical switches
- All switches are FCoE FCF switches
- Storage VDC Agg, Nexus 5k at acces
- SAN can utilize higher performance, higher density, lower cost Ethernet switches for the aggregation/core
- **Edge-Core-Edge** Topology
- Access for new 10G CEE servers to old FC storage/tape
Let’s see well know SAN vendor / Competitive?

**Brocade FCOE10-24 Blade for DCX and DCX-4S**
- No VE_port support -> limitation of multi-hop FCoE, DC interconnect over FCoE
- No ISSU support
- No L3 support, ISL Trunk max. 8 ports

**Brocade 8000 Fixed** Top-Of-Rack Switch
- No Unified Ports, ability to turn switch into FC regular switch
- No STP neither RSTP, no FCoE N-Port Virtualization
- No L3 support, ISL Trunk max. 8 ports
Cisco FEX – Virtualized Access Switch

Cisco Nexus® 5500 +Cisco Nexus® 7000

Cisco Nexus® 2000 FEX

Cisco Nexus® 2000 FEX

Distributed High Density Edge Switching System (up to 4096 virtual Ethernet interfaces)
Fabric Extender Technology (FEX)

- Single point of management
- IEEE 802.1BR*
- Consolidates network management
- FEX managed as remote line card
- Extends Cisco unified fabric into OEM partner blade chassis
- Adapter FEX
- Consolidates multiple 1GbE interface into a single 10GbE interface
- Extends network into server VM-FEX
- Consolidates virtual and physical network
- Each VM gets a dedicated port on switch

Distributed Modular System for Virtual Machine

ONE NETWORK
- Parent switch to top of rack
- Parent switch to adapter
- Virtual same as physical

- Single point of policy
- Single point of management
- Reduction in cables
- Consistency across rack and blade servers
- Interoperable—standards-based
Automation, virtualization and high-availability
NX-OS Statefull Process Restart

- NX-OS services checkpoint their runtime state to the PSS for recovery in the event of a failure

If a fault occurs in a process…
- HA manager determines best recovery action (restart process, switchover to redundant supervisor)
- Process restarts with no impact on data plane
- Total recovery time: ~10s ms
- Statefull restart/NSF for OSPFv2/3, BGP, EIGRP, IS-IS, FHRP
In-Service Software Upgrade

N7K# install all kickstart bootdisk:4.1-kickstart system bootdisk:4.1-system
N7K#

Upgrade and reboot
Initiate stateful failover

Upgrade and reboot
Upgrade and reboot I/O modules

Needed for animation, don’t remove!

I/O Module Images

N7K Data Plane

HA Manager
Linux Kernel
N7K#

Release 5.1

OSPF BGP PIM etc.
HA Manager
Linux Kernel

Active

OSPF BGP PIM etc.
HA Manager
Linux Kernel

Standby

Release 5.1

Release 5.2

Upgrade and reboot I/O modules

Upgrade and reboot
Managing the Fabric
Nexus Port Profiles – Provisioning

- A new model of provisioning the access ports
- Enables the application of **common configuration** across groups of ports
- A port-profile can **inherit attributes** from other port-profiles (nested profiles)
- A change to a port-profile **automatically updates configuration** of all member ports
- Any interface command available on a Nexus interface can be a part of a port-profile e.g. ACL, L3, VLAN, etc.
- Configuration precedence/order:
  - Default config. < Port-profile < Manual config.
Programmability – SDN
Potential growth in System Control Plane Capabilities

Network Automation

On-Box Scripting
- TCL
- PERL
- Python

XML Mgmt. Interface
- NETCONF & YANG models
- XMPP, REST

onePK
- SDK/APIs in multiple languages.
- Extensibility & Programmability

Network Admin
- Quick automation
- Manual or home grown scripts

NM App Developer
- NM app/OSS
- Standards driven
- Typical net mgmt.

App Developer
- VARs/Partners/CU
- Prog. APIs
- Extend and leverage
DCNM DEMO
Resume

• DC Fabric Evolution
  Exceptional performance while maintain High Availability
  Zero Service Disruption Upgrades

• Simplicity in DC operations
  Switch Virtualization via VDC
  Consolidate network layers for medium-business
  Converged networks
  Remote line card FEX architecture model

• Virtualization and High Availability
  Automation support
  High Availability with Zero packet loss
  Virtualization
Knowledge Test

Pre aký jediný protokol NX-OS nepodporuje statefull-restart/nonstop forwarding (NSF)?

A. BGP
B. OSPFv2
C. IS-IS
D. RIP
Thank you.