vPC Failure Detection & Recovery

In This Section

- Correct vPC Failure Order of Operations
- vPC Failure Problem Cases
- vPC Failure Design Workarounds

vPC Initialization Order of Operations

- vPC process starts
- IP/UDP 3200 Peer Keepalive connectivity established
- Peer-Link adjacency forms
- vPC Primary/Secondary role election
- vPC Consistency Checks performed
- Layer 3 SVIs move to up/up state
- vPC member ports move to up/up state

vPC Consistency Checks

- vPC Peers sync control plane over Peer Link with Cisco Fabric Services (CFS)
- Includes advertisement of "Consistency Parameters" that must match for vPC to form successfully
  - E.g. Line Card Type (M or F), Speed, Duplex, Trunking, LACP mode, STP configs, etc.
- Three types of Consistency Checks
  - Type 1 Global
    - Mismatch results in vPC failing to form
    - E.g. STP mode Rapid-PVST vs. MST
  - Type 1 Interface
    - Mismatch results in VLANs being suspended on vPC member
    - E.g. STP port type network vs. normal
  - Type 2
    - Mismatch results in syslog message but not vPC failure
    - Can result in failures in the data plane
    - E.g. MTU mismatch
- Verification
  - show vpc
  - show vpc consistency-parameters

Graceful Consistency Check

- Consistency failure results only vPC Secondary disabling vPCs
  - 50% bandwidth reduction in favor of 0% packet loss
- Enabled by default in 5.2 and later
  - **show vpc**

Modifying vPC Initialization

- vPC Primary/Secondary election
  - Can be influenced by **role priority**
  - Lower value is better
- Layer 3 SVI activation
  - Timer controlled **delay restore interface-vlan**
- vPC member port activation
  - Timer controlled by **delay restore**

vPC Member Port Failure Detection

- vPC Peers exchange vPC Member status over Peer Link
- Failed Member Ports result in “Orphan Ports”
  - Orphan Ports are single attached ports that use a vPC VLAN
  - vPC VLANs are any VLANs allowed on the Peer Link
  - show vpc orphan-ports

vPC Orphan Ports

- Traffic to Orphans use vPC Peer Link as a last resort
  - Orphan Ports use modified loop prevention
  - Traffic from remote Orphan is allowed to enter Peer Link and exit via local Member
  - Traffic from remote Member is allowed to enter via Peer Link and exit via local Orphan
  - Traffic from remote Member is not allowed to enter via Peer Link and exit via local Member
- Orphan ports should be avoided at all costs
  - vPC Peer Link is the bottleneck of the system
    - Should be used only for control plane under ideal circumstances
- Orphan ports can result in traffic black holes
  - Orphans connected to vPC secondary can be isolated from their default gateway if vPC Peer Link fails
  - More on this later...

vPC Peer Link Failures

- vPC Peer link synchronized control plane between vPC Peers
  - I.e. which MAC addresses are reachable via which vPCs
- Peer Link failure can result in “Split Brain”
  - Split Brain is when control plane sync is broken and both vPC Peers assume vPC Primary Role
  - Active/active forwarding improperly occurs and results in black hole or layer 2 loop
- vPC Peer Keepalive and Peer Link work together to prevent this

vPC Peer Link Failure Detection

- vPC Peer Link fails
  - E.g. line card outage
- vPC Secondary pings Primary over Peer Keepalive
  - If vPC Primary is alive…
    - Disable vPC member ports on Secondary
    - Disable SVIs on Secondary
    - Goal is to force end host to forward via Primary
  - If vPC Primary is dead…
    - Promote vPC Secondary to Operational Primary
    - Continue to forward traffic on new Primary
- Peer Keepalive and Peer Link must not share fate in order to prevent Split Brain
  - E.g. separate MGMT switch, separate Port Channels on separate line cards

vPC Auto Recovery

- Certain failures can result in neither vPC Peers forwarding
- Power outage with node failure problem case
  - Power outage on both Peers
  - Only one Peer is restored
  - vPC Peer Keepalive never comes up
  - Means vPC Peer Link can never come up
  - Means vPC Member Ports can never come up
  - Servers are isolated
- vPC Auto Recovery allows single Peer to promote itself to Primary
  - If Peer Link does not initialize before Auto Recovery timeout, promote myself to Primary and bring up Member ports

vPC Auto Recovery (cont.)

- Gradual failure problem case
  - vPC Peer Link goes down
  - vPC Secondary pings vPC Primary and gets response
  - vPC Secondary disables vPC Member Ports
  - vPC Primary completely fails
  - vPC Secondary does not re-activate Member Ports
  - Servers are isolated
- vPC Auto Recovery allows Secondary to detect this
  - vPC Primary is continually tracked over vPC Peer Keepalive
  - Peer Keepalive failure at later time results in Secondary promoting itself to Primary
  - Secondary re-activates Member Ports

Peer Keepalive & Peer Link Fate Sharing

- Fate sharing problem case
  - Keepalive configured via Layer 3 SVI
  - SVI VLAN is allowed on Peer Link
  - STP always prefers Peer Link
  - Peer Link fails, but Primary still up
  - No layer 2 path for SVI exists or Secondary disables SVI
  - Secondary cannot ping Primary
  - Secondary promoted to Operational Primary
  - Split Brain occurs

Peer Keepalive & Peer Link Path Diversity

- Keepalive and Peer Link must have path diversity to ensure proper failure detection
  - Peer Link failing must not affect Keepalive
  - I.e. not the same Shared Risk Link Group (SRLG)
- Example fixes:
  - Keepalive on mgmt0 with separate OOB MGMT switch
    - Assumes mgmt0's are not wired directly to each other
  - Keepalive on separate Layer 3 Port Channel
    - Assumes port density is available
  - Keepalive as subinterface of East/West routing peering link
    - Reduces need for port density
  - Keepalive as SVI that is not a vPC VLAN
    - Assumes alternate layer 2 path besides Peer Link

vPC Peer Link Failure & Orphan Isolation

- Problem case:
  - vPC Primary and Secondary are default gateways for vPC VLAN
  - Orphan Port exists on Secondary
  - vPC Peer Link fails, but Primary remains up
  - Secondary pings Primary, gets response, and…
    - Disables vPC Member Ports
    - Disables SVIs
  - Orphan is now isolated from its default gateway

vPC Peer Link Failure & Orphan Isolation

- Example fixes:
  - Dual home all end hosts
    - I.e. don't have Orphans
  - Single attached hosts could attach via dual-homed access switch
    - E.g. IPMI/ILO ports goto 3750, 3750 dual homes to vPC Peers
  - Single attached ports could use non-vPC VLANs
    - Port only counts at Orphan if using vPC VLANs
    - Non-vPC VLANs require additional East/West trunk between vPC Peers
  - Don't disable SVI when Peer Link fails on Secondary
    - "dual-active exclude interface-vlan"

vPC Peer Link Failure & Orphan A/S Failover

- Problem case:
  - Active/Standby Failover device connects via Orphans
    - E.g. Firewall, Load Balancer, etc.
  - Active device connects to vPC Secondary
  - vPC Peer Link fails, but Primary remains up
  - Secondary pings Primary, gets response, and…
    - Disables vPC Member Ports
    - Disables SVIs
  - Active device sees port as still up/up and does not failover
  - Active device is now isolated from its default gateway
- Example fixes:
  - Run active/active not active/standby
    - E.g. ASA Clustering, etc.
  - Dual home the device
    - Both active and standby connect to vPC Primary and Secondary via vPC Member Ports
  - Force the active device to failover to the vPC Primary
    - Interface level "vpc orphan-port suspend"

vPC Peer Link & Northbound Routing

- Problem case:
  - Peer Link and northbound routing links share same linecard
  - Peer Keepalive does not fate share with Peer Link
  - vPC Primary linecard fails
    - Peer Link is lost
    - Northbound routing links are lost
  - Secondary pings Primary, gets response, and…
    - Disables vPC Member Ports
    - Disables SVIs
  - Layer 2 traffic is collected via Primary, but cannot route to WAN
  - Servers are isolated

vPC Peer Link & Northbound Routing

- Example fixes:
  - Keepalive, Peer Link, & Routing Links should not share fate
    - Might not be possible based on port density
  - Enhanced Object Tracking on Primary
    - If Peer Link && WAN == Down, failover to Secondary
  - vPC Self Isolation
    - If Peer Link && WAN == Down, tell Secondary over Keepalive
    - Available in NX-OS 7.2 and later

In Today's Class

- vPC & FHRPs
- vPC Peer Gateway
- Fabric Extenders
- Managing C-Series Servers Using the Cisco Integrated Management Controller (CIMC)
- Enhanced vPC (EvPC)
- FCoE over Enhanced vPC

vPC and FHRPs

- vPC Peer connected towards access layer is typically L2 & L3 network boundary
- For gateway redundancy, vPC Peers run a First Hop Redundancy Protocol (FHRP)
  - i.e. HSRP, VRRP, or GLBP
- FHRP behavior changes to accommodate active/active forwarding over vPC
  - Traffic received in vPC Member Port of FHRP Standby to FHRP Virtual MAC is not forwarded over Peer Link to Active FHRP member
  - Both vPC Peers proxy for VMAC regardless which one is active/master router
  - Result is active/active forwarding without any special configuration
  - Better traffic flow distribution than GLBP

vPC Peer Gateway

- With vPC + HSRP, HSRP Standby acts the same as HSRP Active
- FHRP vPC can break in certain non-standard vendor applications
  - F5 Auto-Last Hop, EMC Packet Reflect, etc.
  - Load balancers attempt a Layer 2 SNAT in order to prevent asymmetric flows
- In vPC, frames sent to FHRP Standby with physical DST MAC of FHRP Active are sent out the Peer Link
  - If final destination is another vPC Member, traffic is dropped
- Possible solutions
  - Disable Auto-Last Hop, Packet Reflect, etc.
  - Use "peer-gateway" to proxy for physical and virtual MACs on both vPC Peers

Recommended Reading

- Design and Configuration Guide: Best Practices for Virtual Port Channels (vPC) on Cisco Nexus 7000 Series Switches
- BRKDCT-2378 - VPC Best Practices and Design on NX OS
- BRKCRS-3146 - Advanced VPC operation and troubleshooting