BROCADE

RESILIENT PROTOCOL APPLICATION AND DEPLOYMENT CONSIDERATIONS

Muhammad Durrani (mdurrani@brocade.com)

Principal Engineer, Brocade Communications Inc.

Daisuke Touda (touda@mfeed.ad.jp)

Senior Manager, IMF (Internet Multi-Feed Japan)

ABSTRACT

Technology plays an essential role in building network resiliency, Technology is available to support flexible and resilient networks, The most critical technology components you should consider when planning for resiliency are data , application, networks, security and end user devices. This presentation will discuss deployment challenges to achieve resiliency at various OSI layer, their advantages and limitations and available technologies to meet the challenges along with various deployment case studies

Agenda

- Overview
- Layer 1 Resiliency
- Layer 2 Resiliency
- Layer 3 Resiliency
- Deployment Case Study
- Conclusion

Overview

Resiliency	Requirement	Feature
Layer 1	SP needs to offer 100% availability in case of layer 1 failures	Optical Switching (OS)
	Minimize outage in order of milli-seconds by detection of optical signal strengths (Optical switching) and upon TX or RX fiber cut aka @ layer 1 LOS	Link Fault Signaling (LFS)
Layer 2	MAC layer redundancy to avoid or Minimize MAC learning upon device or link failure	802.3ad-Extensions
Layer 3	Achieve milli-seconds convergence for layer 3 protocols	Bi-directional forwarding detection (BFD)
	Fast re-route in MPLS layer	MPLS Fast-Reroute – Protection against Link and Router failure

Overview – Cont ..

- Deployment Case Studies

  - Layer 1:
    - Deploying LFS along with Optical Switching module to achieve layer 1 resiliency
  - Layer 2:
    - Deploying Layer 2 / Layer 3 Data Center using 802.3ad-Extension aware VPLS to extend layer 2 network across geographical separated layer 2 domains over MPLS cloud
  - Layer 3:
    - Deploying Bi-directional fault detection to achieve layer 3 Protocol resiliency in conjunction with MPLS fast re-route link protection.

Agenda

- Overview
- Layer 1 Resiliency
- Layer 2 Resiliency
- Layer 3 Resiliency
- Deployment Case Study
- Conclusion

Layer 1 Resiliency – Link Fault Signaling and OS Problem Definition

- SP requires protection at layer 1 to detect failures in milli-seconds on Ethernet Media where layer 1 alarms are not available unlike SONET

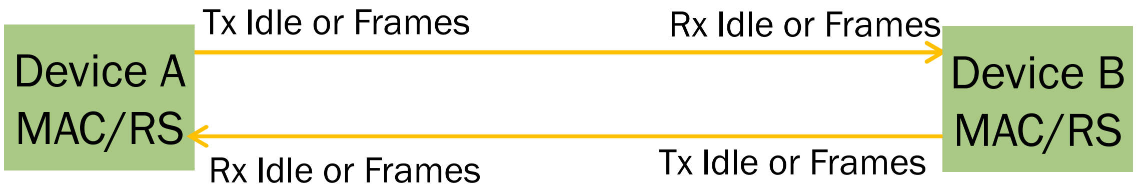
- Needs to have a way to signal remote end when Partial failure happens in layer 1 Ethernet media

- Avoid traffic black holing as soon as layer 1 failure happens
  - Partial or complete layer 1 media failure

- Switch to backup path once active path fails by detecting optical signal levels
  - Single Link Failure
  - Link Aggregation failure (with and without LACP)

Layer 1 Resiliency – Link Fault Signaling Concept

- Link Fault Signaling:
  - Link fault signaling (LFS) is a physical layer protocol that enables communication on a link between two (1 / 10 or 100G) Ethernet devices.

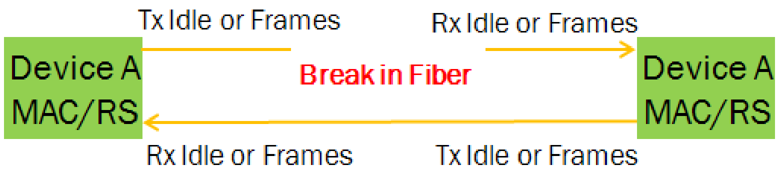


- Device A and B both powered up and operating properly
- Both Devices are capable of Transmitting MAC frames

Layer 1 Resiliency – Link Fault Signaling Operations

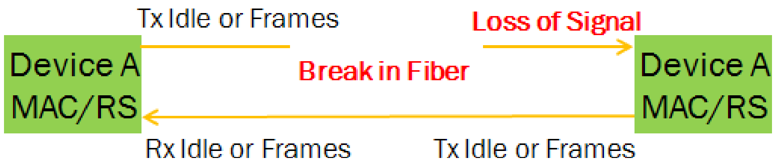
- LFS Fault Operation:

  - Break in RX fiber of Device B



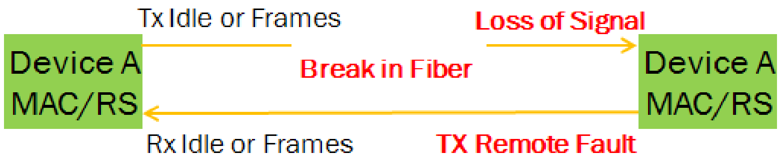
Tx Idle or Frames
Rx Idle or Frames
Device A MAC/RS
Break in Fiber
Device A MAC/RS
Rx Idle or Frames
Tx Idle or Frames

  - Device B detects loss of signal. Local fault is signaled by PHY of Device B to RS of Device B.



Tx Idle or Frames
Loss of Signal
Device A MAC/RS
Break in Fiber
Device A MAC/RS
Rx Idle or Frames
Tx Idle or Frames

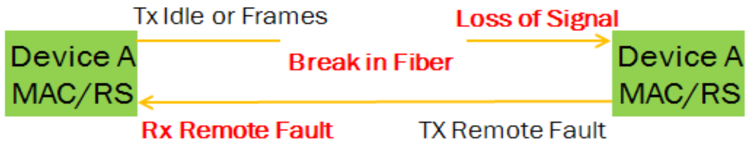
  - RS of Device B ceases transmission of MAC frames and transmits remote fault to Device A.



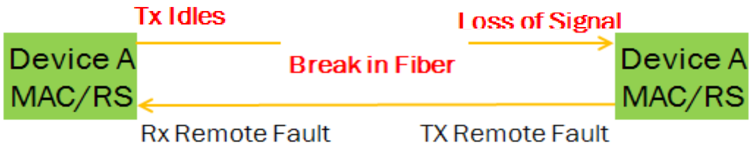
Tx Idle or Frames
Loss of Signal
Device A MAC/RS
Break in Fiber
Device A MAC/RS
Rx Idle or Frames
TX Remote Fault

Layer 1 Resiliency – Link Fault Signaling Operations

- Device A receives remote fault from Device B.



- Device A stops sending frames, continuously generates Idle



Layer 1 Resiliency – Optical Switching Concepts

- Operator has to configure protection switching between active and backup path

  Backup path remains DOWN unless switching happens

- Optical switch operates in

  Automatic switching mode

  Manual switching mode

- Optical switch switching mechanism includes

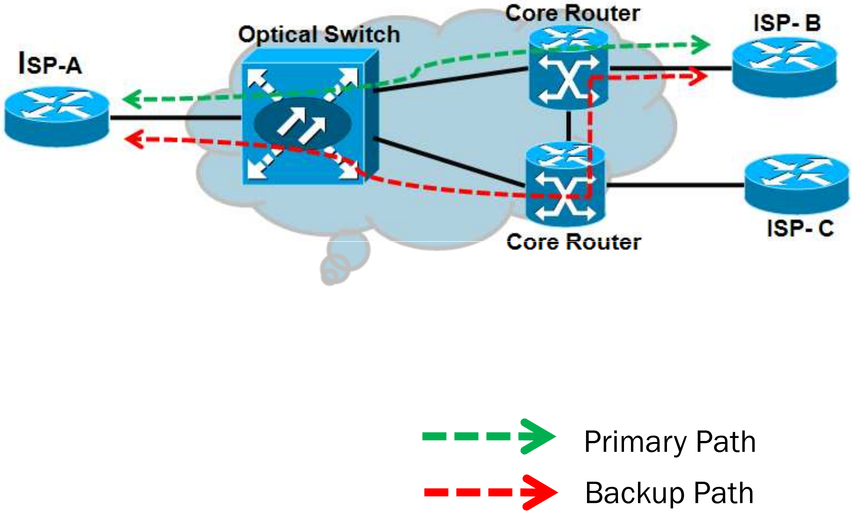
  If optical signal strength is below threshold

  If layer detects LOS

- Layer-1 Optical Switching Protects

  Single Physical Link

  Link Aggregation with and without LACP enabled



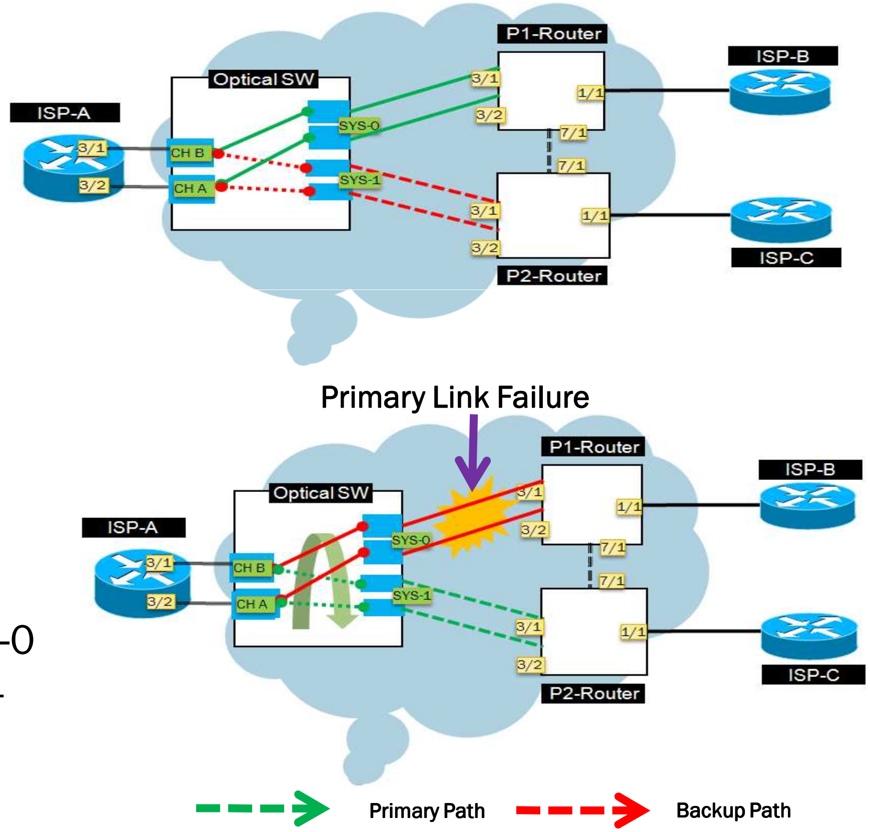
Layer 1 Resiliency – Optical Switching Operations

- One link from each CH-A and B connects to primary path to top P-router

- One link from each CH-A and B connects to backup path to bottom P-router

- Back up Path remains down because OS will connect ISP-A TX/RX to SYS-0
  - SYS-1 will remain un-connected hence link will remain down due to LOS

- Optical Switch once detects (in ms)
  - LOS @ SYS-0
  - Optical signal degradation below threshold at SYS-0
    - automatically connects ISP-A TX/RX to SYS-1
    - This brings up SYS-1 physical layer
    - Traffic continue to flow with ms drop



Layer 1 Resiliency – Deployment Case Study

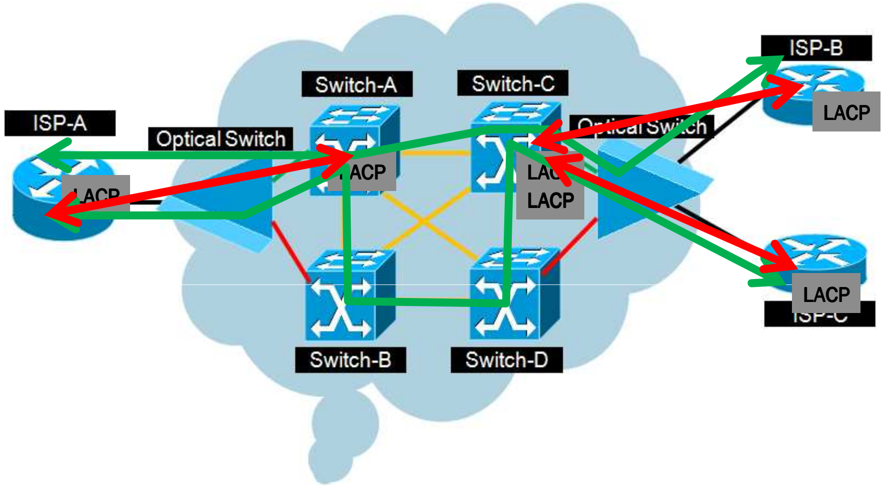
- Requirements
  - Minimize management overheads
  - Provide Link layer redundancy
  - Customer traffic should not be black holed
  - Active and backup path should be available
    - Switch traffic to backup path iff
      - Primary Path fails
      - HW & SW upgrades to nodes in Primary Path
  - Exchange Point Switching/recovery should be in tens of milli-seconds

- Solution
  - Select hardware based solution
  - 802.3ad LAG with LACP short option
  - Link Fault Signaling (LFS)
  - Optical switching
    - hardware based solution
  - LFS with LACP LAG along with optical switching

Layer 1 Resiliency – Deployment Case Study

- Ease of management
  Deploy hardware based solution

- Link Level Redundancy
  Deploy 802.3-ad LAG
  with LACP short
  with trunk-threshold



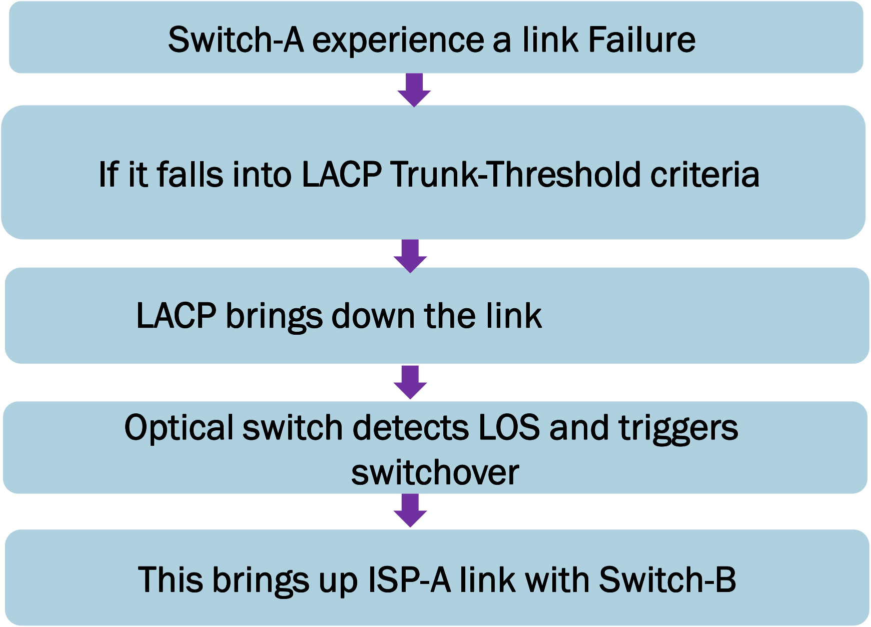
- Customer traffic should not be black holed
  Deploy Link Fault Signaling

- Active and backup path should be available

🟨 Switching Path
🟩 Primary Path
🟥 Backup Path

Layer 1 Resiliency – Deployment Case Study

- Automatic Exchange Point failure recovery
  - Sequence of events



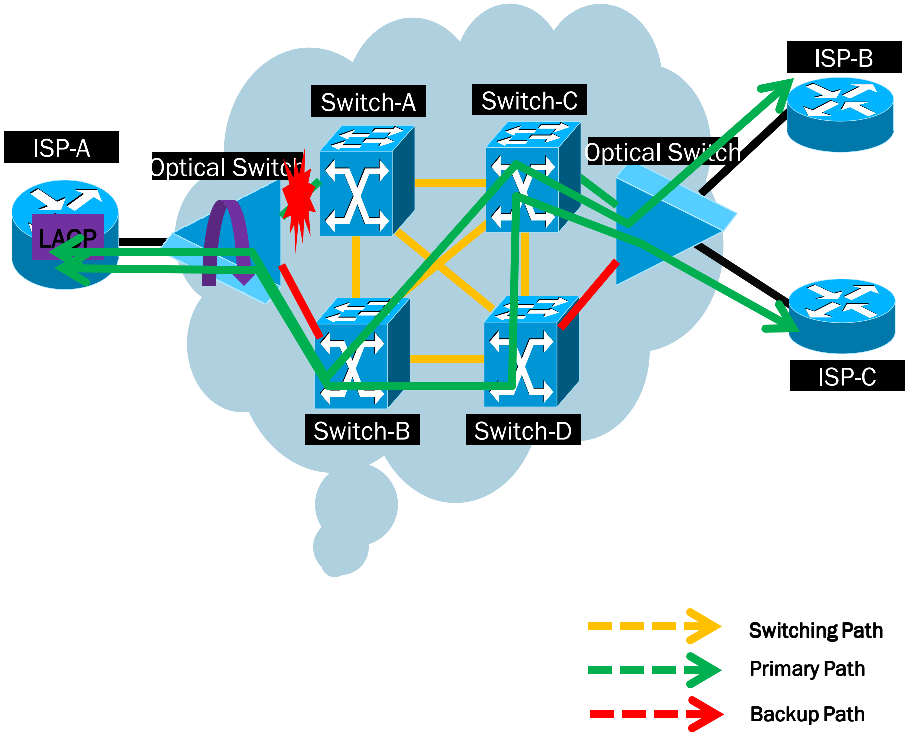
Switch-A experience a link Failure

If it falls into LACP Trunk-Threshold criteria

LACP brings down the link

Optical switch detects LOS and triggers switchover

This brings up ISP-A link with Switch-B



Take Away

- Layer 1 alarms are not available for Ethernet unlike Sonet
  - Mechanism's are in place to achieve similar in Ethernet world
- LFS provides mechanism to bring down link
  - When Partial failure detected (Either TX or RX fiber fails)
  - When Ethernet links are not directly connected
  - Avoid Traffic Black holing
- Hardware based optical switching
  - Provides faster switching mechanism (Hot switch over)
    - Upon link failures (LOS)
    - Upon optical signal degradation
    - Can be integrated with 803.ad (Link Aggregation) & LFS for robustness

Agenda

- Overview
- Layer 1 Resiliency
- Layer 2 Resiliency
- Layer 3 Resiliency
- Deployment Case Study
- Conclusion

Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad–Extensions – Problem Definition

- SP requires to provide link layer and switch level redundancy to their customers
- Provides traffic restoration in tens of milliseconds in case of link or switch failures.
- Allows servers and switches to have redundant connections to both Active and Backup switch and to fully utilize all links (including redundant ones) for traffic transport.
- Allows servers and switches to use standard link aggregation (802.3ad) to connect to redundant switches
- No MAC learning on servers and switches upon gateway device or link to gateway device fails

Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad-Extensions – Concept

- Dynamic LAGs

  Client creates a single dynamic LAG towards the Trunking nodes.

  For Trunking (Logical Switch A and B) nodes the dynamic Lag consists of two LAGs, each is configured on one of the Trunking devices.

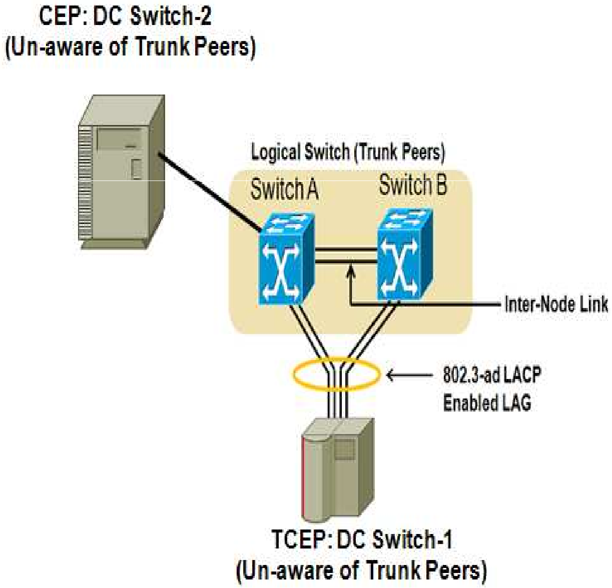
  A dynamic LAG runs Link Aggregation Control Protocol (LACP).

- Trunk peers

  Each trunk physical node, A and B, will act as an Trunk peer

  Trunk Peers are connected using an Inter-node link.

  The pair of Trunk nodes will act as one logical switch for the access switch or server so that the Trunk pair can connect using standard LAG to them.

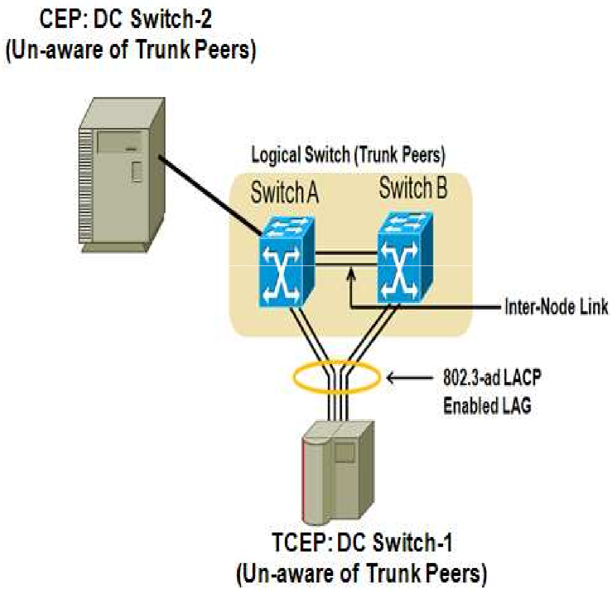

CEP: DC Switch-2
(Un-aware of Trunk Peers)
Logical Switch (Trunk Peers)
Switch A
Switch B
Inter-Node Link
802.3-ad LACP Enabled LAG
TCEP: DC Switch-1
(Un-aware of Trunk Peers)

Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad-Extensions – Concept

- Inter-Node link (INL) traffic handling
  An INL link can be a single port or a static or LACP LAG.
  Normal VLANs can co-exist with Trunk VLANs on the INL .
  For Inter-node VLANs, MAC learning is disabled on INL ports.

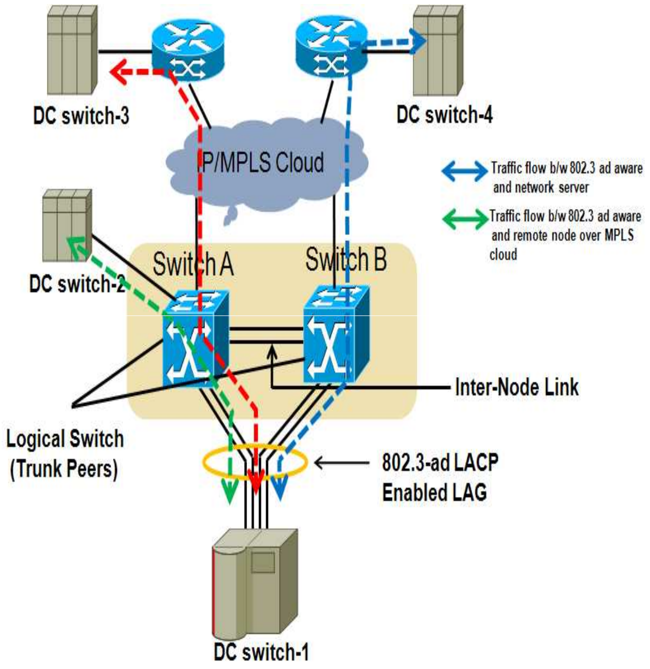
- Trunk Client End Point (TCEP)
  Device running 802.3-ad LAG with Trunk peers is called TCEP

- Client End Point (CEP)
  Device directly connected to Trunk peers is called Client End-Point


CEP: DC Switch-2
(Un-aware of Trunk Peers)
Logical Switch (Trunk Peers)
Switch A
Switch B
Inter-Node Link
802.3-ad LACP Enabled LAG
TCEP: DC Switch-1
(Un-aware of Trunk Peers)

Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad-Extensions – Concept

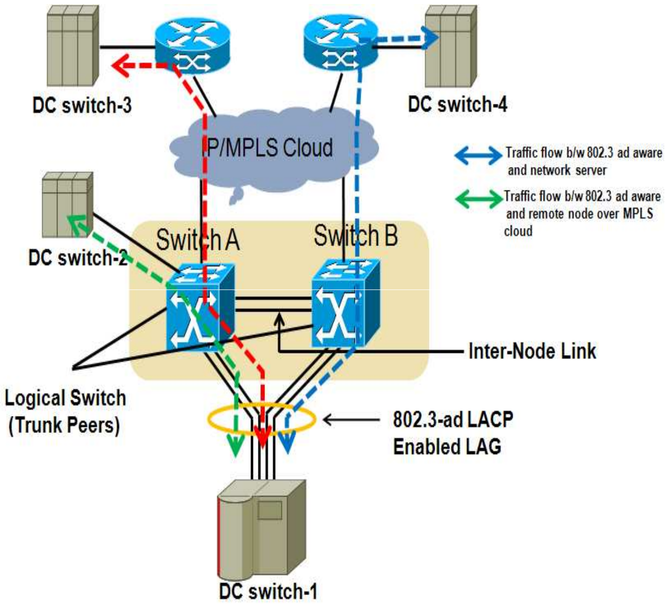
- Sub-second failover in the event of a link, module, switch fabric, control plane, or node failure
- Layer 2 and Layer 3 forwarding between 802.3-ad aware Node and Network server / remote node over IP/MPLS cloud
- Flow based load balancing rather than VLANs sharing across network links
- Ability to provide the resiliency regardless of the traffic type layer 3, layer 2 or non-IP (legacy protocols).


DC switch-3
DC switch-4
IP/MPLS Cloud
Traffic flow b/w 802.3 ad aware and network server
Traffic flow b/w 802.3 ad aware and remote node over MPLS cloud
DC switch-2
Switch A
Switch B
Inter-Node Link
Logical Switch (Trunk Peers)
802.3-ad LACP Enabled LAG
DC switch-1

- Provides nodal redundancy in addition to link and modular redundancy
- Operates at the physical level to provide sub-second failover

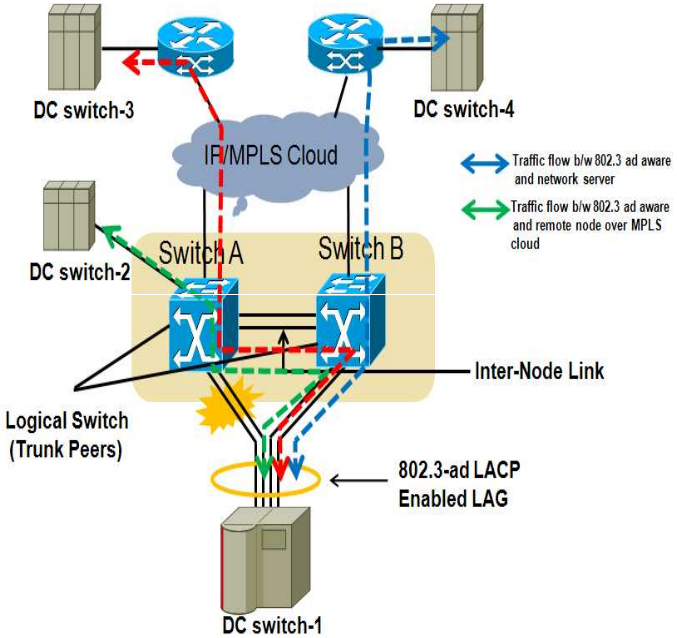
Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad-Extensions – Concept

- DC-Switch-2 and 3 source MAC is learned via Switch-A
- Traffic from DC Switch-1 to DC Switch 2 and 3 will flow via switch A
- DC-Switch-4 source MAC is learned via switch-B
- Traffic from DC Switch-1 to DC Switch 4 will flow via switch A
- MAC learned by Switch A is synched to Switch B using INL link and vice versa hence both switches will maintai same copy of MAC table
- Trunk peer runs MPLS/IP towards the cloud and layer 2 towards DC server farm (DC Switch-1)


DC switch-3
DC switch-4
P/MPLS Cloud
Traffic flow b/w 802.3 ad aware and network server
Traffic flow b/w 802.3 ad aware and remote node over MPLS cloud
DC switch-2
Switch A
Switch B
Inter-Node Link
Logical Switch (Trunk Peers)
802.3-ad LACP Enabled LAG
DC switch-1

Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad-Extensions – Concept

- When DC Switch-1 link to Switch fails
  only Source MAC of DC Switch-1 will be flushed
  DC Switch-3 and 2 will still be reachable via Switch-A
- Switch A will install DC Switch-1 reach-ability via Switch B learned via INL link without waiting for Data traffic to learn the source MAC
- DC Switch-1 will not know about the failure and will continue to forward traffic over available LAG link
- Traffic from DC Switch-1 to DC Switch-2 & 3 will flow via Switch-B → INL → Switch-A


DC switch-3
DC switch-4
IP/MPLS Cloud
Traffic flow b/w 802.3 ad aware and network server
Traffic flow b/w 802.3 ad aware and remote node over MPLS cloud
DC switch-2
Switch A
Switch B
Inter-Node Link
Logical Switch (Trunk Peers)
802.3-ad LACP Enabled LAG
DC switch-1

Layer 2 Resiliency – Deployment Case Study

- Requirements
  - Provide MAC layer redundancy
    - When Provider Edge Node failure occurs
    - When Provide Edge Node's link failure occurs
    - When Provider Edge routing failure occurs
    - Data-Center switch should not trigger MAC learning upon Edge device/Gateway failure
  - Solution should be applicable for
    - Geographically dispersed layer 2 domain over IP or MPLS domains
    - Locally collocated layer 2 domain connected via Provider Edge router

- Solution
  - 802.3-ad Extensions based solution
    - Trunk Peers provides Redundancy
      - Node level redundancy
      - Level redundancy
      - Routing and MPLS failures
      - Keep synch database of remote MAC eliminates the MAC learning for connected end devices
  - 802.3-ad Extensions based solution
    - Extends layer 2 domain and Provides connectivity over Layer 2 , IP or MPLS domains

Layer 2 Resiliency – Deployment Case Study

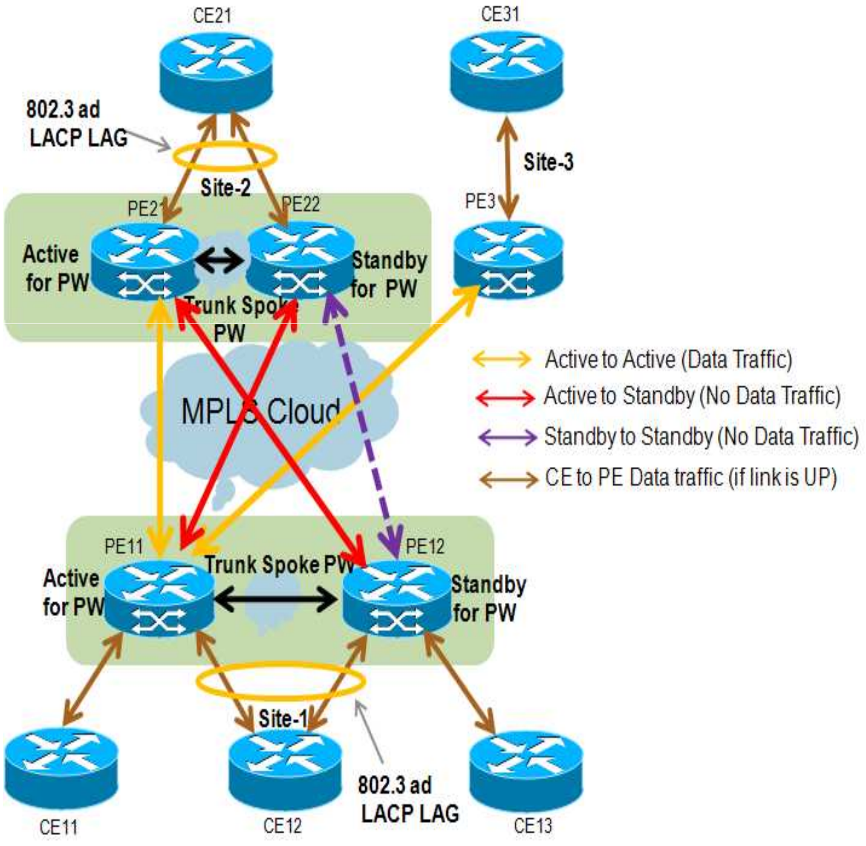
- Requirements
  - Continue interoperability with other vendor
    - 100% interoperable with third party vendors
    - No configuration changes required on remote nodes
    - Solution should be transparent to remote nodes
  - Faster convergence time
    - Convergence within 50ms
      - link and node failure
      - Layer 2 failures
      - IP and MPLS layer failure

- Solution
  - 802.3-ad Extensions based solution
    - 100% interoperable with third Party vendors
    - Configuration only needed on Trunk Peers
    - Remote nodes are not aware of Trunk Peers
    - Solution is transparent to remote nodes
  - 802.3-ad Extensions based solution
    - Trunk Peer link and Node failures are transparent to end node
      - No MAC learning triggered upon failures
      - In Conjunction with MPLS FRR and BFD provides ms convergence in IP / MPLS layer failure

Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad-Extensions – Deployment Case Study

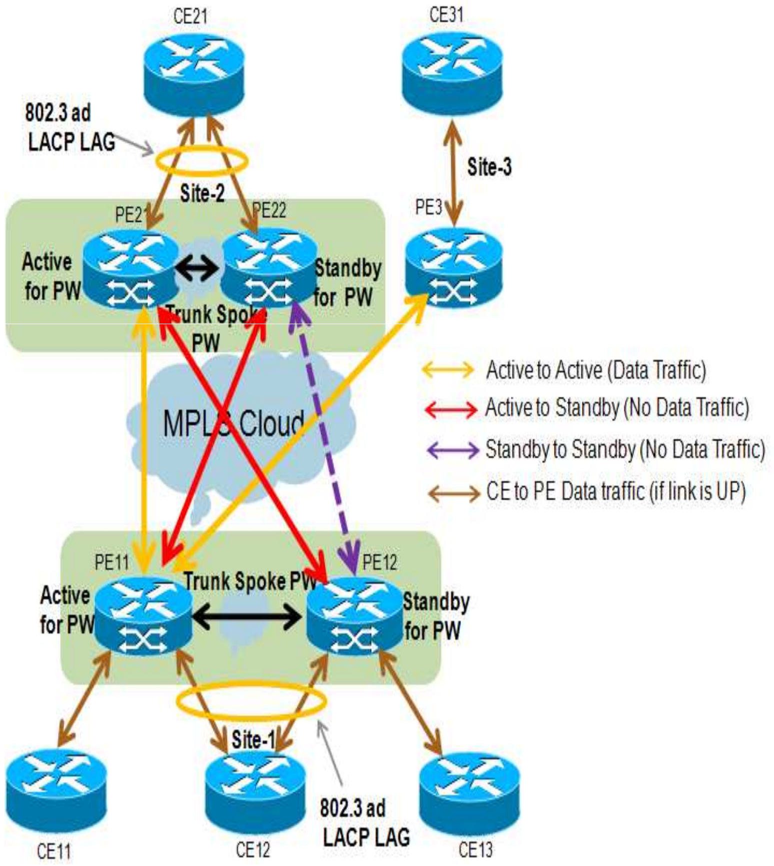
Topology Description:

- PE11 and PE12 are the two trunk Nodes.
- CE12 is connected to the trunk nodes using LAG.
- From Trunk nodes perspective the links connected to CE12 are called Trunk Client end-points (TCEP)
- CE11 and CE13 are single homed to PE11 and PE12 respectively. These are called Client end-points



Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad-Extensions – Deployment Case Study

- This solution requires MPLS connectivity between the two Trunk nodes and should be able to setup a VPLS Peering session.
- Remote PE's may not be aware of two co-located Trunk PE nodes but they need to be configured as two independent PE's.
- When one of the Trunk PE node fails it doesn't affect the connectivity for Trunk CEP end-points
- In the topology PE11 and PE12 could be treated as two separate peers from PE21, PE22 and PE3 point of view.



Layer 2 Resiliency – MAC Layer resiliency using 802.3 ad-Extensions – Deployment Case Study

- It is expected that PE11 and PE12 both have MPLS connectivity to remote PE's

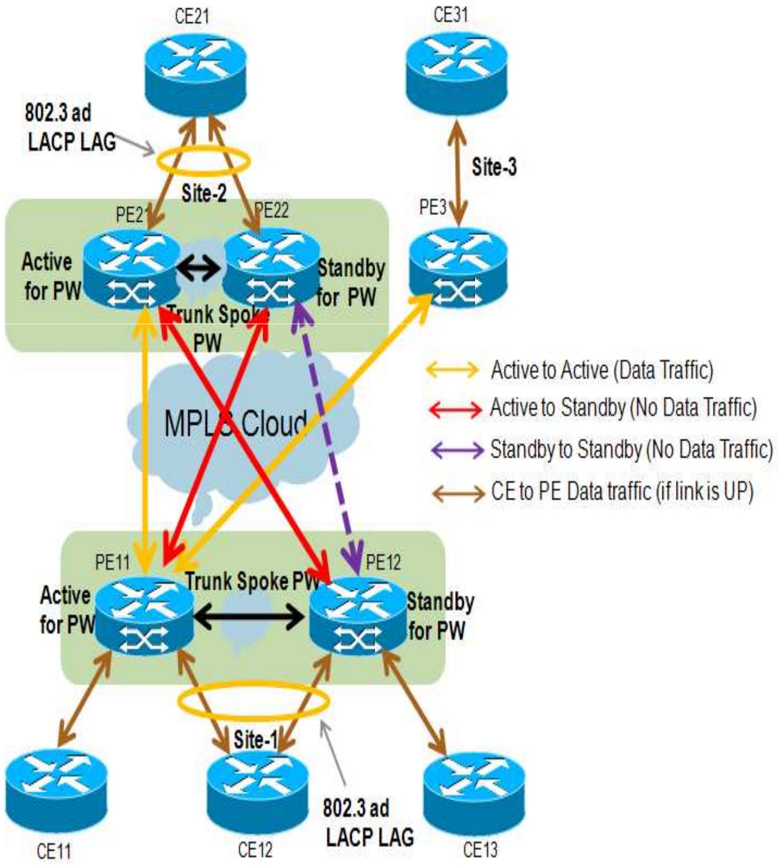
- when one of the Trunk node cannot reach a remote PE then the other one also cannot.

- Both trunk nodes should have connectivity to remote PE's for VPLS to work in all cases.

- Upon CE12 link failure to PE11 traffic will flow via CE12 → PE12 → PE11 → MPLS Cloud → remote end

  No MAC learning needed for CE12

  VPLS Link and Node convergence will be in ms



Take Away

- 802.3-ad extension provides Layer 2 redundancy similar to Layer 3

  To Achieve Fast convergence in layer 2 Ethernet networks

  upon link and Node failures

  host doesn't have to re-learn the Source/Destination MAC

  Easy to implement

  implementation is Transparent to Datacenter servers/host devices

- Extends redundancy of layer 2 datacenter networks over geographically distributed regions

  Can interoperate with Layer 2/Layer 3 networks

  802.3-ad extensions can be extended to VPLS to take advantage of

  built-in MPLS redundancy (Link and Node Protection)

Agenda

- Overview
- Layer 1 Resiliency
- Layer 2 Resiliency
- Layer 3 Resiliency
- Deployment Case Study
- Conclusion

Layer 3 Resiliency – Bi-direction Forwarding Detection Problem Definition

- Methods needed to quickly determine forwarding failure
- Ethernet needs a solution for failure detection
- Layer 3 Data Forwarding plane needs a check and Checking should not be bound to single hop
- Fast Hello needed for LDP, OSPF, ISIS, PIM, RSVP, BGP etc to catch same types of issues.
- BFD is a single Layer 3 protocol for detecting forwarding failures
- Other protocol timers can now be left at defaults

Layer 3 Resiliency – Bi-direction Forwarding Detection Operations

- Routing Protocol (BFD client) bootstraps BFD to create BFD session to a neighbor,

  and to receive link status change notification.

- Receive and Transmit intervals are negotiated and configurable

- Two systems agree on method to detect failure

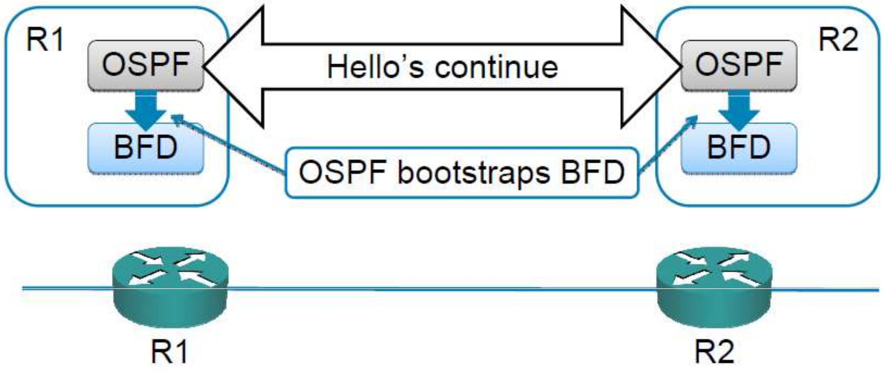
  Via sending packets, watching counters etc

- In case of failure, BFD notifies BFD client

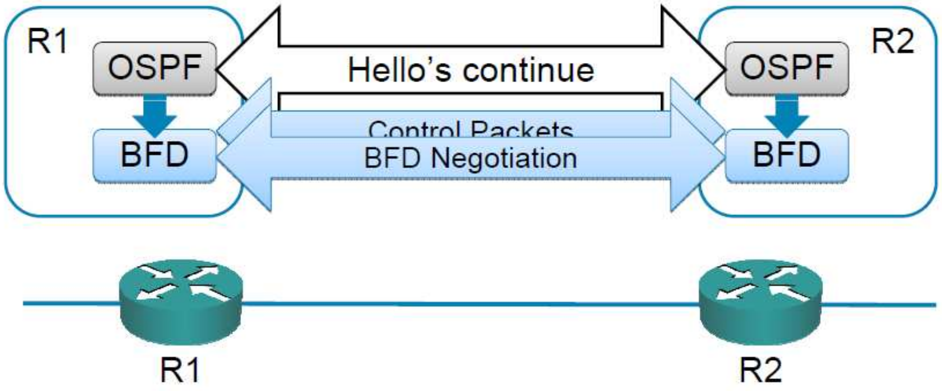
- BFD Client independently decides on action (if any)

Layer 3 Resiliency – Bi-direction Forwarding Detection Operations

- OSPF adjacency comes up
- OSPF bootstraps BFD once session is UP
- BFD establishes session with peer router



- OSPF hello's at slower rate
- BFD control Packet maintain state and verify forwarding plane liveliness

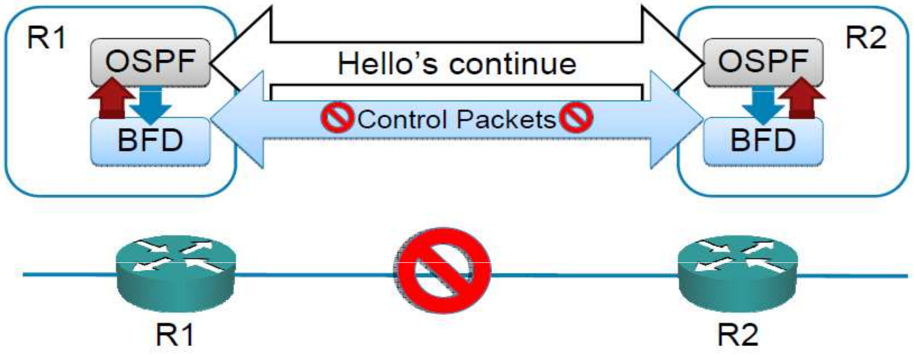


Layer 3 Resiliency – Bi-direction Forwarding Detection Operations

- BFD notifies OSPF of failure

- OSPF declares neighbor dead

- Other protocols (ISIS, BGP) may take more granular actions

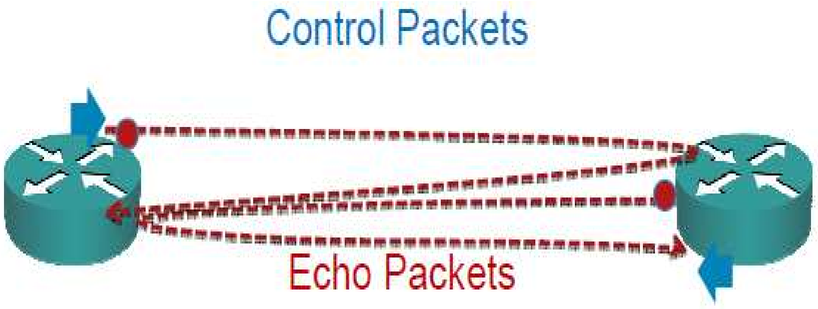


- If echo function is not negotiated

  control packets sent at high rate to achieve Detection Time

- If echo function is negotiated

  control packets sent at a slow rate self directed echo packets sent at high rate (Min Echo Rx Interval)



Layer 3 Resiliency – MPLS Traffic-engg Problem Definition

- Congestion in the network due to changing traffic patterns

  Election news, online trading, major sports events

- Better utilization of available bandwidth

  Route on the non-shortest path

- Route around failed links/nodes

  Fast rerouting around failures, transparently to users

  Like SONET APS (Automatic Protection Switching)

- Build new services—virtual leased line services

  VoIP toll-bypass applications, point-to-point bandwidth guarantees

Layer 3 Resiliency: MPLS TE Application Fast Re-route
MPLS Link and Node Protection

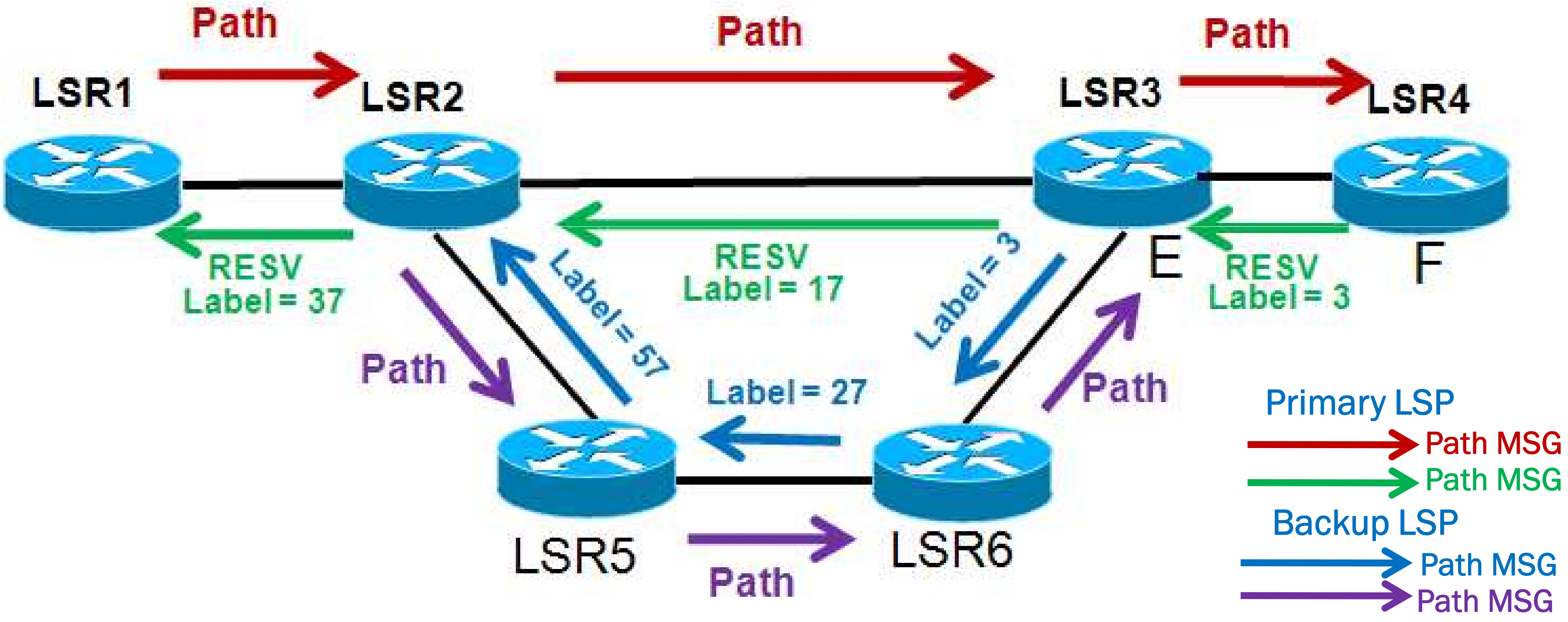
- **Link protection** take out the link being protected and recalculate best shortest path to the next-hop satisfying the constraints

- **Node protection** take out the node being protected and recalculate best shortest paths to termination points (usually next-next-hops) satisfying the constraints

- Types of MPLS FRR

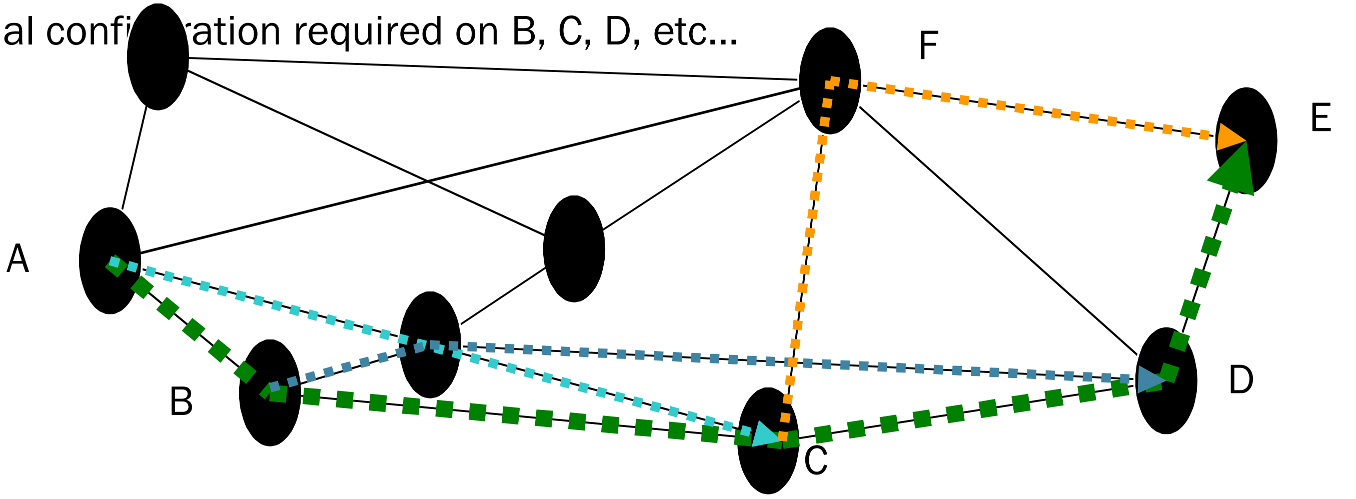
One-to-one Backup	Many-to-one Backup
• Backup each LSP separately. • More flexible • Simple to configure • Detours are setup automatically	• Backup a bunch of LSPs with one LSP • label stacking • Requires configuring bypass LSPs

Layer 3 Resiliency: MPLS Fast Re-route
Protected and Backup Protection LSP Setup



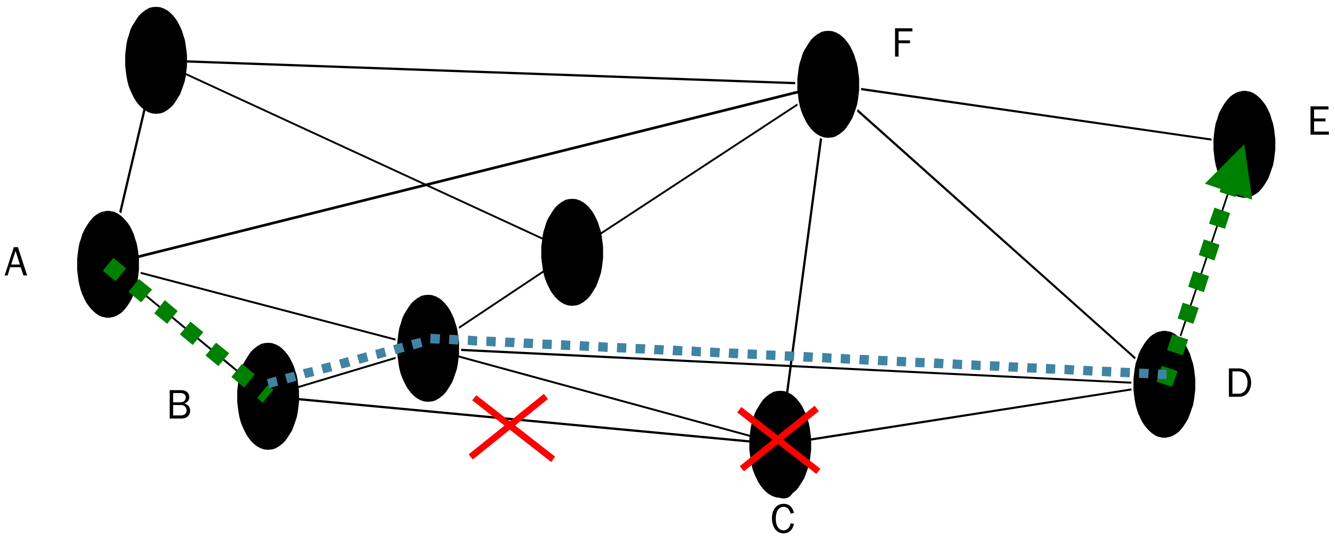
Layer 3 Resiliency: MPLS Fast Re-route One-to-one backup: example

- LSP setup Between Node A and E
- Enable fast reroute on ingress
  - A creates detour around B
  - B creates detour around C
  - C creates detour around D
  - No additional configuration required on B, C, D, etc…



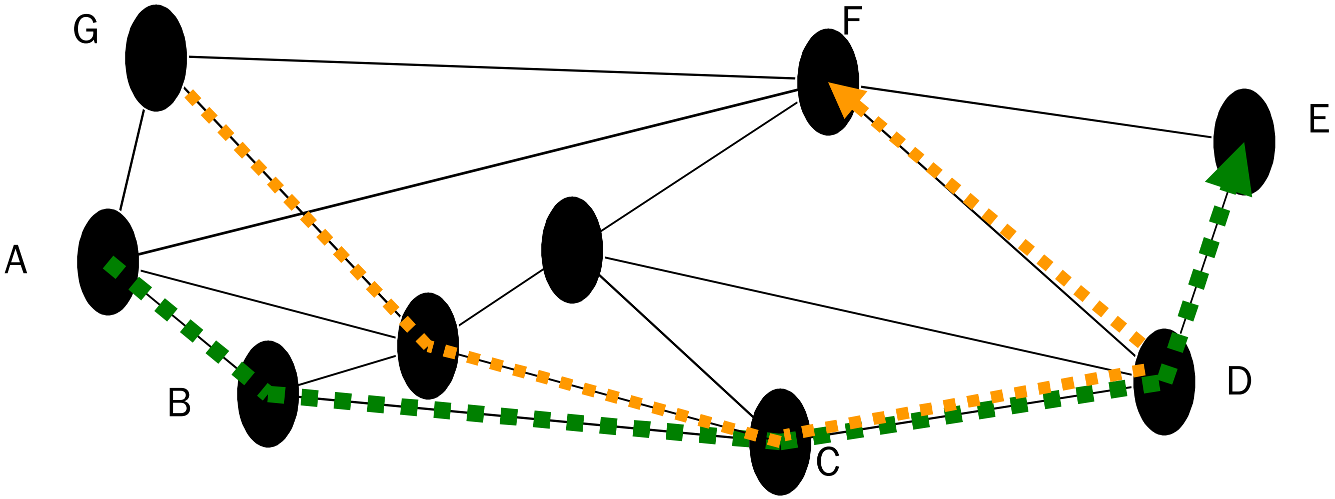
Layer 3 Resiliency: MPLS Fast Re-route One-to-one backup: example

- Node C or/and link B-C fail:
  - B immediately detours around C
  - B signals to A that failure occurred



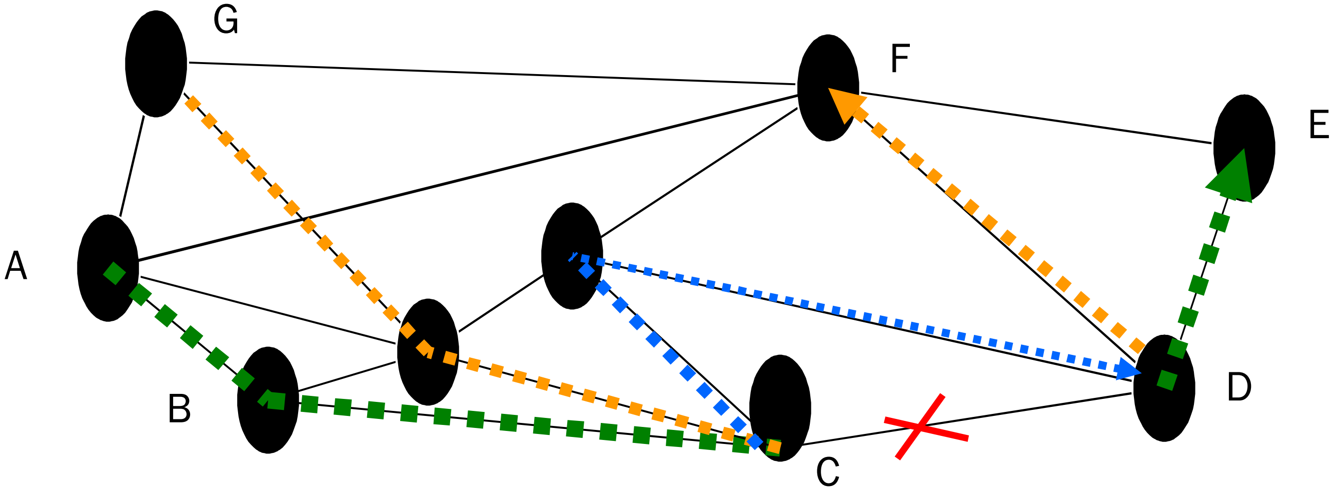
Layer 3 Resiliency: MPLS Fast Re-route One-to-one backup: example

- Two User LSPs(G-F and A-E) going over link C-D.
- Bypass lsp is created on C to avoid C-D(direct link)



Layer 3 Resiliency: MPLS Fast Re-route
Many-to-one backup: example

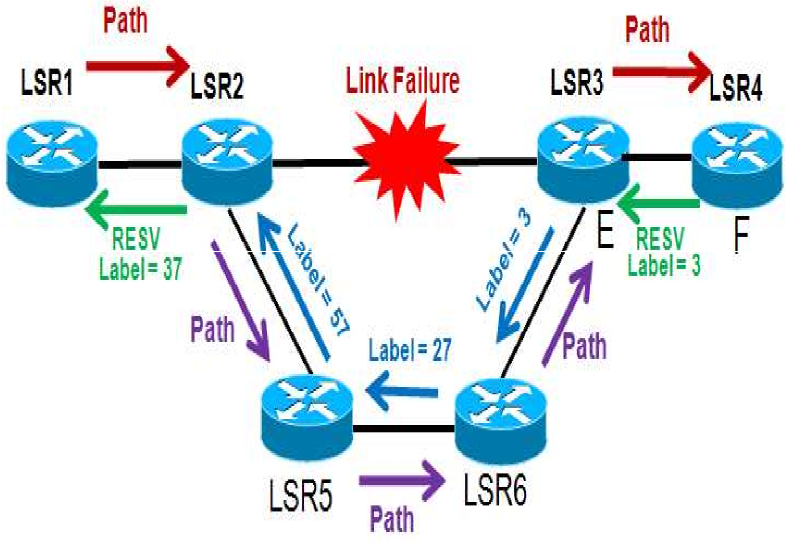
- Link C-D fails
  - C reroutes user traffic with label-stacking (“outer” label + “inner-1” or “inner-2” labels)
  - C signals to A and G that failure occurred



Layer 3 Resiliency: MPLS Fast Re-route
Backup Protection LSP Setup
Link Failover and Label stack walk

- Backup Protection LSP:

To protect link between **LSR2** and **LSR3**
The Backup protection LSP will be **LSR2 → LSR5→LSR3**



Label Stack	LSR1	LSR2	LSR5	LSR6	LSR3	LSR4
	37	57	27	3	3	
		17	17	17		

Take Away

- BFD Provides milli-seconds convergence mechanism for
  - Higher layer Protocol such as OSPF, ISIS , BGP, MPLS and RSVP
  - Detects Control and Data Path failures and informs upper layer protocols

- Fast Re-route provides convergence mechanisms
  - For MPLS applications in case of Link or Node failures
  - Provides one to one and one to many LSP Protection mechanism
  - Easy to configure and Manage
  - Can be used in conjunction with BFD

Questions & Answers