

Path Failure Detection

Fast Rerouting, BFD, Dampening

(C) Herbert Haas 2009/05/02

General Problem Description



- **Layer 2 and layer 3 topologies may be different**
- **Concurrent connectivity detection and protection/re-routing mechanisms**
 - ◆ They are not aware of each other!
- **Even worse: Multiple transport technologies**
 - ◆ That is "multiple Layer 2" in OSI speak
 - ◆ SDH, MPLS, POS, ATM, ...
 - ◆ Oscillation effects possible

(C) Herbert Haas 2009/05/02 www.perihel.at

2

Protection switching



- Originally, the ITU-T requires SONET/SDH rings to recover within 150 ms upon a link/path failure
 - ♦ Ring topology as required redundancy method
- These stringent 150 ms requirement becomes also mandatory for IP networks
- Today Packet over SONET (POS) provides the shortest recovery time
 - ♦ Typically detects and reacts to media or protocol failures within 50 ms
 - ♦ This has become the benchmark against which other protocols are measured

Network convergence



1. Detect the event (=failure)
 - ♦ Failure may happen locally or remote
2. Propagate the event
 - ♦ Communicate the failure event to other devices in the network
3. Process the event
 - ♦ Repair delay
 - ♦ Notified devices must calculate an alternate path
4. Update the routing table/FIB

Failure detection delay



- **Failure detection is the most critical delay component**
 - ♦ Long delays
 - ♦ False negatives, inconsistent detection methods
 - ♦ Ambiguous results possible
- **Routing protocols use varying methods and timers to detect the loss of a routing adjacency with a peer**

Fast Reroute (FRR)



- **Originally an extension to MPLS-TE**
 - ♦ Often the only reason to use MPLS-TE
- **Pre-calculates shortest paths around individual nodes and links**
- **Upon failure, traffic can be quickly switched to the reroute path**
 - ♦ Within 50 ms or faster!
 - ♦ Typically no best-path chosen but enough to allow IGP to calculate a new best route in the meantime

FRR Details



- **Path versus Local protection**
 - ◆ Local protection can be either Link protection or Node protection
- **Local protection**
 - ◆ Allows 1:N relationship between one backup LSP and N primary LSPs

Link failure detection



- **Either via RSVP hello extensions**
 - ◆ Especially needed when PHY does not support keepalives (such as Ethernet)
- **Or 'native' methods, e. g. SONET/SDH mechanisms or other keepalives (BFD)**

Local Protection – Basic Principle



- Intermediate nodes may announce backup paths to the head-end using RSVP (Resv) messages
- In case of a link problem, the affected upstream router changes the LFIB entries: An additional label is inserted to reroute the packets over the backup path
- Post-Failure Signaling is done to inform all nodes about the new situation in the network; this consists of
 - ♦ Upstream signaling (code 25/3 = "tunnel locally repaired")
 - ♦ IGP notification
 - ♦ Downstream signaling - from intermediate 'repair point' tell downstream to keep remaining primary path up

Path protection – Basic Principle



- Provides additional LSP in parallel with an existing LSP
- Faster convergence than local protection
- But every backup node must reserve the same amount of BW as currently used by the primary LSP

FRR Extensions to RSVP-TE (RFC 4090)



- **One-to-one backup method**
 - ◆ Creates detour LSPs for each protected LSP at each potential point of local repair
 - ◆ Might lead to scalability problems
- **Facility backup method**
 - ◆ Creates a bypass tunnel to protect a potential failure point
 - ◆ Can protect a set of LSPs that have similar backup constraints
 - ◆ Utilizes MPLS label stacking

IP Fast Reroute Framework



- **A draft to allow FRR without MPLS**
 - ◆ "...provides a framework for the development of IP fast-reroute mechanisms which provide protection against link or router failure by invoking locally determined repair paths. Unlike MPLS fast-reroute, the mechanisms are applicable to a network employing conventional IP routing and forwarding."
- **Failure detection**
 - ◆ Loss of light
 - ◆ Bidirectional Forwarding Detection (BFD)

Bidirectional Forwarding Detection (BFD)

(C) Herbert Haas 2009/05/02

Link failure detection issues



- Link-layer failure detection times can vary widely *depending on the physical media and the Layer 2 encapsulation used*
 - ♦ Switches can hide link-layer failures from routing protocol peers
- **Various network applications cannot reliably detect bidirectional failures**
 - ♦ And the detection times are often too long (OSPF or EIGRP: down to 1 second only)

(C) Herbert Haas 2009/05/02 www.perihel.at

14

About BFD



- **Current Internet Draft proposed by Juniper and Cisco**

"...to detect faults in the bidirectional path between two forwarding engines, including interfaces, data link(s), and to the extent possible the forwarding engines themselves, with potentially very low latency. It operates independently of media, data protocols, and routing protocols."

draft-ietf-bfd-base-09 (Feb 2009)

About BFD (cont.)



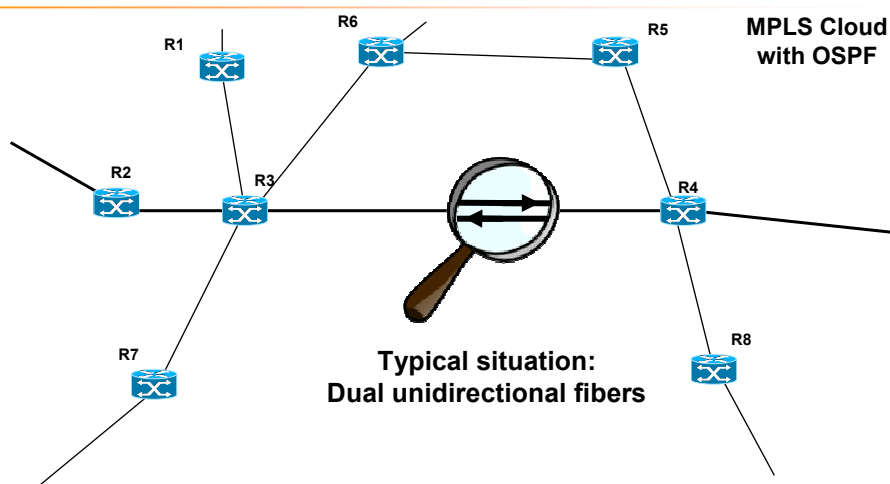
- BFD provides a low-overhead and **fast two-way connectivity verification** between two systems
 - ◆ Separate BDD session for each path and data protocol
- Common BFD applications include
 - ◆ Control plane liveness detection
 - ◆ Tunnel endpoint liveness detection
 - ◆ A trigger mechanism for IP/MPLS Fast ReRoute
 - ◆ MPLS Label Switching Protocol data plane failure detection

MPLS Example – In short



- LDP relies on TCP and detects link failures in either direction
- But OSPF might only detect unidirectional failures
- Now assume one fiber (of a dual-fiber link) breaks
 - ◆ One OSPF peer may still think the adjacency is up and continues to distribute routes for that path
 - ◆ Also distributes labels (via LDP) to other neighbors!
 - ◆ Results in black hole routing!

MPLS Example (1)



MPLS Example (2)

We use OSPF to discover the topology and LDP to announce labels for each network.
Note that OSPF Hellos are unidirectional mechanisms!

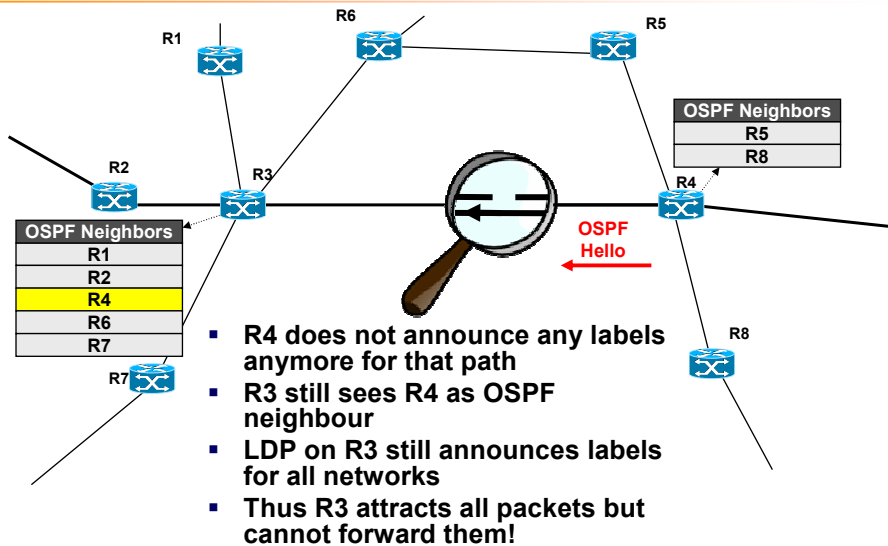
(C) Herbert Haas 2009/05/02 www.perihel.at 19

MPLS Example (3)

Now one fiber breaks
(e. g. connector problems)

(C) Herbert Haas 2009/05/02 www.perihel.at 20

MPLS Example (4)



Session Details

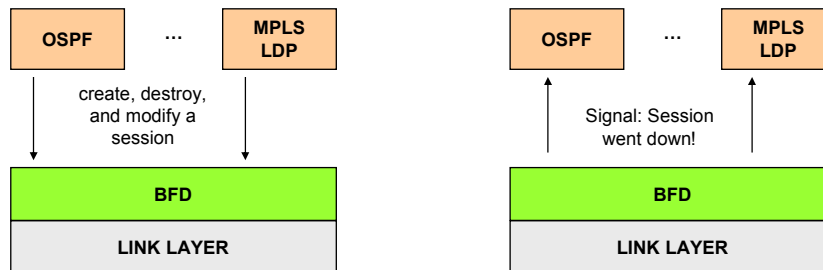


- Multiple BFD sessions can be established between the same pair of systems
 - ◆ When multiple paths between them are present in at least one direction (less return paths possible)
- Always runs in unicast, point-to-point mode
- Always uses three-way handshake for establishment and tear-down
 - ◆ To ensure that both systems are aware of the state change

BFD as service



- BFD can be considered as service for network applications
- Provides
 - ♦ Service primitives to maintain a session to neighbors
 - ♦ Signalling to its clients when session goes up or down
- Peer address must be provided by client
 - ♦ BFD does NOT discover neighbors!



Automatic rate adaptation



- Peers determine max rate of BFD transmissions and receptions
 - ♦ Can be adjusted in real-time
- Thus minimizes fault recognition time
- Also used to avoid link congestion

Operating Modes



- **Asynchronous mode**
 - ♦ Periodic BFD Control packets in both directions
 - ♦ If some packets in a row are missing, the session is declared down
- **Demand mode**
 - ♦ Either in one or both directions
 - ♦ One peer may start/stop the other from sending control packets
 - ♦ Thus resembling "on demand" verification

Echo function



- **Additional optional function for both modes: "Let the peer loop back our own BFD *Echo* packets"**
- **Reduces BFD *Control* packet rates**
 - ♦ Even no control packets required in Demand mode
- **Can be enabled in one or both directions**
 - ♦ "Loop-back" peer must signal to support the echo function

Note



- **Plain asynchronous mode**
 - ♦ Has best detection/overhead ratio
- **Demand mode**
 - ♦ Reduces the overhead, e. g. when the number of BFD sessions is large
 - ♦ Detection time depends on heuristics of the client
 - ♦ Should only be used when the RTT is small
- **Consider firewalls that may block BFD**

BFD Session Initiation



- **Peers have either active or passive role**
 - ♦ At least one peer must be in active role
- **Active peer starts sending BFD Control packets**
 - ♦ First starting at slow rate
 - ♦ When bidirectional communication is achieved, the BFD session comes up
- **If the Echo function is NOT active, the transmission rate of Control packets may be increased**
 - ♦ To a level necessary to achieve the detection time requirements for the session

Protocol Details



- **Small messages: 24 bytes**
- **BFD messages can be carried on arbitrary layers**
 - ◆ "Encapsulation out of scope...and should be appropriate to the environment"
- **Message authentication optional**
 - ◆ Currently keyed-MD5 or keyed-SHA1 defined (plus plain password)

Timer negotiation



- **Each peer reports the desired TX and RX rate**
 - ◆ The peers agree on the slower rate
- **A jitter of 25% should be *subtracted* to the actual TX interval**
 - ◆ To avoid self-synchronization phenomena
 - ◆ That is, on average, the actual interval is 12.5% less than negotiated

Caveats



- **BFD might cause false positives**
 - ◆ Upon packet corruption, queue congestion
- **CPU consumption may be significant on certain implementation**
 - ◆ Cisco claims <2% upon 100 BFD sessions
- **Non-stop Forwarding (NSF) switchover to standby route processor may cause a few lost packets**
 - ◆ **BFD-session may break**
 - => OSPF is (wrongly) notified about broken link
 - => NSF fails

BFD is NOT appropriate...



- **In Dual-ring SONET/SDH Automatic Protection Switching (APS)**
 - ◆ APS already supports ~ 50ms switchover

Cisco's (current) implementation



- **Uses UDP/IP**
 - ◆ 20+8+24=52 bytes total
 - ◆ Destination port 3784
 - ◆ Source port within [49152..65535]
- **Highest rate: 50 ms interval**
- **Supported for OSPF, IS-IS, EIGRP, and BGP clients**
- **Outgoing BFD packets enjoy the highest possible queue priority**
 - ◆ BFD packets bypass the QoS subsystem
 - ◆ Therefore QoS features **CANNOT** be applied to outgoing BFD packets

Other Cisco-related issues



- **Some protocols like HSRP and Multicast are not currently BFD-enabled**
 - ◆ But benefit indirectly: BFD notifies IGP and HSRP or Multicast relies on IGP
- **Good BFD implementation in high-end platforms (GSR, CSR, ...)**
- **ISR platforms should use IOS release 12.4(15)T or newer**
 - ◆ Only supported on Ethernet interfaces!
- **BFD does not work for multihop BGP neighbors**

Configuration



```
! Enable BFD on an interface: Specify sending and receiving rates, as well  
! as the max number of packets (in a row) that can be lost before a failure  
! is assumed.  
(config-if)# bfd interval <50-999> min_rx <1-999> multiplier <3-50>  
  
! Finally, select a "BFD-client"  
(config-router)# bfd all-interfaces | interface <if-id>
```

Typical BFD echo-function configuration:

```
! Specify a global interval when control packets should be send:  
(config)# bfd slow-timers 10000  
(config-if)# bfd interval 50 min_rx 100 multiplier 3  
(config-if)# bfd echo
```

Interface Dampening

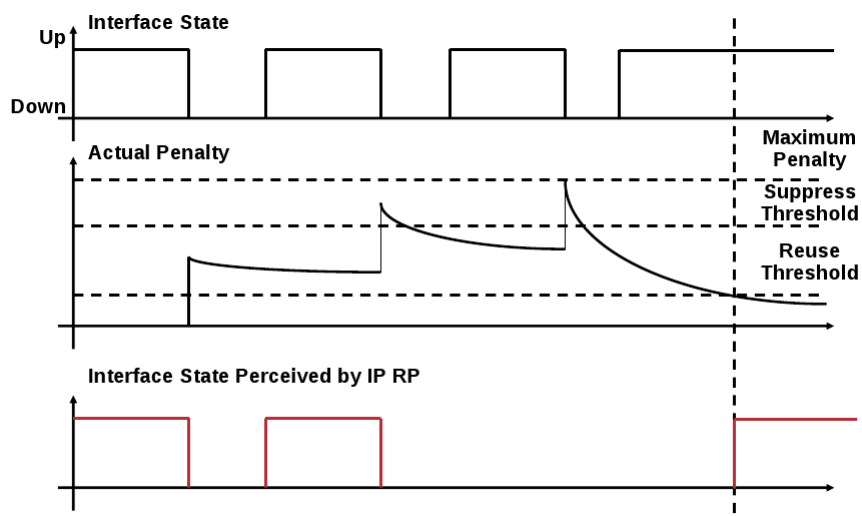
Interface dampening



- Flapping interfaces cause lots of
 - ◆ Routing table processing
 - ◆ Routing protocol updates
- "Dampen" the interface

```
(config-if) # dampening [<half-life-period> <reuse-threshold>]  
                    [<suppress-threshold> <max-suppress-time>]  
                    [<restart-penalty>]
```

Interface dampening - example



Note!



- **Fast failure detection and reaction may cause routing instabilities**
- **Even short congestion periods may result in peers which declare each other dead**
- **May be abused by attackers**