BFD

(Bidirectional Forwarding Detection)

Does it work and is it worth it?

Tom Scholl, AT&T Labs

NANOG 45

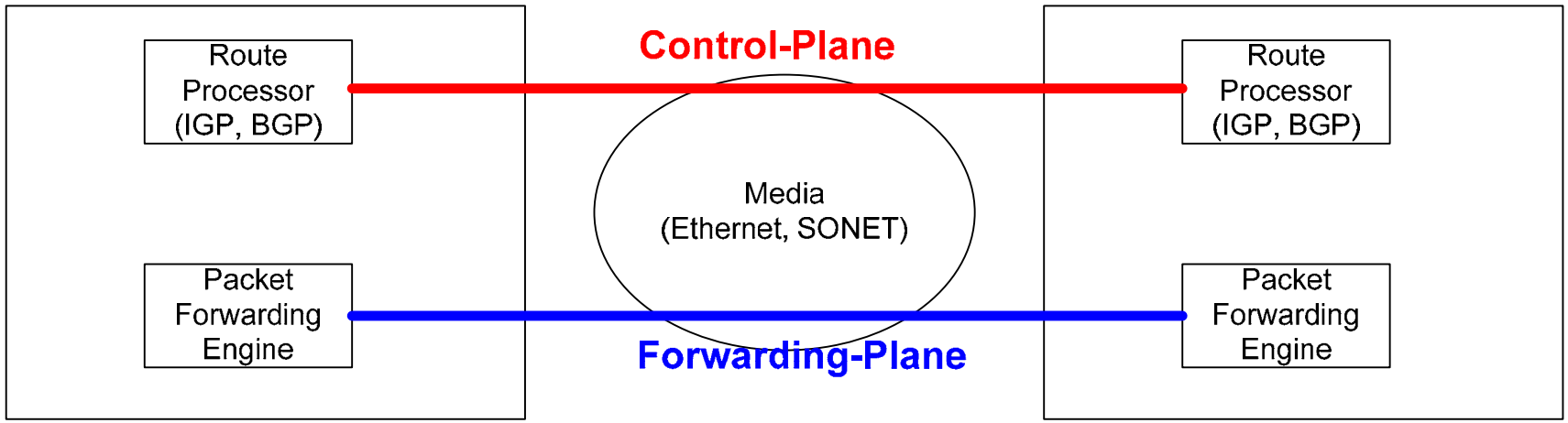
What is BFD?

- BFD provides a method to validate the operation of the *forwarding plane* between two routers.
- Upon detecting a failure, triggers an action in a routing protocol (severing a session or adjacency).
- Operates in two modes:
  - Asynchronous
  - Demand
- In either mode, BFD provides an Echo function in which one side can request its neighbor to loop back a series of packets.

Why would an operator use this?

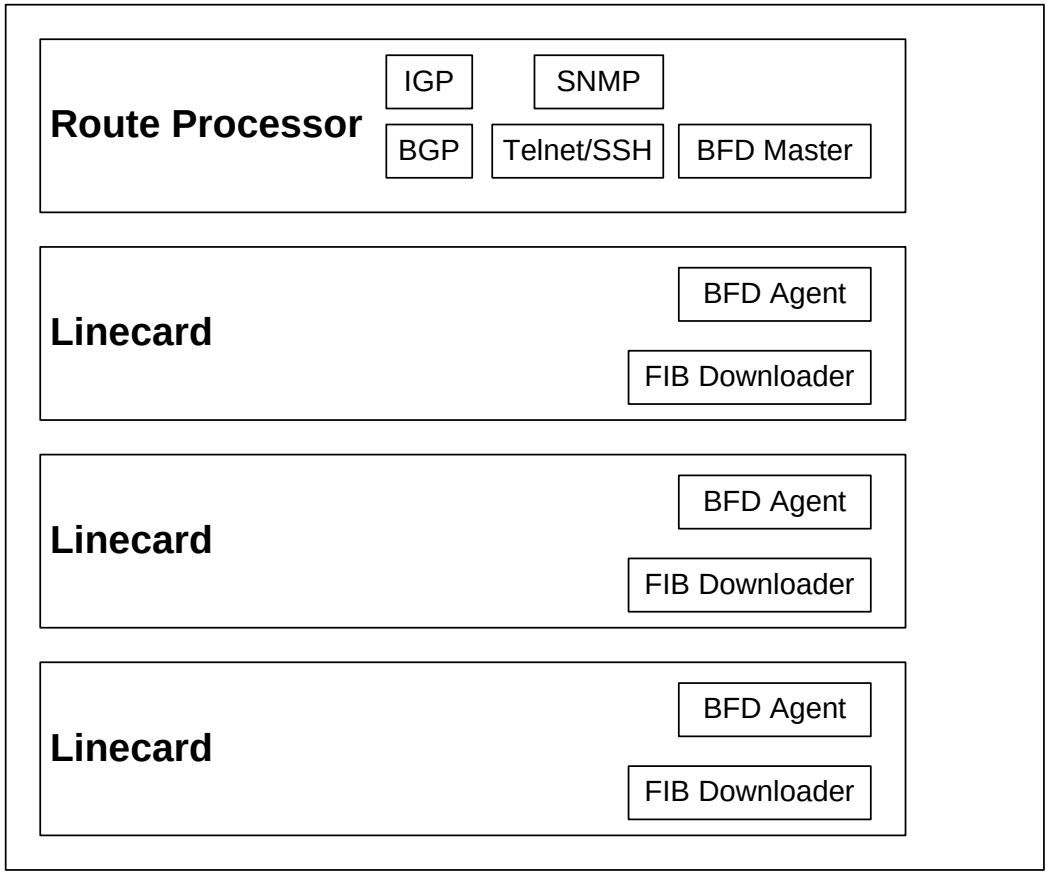
- BFD can rapidly propagate awareness of *forwarding plane* failures up to routing/signaling protocols.
- Relying solely upon hellos, KEEPALIVEs, etc. to validate forwarding behaviors can be a *bad* idea.
  - Routing/signaling protocols tend to be treated differently than forwarded traffic.
  - Most routing/signaling protocol implementations are not designed to operate with sub-second keepalive intervals.
- Often, BFD runs on the line card, not the route processor, so it is unaffected by RP CPU utilization.

Understanding the layers



Router architectures and BFD

- An example of BFD in a distributed router architecture



What protocols does BFD work with?

- Static routes
- IGPs (OSPF, IS-IS)
- BGP (eBGP, iBGP)
- LDP
- RSVP

Static Routes

- Static routes only use next-hop reachability information to determine whether they are valid.
- BFD provides a nice alternative to validate the forwarding path and provide liveliness detection for the actual next-hop.

IGP

- Some mechanisms exist within the IGP to determine a failure rapidly (even at sub-second intervals).
  - These capabilities (“fast hellos”) only work by verifying the IGP keepalive mechanisms.
  - IGP protocols generally are punted to the route-processor in a distributed system, often bypassing standard packet forwarding.
  - Because IGPs generally run on the route-processor, heavy CPU usage can cause IGP adjacencies to fail.
- BFD can help by severing an IGP adjacency in the event of forwarding path failure.

BGP

- Like IGPs, BGP has its own keepalive mechanism.
  - BGP tears down a session when it has not received a KEEPALIVE message from its neighbor before the hold timer expires.
  - BGP is generally executed on route-processors, just like an IGP, so high RP CPU utilization can also cause BGP failure.
- BFD can shutdown the BGP session in under a second after a forwarding path failure.

iBGP

- BFD can be enabled on an iBGP session between router loopbacks to verify forwarding path.
- Can be an alternative to reliance upon the IGP to notify you of a router going offline.
  - No longer need to rely upon event-driven or periodic next-hop scanning.
  - Can improve iBGP convergence by rapidly detecting BGP neighbor failure.

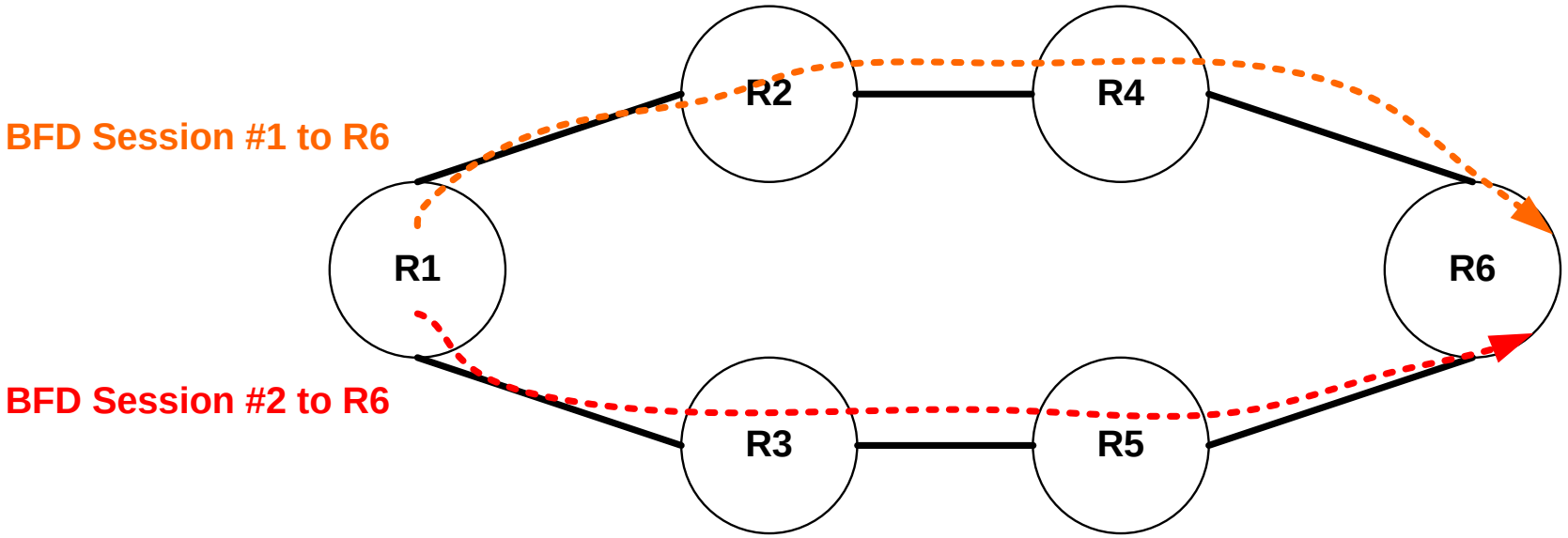
eBGP

- BGP timers aren't great for fast failure detection.
- BFD is great for situations where:
  - You and your neighbor have an L2 device in the middle. (like Internet Exchange LANs or MPLS transport)
  - Transport between neighbors lacks reliable link state notification. (wavelengths)
- BFD allows for ranges to be specified for minimum detection thresholds.
  - Neighbors may have various timers due to their own limitations or preferences.
  - Timers are continuously negotiated and can be altered at any time.

MPLS LDP

- BFD can be enabled to provide OAM on a particular LDP FEC.
  - The LSP is bootstrapped with LSP-Ping and BFD can be operated at a variety of intervals.
- This is useful for informational purposes as LDP really doesn't have a mechanism to select an alternate path (it sticks with what the IGP tells it).
- One benefit is the ability for LDP to "fork" across ECMP paths in a network, providing validation across the ECMP tree.

MPLS LDP and ECMPs (cont'd)



MPLS RSVP

- BFD can be used with RSVP to provide liveliness detection on a path built by RSVP-TE.
- Upon BFD declaring a failure on a particular RSVP-TE path, the head-end router (the router initiating the BFD session) can trigger the use of secondary paths.
- This provides an operator with a nice method to verify multiple forwarding paths as well as provide an automated method to select an alternate path.

MPLS RSVP – Point-to-Multipoint LSPs

- BFD can be used to operate within the Point-to-Multipoint environment to support BFD for each downstream router PE.
- P2MP LSPs are very popular for providing linear broadcast of media, typically with the requirement of rapid-convergence (FRR), bandwidth-reservation and explicit routing (SRLG-free paths).

Pseudowires

- BFD can be used with a pseudowires VCCV (Virtual Circuit Connectivity Verification) control channel.
- This provides a rapid method to detect faults between the endpoints of a pseudowire.
- The fault information could then be translated to other protocols native OAM capabilities (ATM, FR, Ethernet).

What are the caveats?

- Two main ones:

1. BFD can have high resource demands depending on your scale.

2. BFD is not visible to Layer 2 bundling protocols. (Ethernet LAGs or POS bundles)

BFD Resource Demands

- The number of BFD sessions on each linecard or router can impact how well BFD scales for you.
- Each unique platform has its own limits.
  - Bundled interfaces supporting min tx/rx of 250ms or 2 seconds have been seen.
  - In some cases, BFD instances on a router may need to be operated on the route-processor depending on the implementation (non-adjacency based BFD sessions).
- Test your platform first before deploying BFD. Attempt to put load on the RP or LC CPU with your configured settings. This can be done by:
  - Executing CPU-heavy commands
  - Flooding packets to TTL expire on the destination

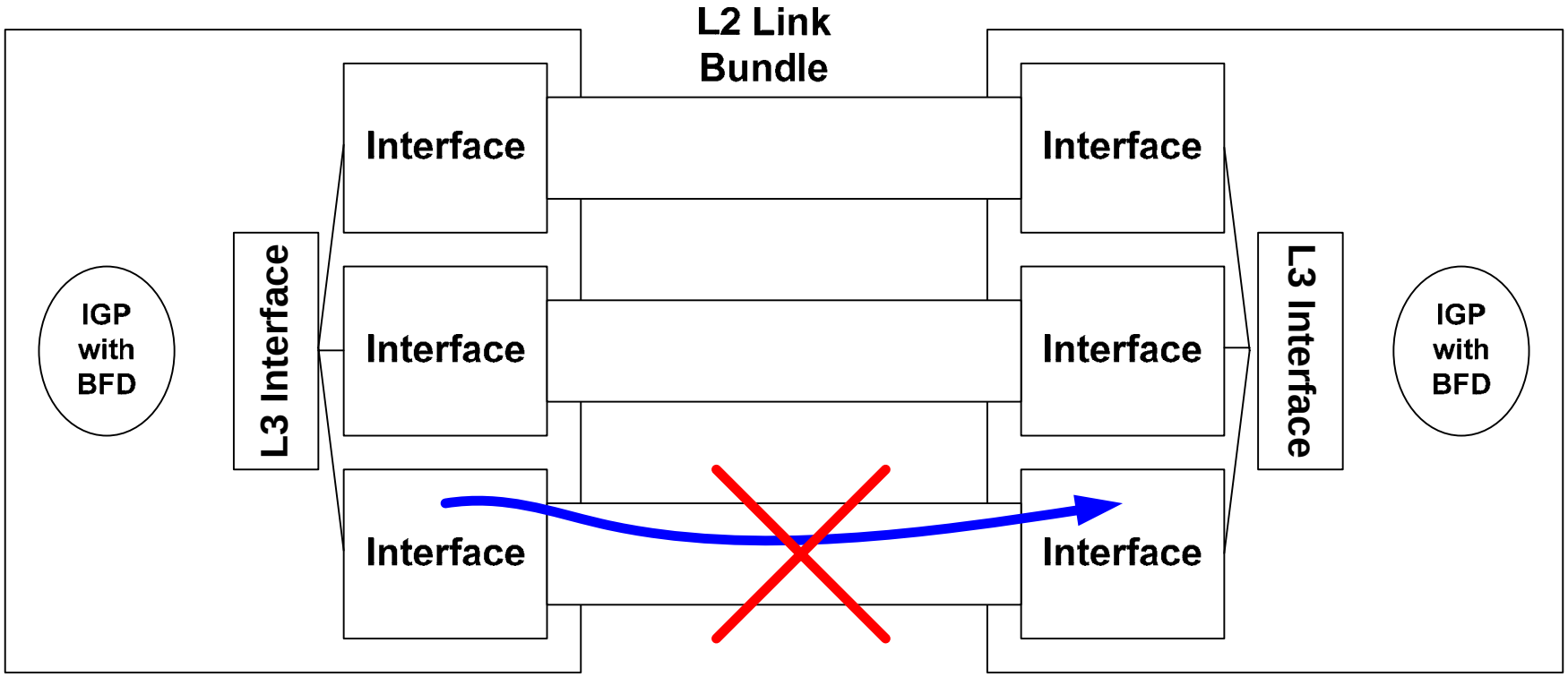
BFD Resource Demands (cont'd)

- What values are safe to try?
- Based upon speaking to several operators, 300ms with a multiplier of 3 (900ms detection) appears to be a safe value that works on most equipment fairly well.
- This is a significant improvement over some of the alternatives.

BFD and L2 link-bundling

- BFD is unaware of underlying L2 link bundle members.
- A 4x10GigE L2 bundle (802.3ad) would appear as a single L3 adjacency. BFD packets would be transmitted on a single member link, rather than out all 4 links.
- A failure of the link with BFD on it would result in the entire L3 adjacency failing.
  - However, in some scenarios the failed member link may result in only a single BFD packet being dropped. Subsequent packets may route over working member links.

BFD and L2 link-bundling (cont'd)



BFD and L2 link-bundling (cont'd)

- This can be a showstopper because it's generally preferable to build L2 bundles rather than to use L3 ECMP, to avoid platform-specific scaling issues and polluting your IGP.

- Having BFD "fork" across each individual link would be great, but it would have its own scaling impact. Each individual member link would have to have a separate BFD session. No vendor currently supports this mode of operation, nor is there a published draft describing it.

Conclusion

- Routers do still have faults in the forwarding plane where IGP and other control-plane protocols continue to work.
  - These events do happen and result in major outages (you've seen some in the press in 2008…)
- The default hello/keepalive intervals of some protocols (BGP, IGP, RSVP) are still too high to be optimal for failure detection.
- There needs to be a way to support L2 link bundling as networks continue to grow links (we don't have 100GE yet, so scaling Nx10G and Nx40G is going to be important).
- Always remember to stress-test your configurations to make sure that you and your equipment is comfortable with what you've selected.

Send questions, comments, complaints to:

Tom Scholl, AT&T Labs

tom.scholl@att.com