FRR for IP and LDP based on Fast Notification

draft-csaszar-ipfrr-fn-02 IETF82, Taipei

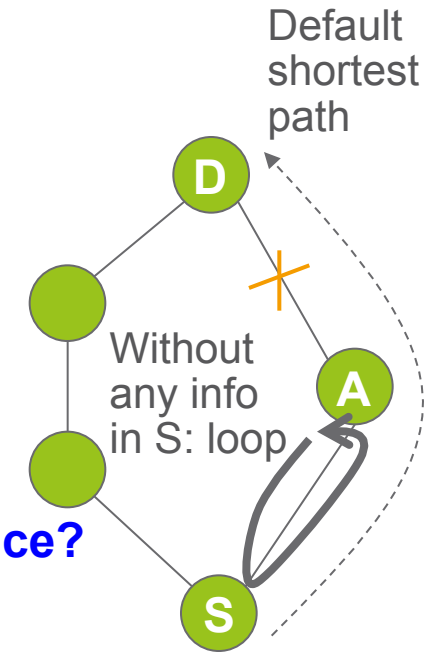
András Császár	Andras.Csaszar@ericsson.com
Gábor Enyedi	Gabor.Sandor.Enyedi@ericsson.com
Jeff Tantsura	Jeff.Tantsura@ericsson.com
Sriganesh Kini	Sriganesh.Kini@ericsson.com
John Sucec	sucecj@telcordia.com
Subir Das	sdas2@telcordia.com

Background

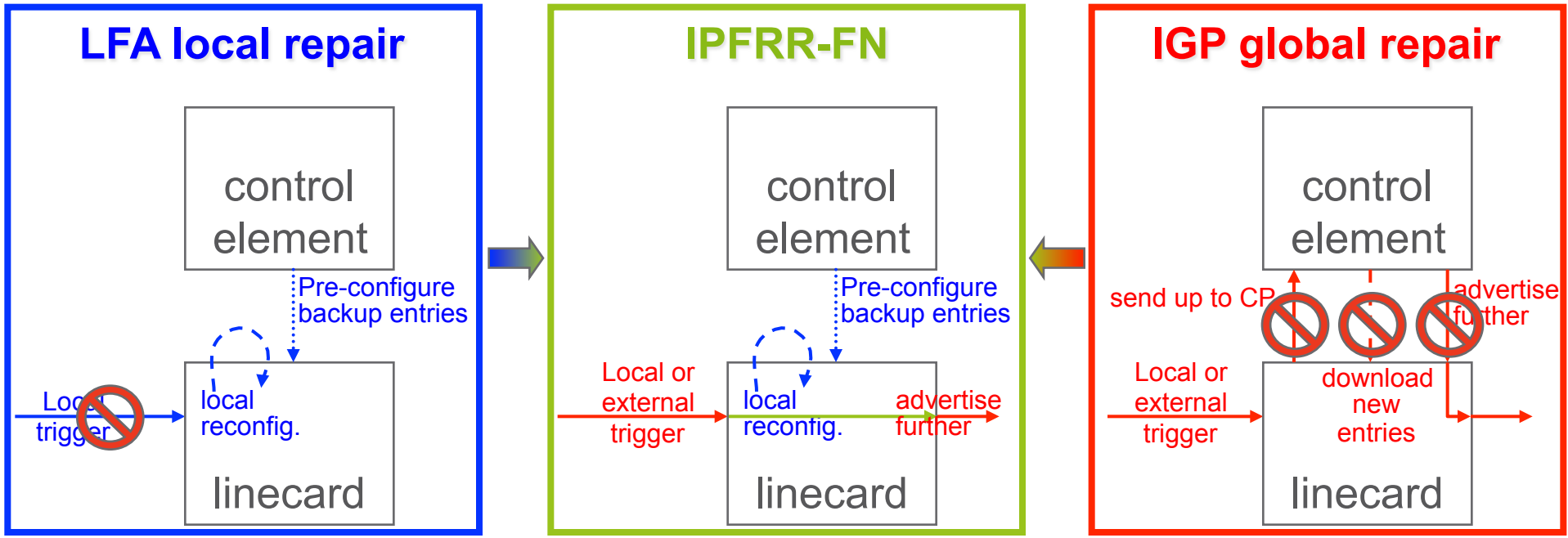
- IP & LDP based on hop-by-hop forwarding
  - Consistency between hops ensured by IGP
- A failure creates inconsistency
  - Wait for IGP global reconvergence (slow)
  - Temporarily use faster means to notify changed routing configuration
    - Encode information into data packet (bits, encaps, label change)
    - Encode info into packet direction (interface specific forwarding)
    - Explicit notification not allowed due to fears of slow performance

Illusion of local repair!

Any difference?



FN ≠ IGP Link State Advertisement



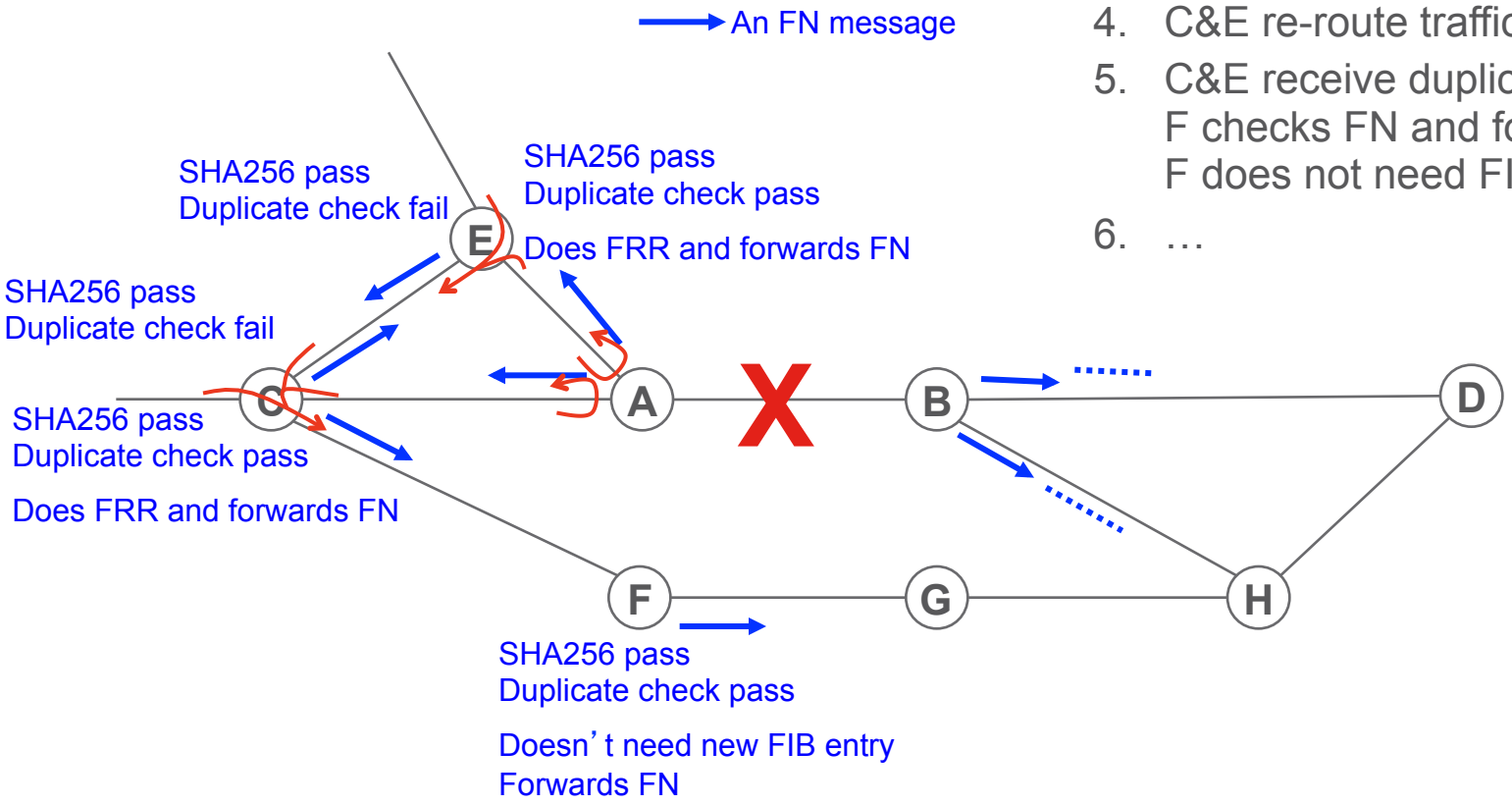
IPFRR-FN Principles

- NOT modifying the IGP/LDP
  - Only using its LSDB
- Pre-computation
  - Let the IGP prepare for each potential (single) failure case
- Pre-installation of backup routes
  - Which deviate from primary routes
- Explicit failure notification in data plane
  - Flooding with duplicate filtering and SHA256 auth check
- IGP after global reconvergence only “confirms” routes
  - Reducing micro-loops (FRR detour identical to final IGP path)

Basic Fail-Over Mechanism with IPFRR-FN

Default path: C-A-B-D

LFA could not handle the failure of A-B link



1. A floods FN (B, too)
2. A reroutes traffic
3. C&E check FN and forward it
4. C&E re-route traffic
5. C&E receive duplicate FN, drop
   F checks FN and forwards it
   F does not need FIB update
6. …

Concerns – The Devil in the Details

- Pre-calculation performance?
- Backup database size?
- Performance of FIB update from the backup database?
- Time to originate an FN packet?
- Time to forward an FN packet?
  - Including duplicate and SHA256 authentication check
- Time to process an FN packet?
- Packet flow disruption time?

Pre-Calculation & Pre-Install

- Non-optimised implementation: ca. 1 SPF for each failure
- A decent implementation should use incremental SPF for each new pre-calculated failure
  - Drastic decrease of overhead
- Only need to pre-install **relevant** cases:
  - For failures downstream on the shortest path(s) towards the destination
    - Only those failures, which result in next-hop change!

Backup Database and FIB Update
An *Extreme* Case

- 1000 nodes
- 20 hop diameter
  - Worst case: *every path* is 20 hops long and *each* link/ node failure results in a *new* alternative next-hop
- 9000 external prefix groups
  - Prefix group = Set of prefixes with the same primary and secondary border routers
    - 9000 prefix groups correspond to *95 BRs*, with each combination serving at least a prefix (95*94≈9000)
- When storing in a very simple structure and assuming a failure impacts *each* route: FIB update can be solved with 50k memory transactions
  - Assuming DRAM with 5MT/sec, and 1 memory controller: 10ms

Underestimate

3.4MB

Comparison: linecards equipped with 1+GB DRAM

FN Packet Performance in Research Prototype

› Prototype: Ericsson SmartEdge with PPA2-based linecards
  – ca. 5-6 years old line card

Linecard requirements:

- Support packet origination locally
- Support packet recognition locally

} Available if card can do BFD or ICMP Echo locally

- Support FIB update locally

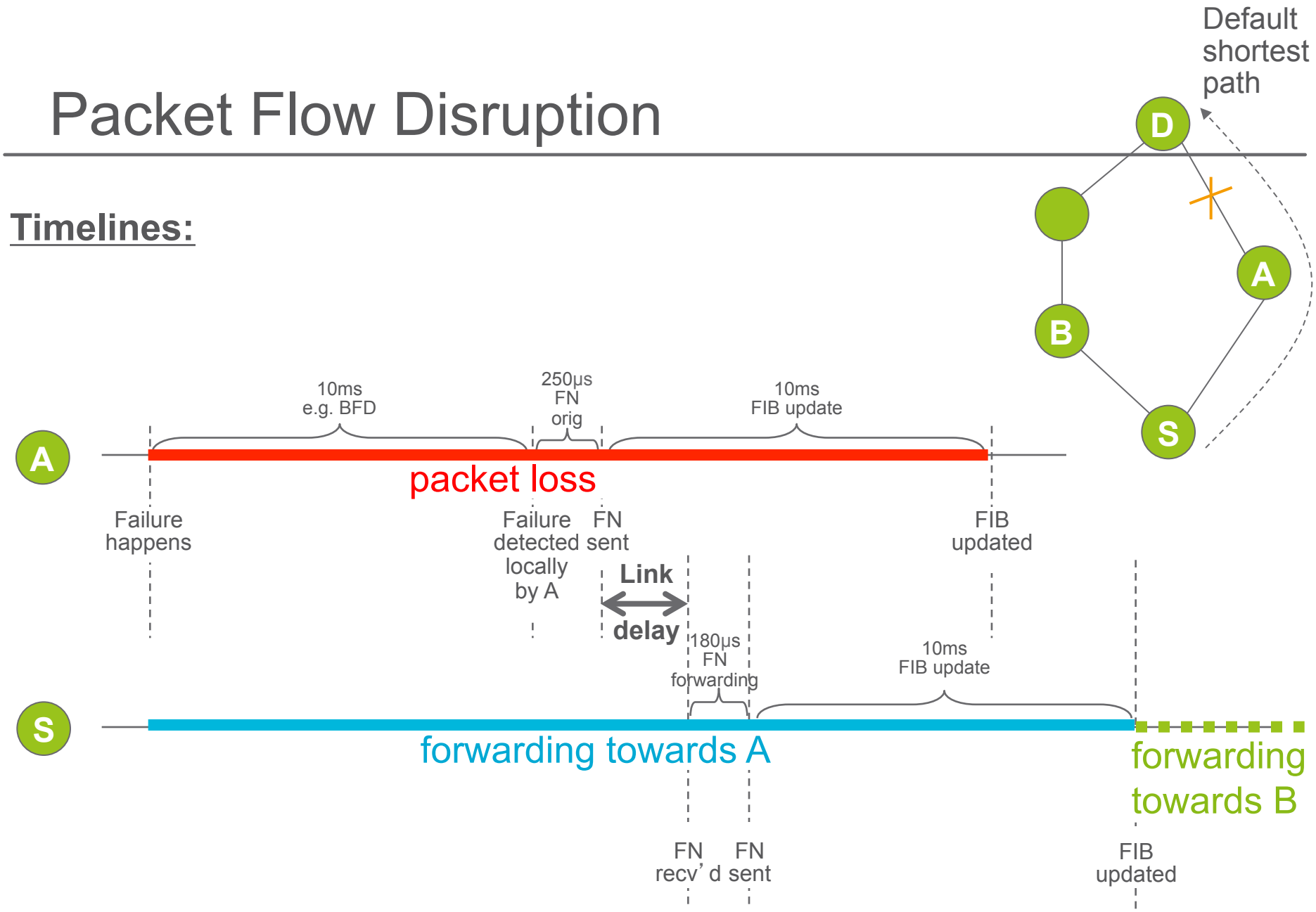
} Available if card can do LFA locally

› FN packet origination < 250µs after failure detection

› FN packet forwarding per hop < 180µs
  – including SHA256 verification and duplicate check in each hop!

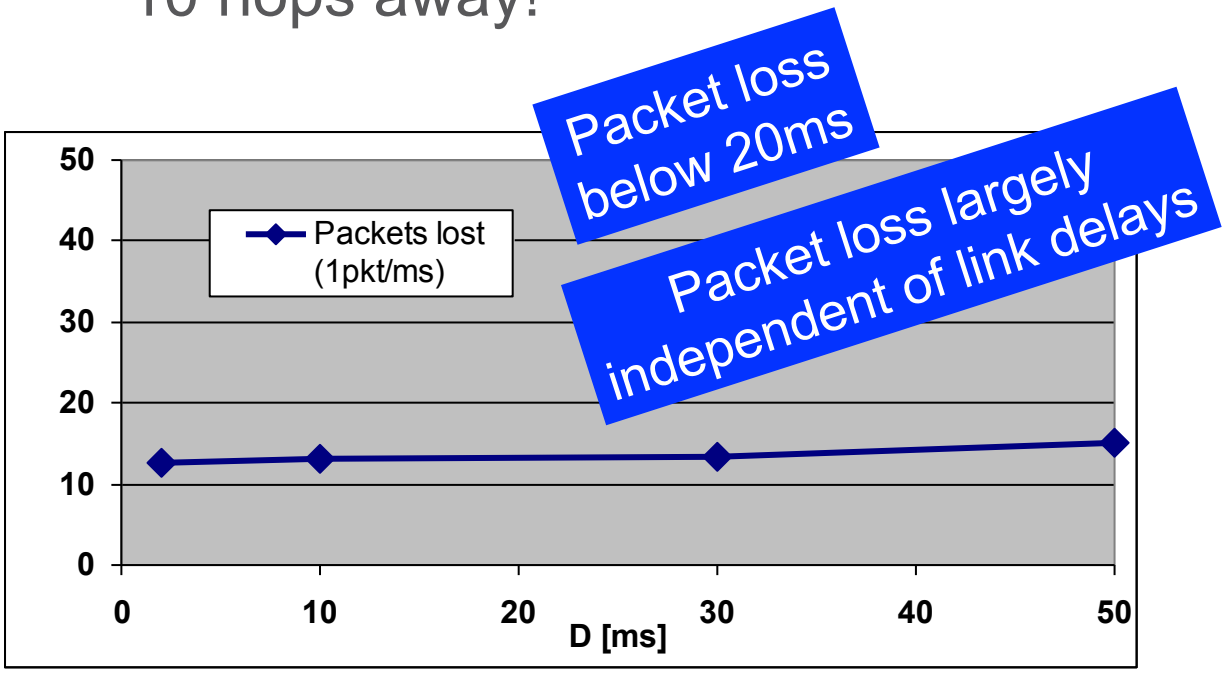
Packet Flow Disruption

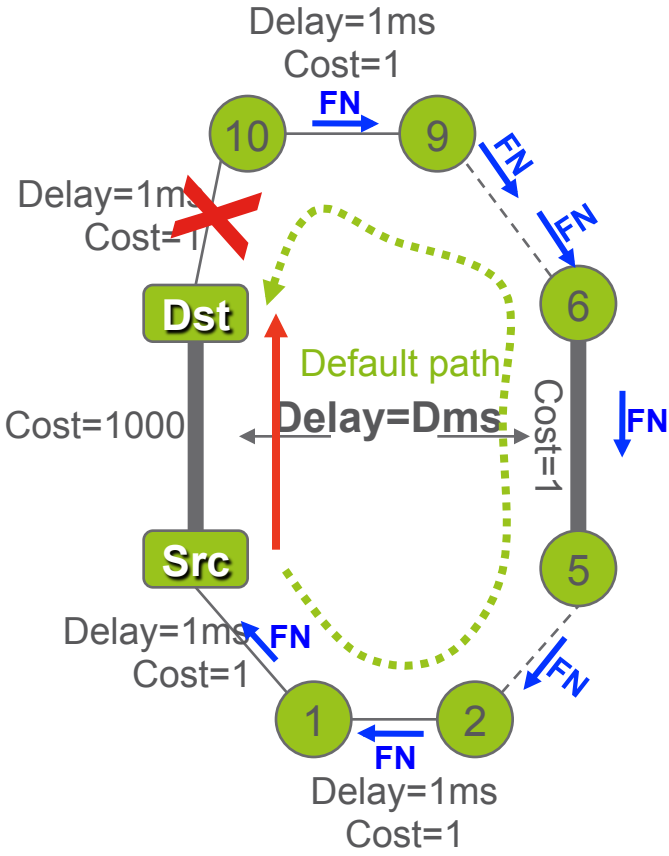
Timelines:



E2E Packet Flow Impact

- Traffic flow: 1 pkt / ms
- Varying delay of "bold" links (D)
- FN results in re-routing 10 hops away!

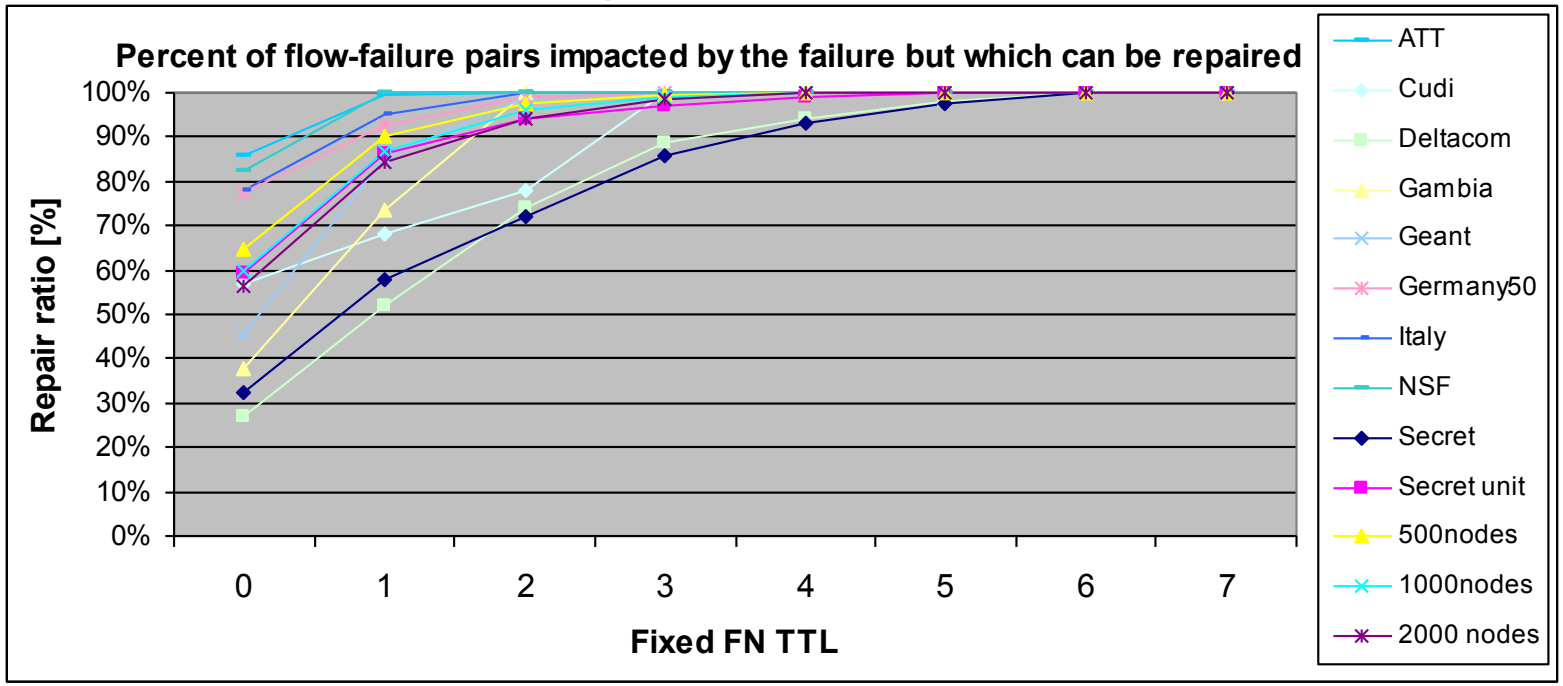




Constraining FN Scope

Any ideas/collaboration is welcome!

A. Static pre-configuration the TTL for FN messages in routers based on best current practices and related studies of available ISP and enterprise network topologies

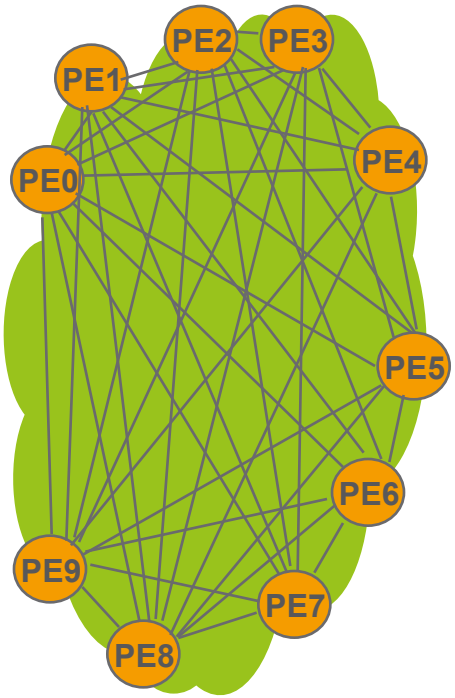


B. Dynamically pre-calculate the TTL value

C. Dynamically pre-calculate the set of neighbours for which a particular FN message should be forwarded

Application to Provider Provisioned VPNs

- Providing FRR for egress PE failure
  - Existing approach: PEs running multi-hop BFD between in each other in (full) mesh
  - E.g. 100 PEs, could be 10k multi-hop BFD sessions (each transmitting BFD packets every, say, 10ms), continuously, all time!
  - Ingress PE router changes egress PE to alternative egress PE
- PW-redundancy: new egress PE needs to activate standby PW with LDP, too
- Why not let the network inform the PEs quickly that a failure happened?
  - FN can distinguish link and node failures!
  - Both ingress and new egress PE receive FN, can modify their routes/ PWs upon primary PE node failure

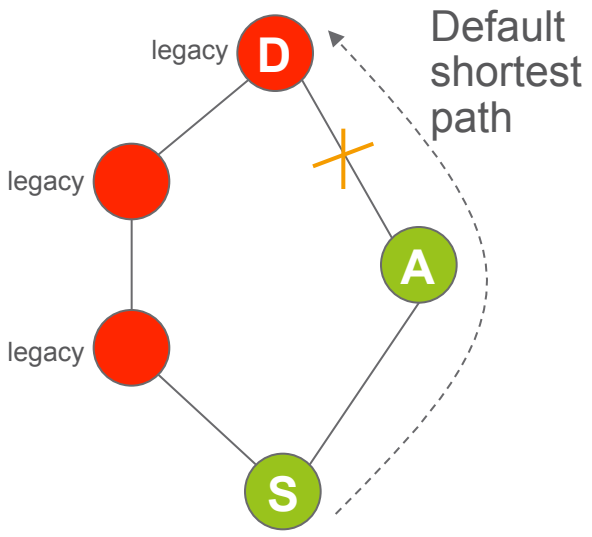


If I haven't been shot down (yet)

And still have time

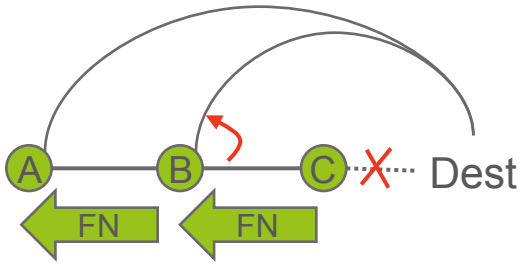
Incremental Deployment

- The more router support IPFRR-FN, the better
- But even two routers can make wonder
- Advertisement of FN capability
  - E.g. Router Capability TLVs
    - OSPF [RFC4970]
    - IS-IS [RFC4971]
- Let's take the example on the first slide that LFA could not solve
  - Even if only A and S support FN, they can start solving failure cases left by LFA
- Remember: TTL=1 or 2 can already greatly improve coverage! (slide 12)

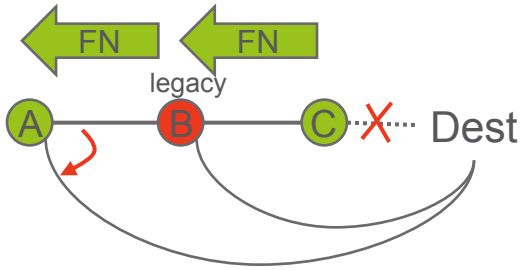


Incremental Deployment – Few Legacy Nodes
Legacy Node Bypass

- Legacy
  - It can at least forward the multicast packets of FN (static conf)
  - FN packets are not recognised/processed → routes are not changed!
- FN-capable nodes
  - When pre-calculating backups, have to consider that legacy nodes won't change routes



- Example:
  - If B is FN capable: it will re-route
  - If B is legacy: C can re-route



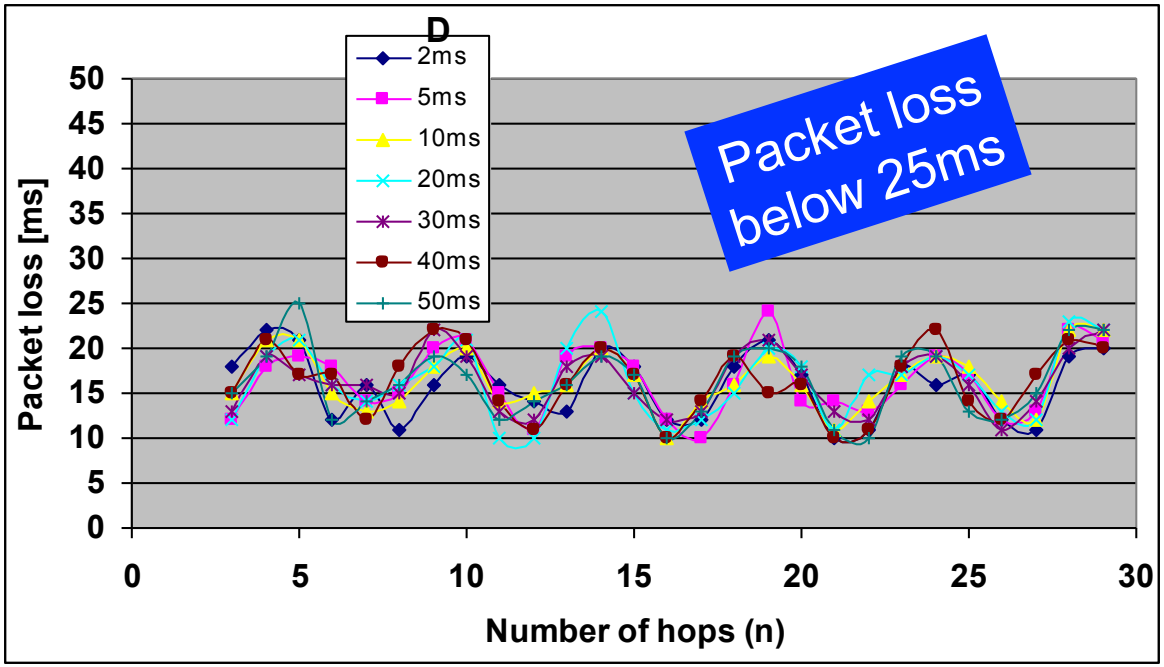
Conclusions

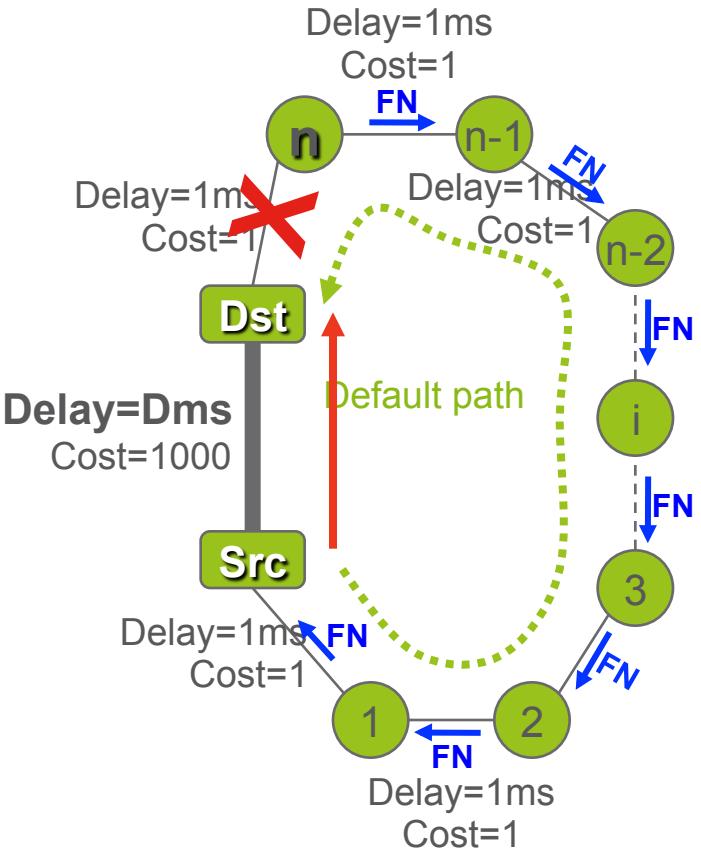
- Fast Notification based IPFRR
  - is feasible
  - has good performance
  - uses the same paths as detours that the IGP will use after global re-convergence (reducing micro-looping)
  - Complete coverage for
    - all single link,
    - all single node and
    - all single SRLG (local and remote) failures and for
    - a reasonable number of pre-configured multiple failure cases deemed important by the operator
  - Does not require total network upgrade to show benefits
  - SIMPLE TO GRASP: just let the routing engine pre-do what it would anyway do after the failure!
- Applicable to
  - IP
  - LDP-MPLS: liberal label retention + downstream unsolicited mode
  - L2VPN and L3VPN PE protection
  - ASBR protection

BACKUP

E2E Packet Flow Impact
Scenario 2

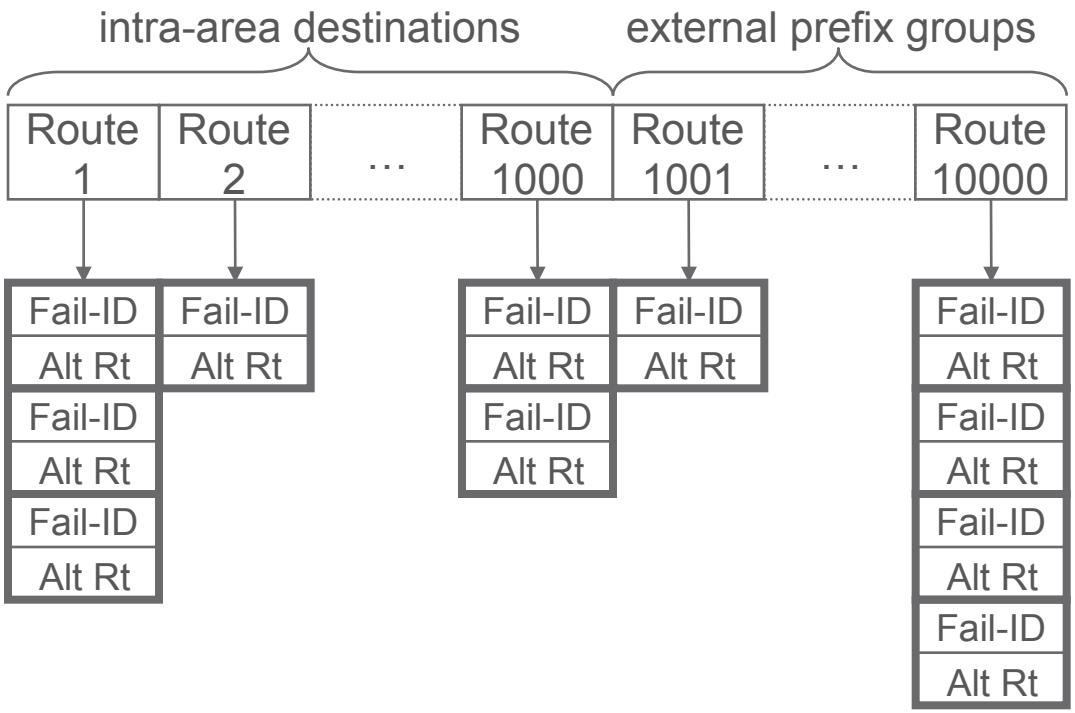
- Traffic flow: 1 pkt / ms
- Varying delay of bold link (D) and length of ring (n)





FIB Update on Linecard from Backup DB

See numbers' origin on slide 8



Main operation: binary search in *each* list

- 50k memory transactions
  - Assuming DRAM with 5MT/sec, and 1 memory controller: 10ms

Underestimate