IOS XR - IP Fast Convergence

Krzysztof Mazepa

kmazepa@cisco.com

Abstract

Mechanisms available on XR platforms (CRS-1, 12k XR) that allows Service Providers to achieve subsecond convergence:

- IGP Fast convergence
- IPoDWDM proactive protection
- BGP Local Convergence Upon PE-CE Link Failure
- BGP Prefix Independent Convergence

Where those mechanisms should be deployed

- Internet IP/MPLS core
- L3 VPN networks (R4 mobile core)
- IPTV networks

Fast convergence (FC)

- The routing protocol detects the failure and computes an alternate path around the failure
- The HW and SW must be optimized
  - **FAST: sub 200msec**
  - **SIMPLE**: without any design or knob tuning, works for any failure, end-to-end
  - **MULTI-SERVICE**: for any service
  - **SCALABLE**

IOS-XR IGP
Fast Convergence

ISIS

- LSP Generation **is optimized by default**
  - `lsp-gen-interval maximum-wait <M>`
    `initial-wait <I> secondary-wait <E>`
  - **Default value of I = 50 msec**
- Flooding & Pacing is optimized by default
- Full SPT computation has been all rewritten and **optimized**
  - `spf-interval maximum-wait <M>`
    `initial-wait <I> secondary-wait <E>`
  - **Default-value if I = 50msec**
  - Full SPT takes [6, 9] msec for a 1000-router Tier1 network (real test)
- Incremental SPT
  - `ispf [level-1 | level-2]`

ISIS

- **Prefix Prioritization**
  - 4 priorities: **Critical**, **High**, Medium, Low
  - /32 IPv4 and /128 IPv6 prefixes are classified by default in **Medium Priority**
  - Rest is classified by default in Low Priority
- **Customization**
  - `spf prefix-priority`
  - This command supports prefix list for the first three priorities. The unmatched prefixes will be updated with low priority.
  - As soon as the “prefix priority” command is used, then the /32 heuristic is no longer applied. If you then want to keep the /32’s in medium, you need to configure the medium ACL so.

ISIS

- **Prefix Prioritization is THE key behavior**
  - CRITICAL: **IPTV SSM sources**
  - HIGH: **Most Important PE's**
  - MEDIUM: All other PE's
  - LOW: All other prefixes
- **Prefix prioritization customization is required for CRITICAL and HIGH**

Reference slide
ISIS: Prefix Priority Customization

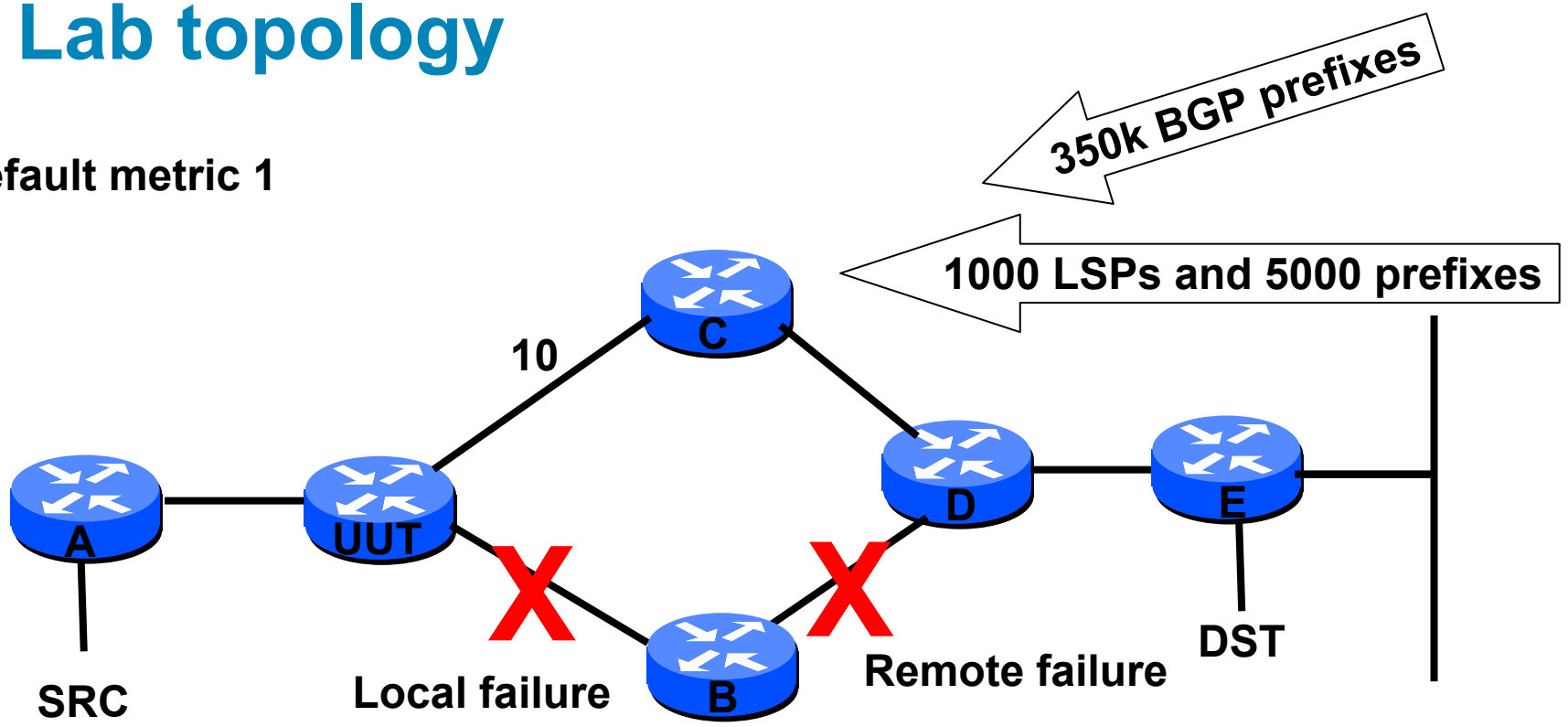
```
ipv4 prefix-list isis-critical-acl
10 permit 0.0.0.0/0 eq 32
ipv4 prefix-list isis-high-acl
10 permit 0.0.0.0/0 eq 30
ipv4 prefix-list isis-med-acl
10 permit 0.0.0.0/0 eq 29

router isis 1
 address-family ipv4 unicast
   spf prefix-priority critical isis-critical-acl
   spf prefix-priority high isis-high-acl
   spf prefix-priority medium isis-med-acl
```

IGP Case Study

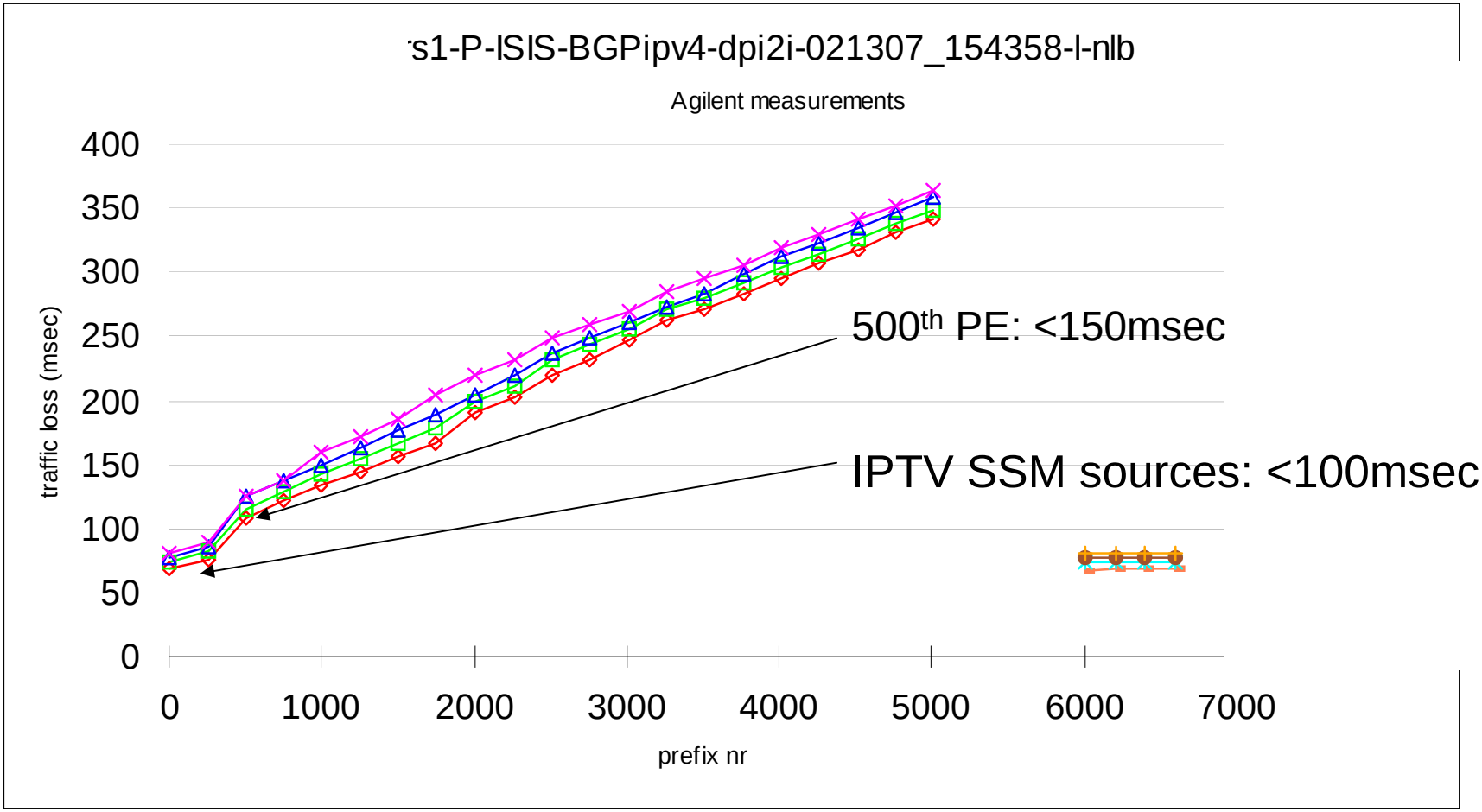
Lab topology

Default metric 1



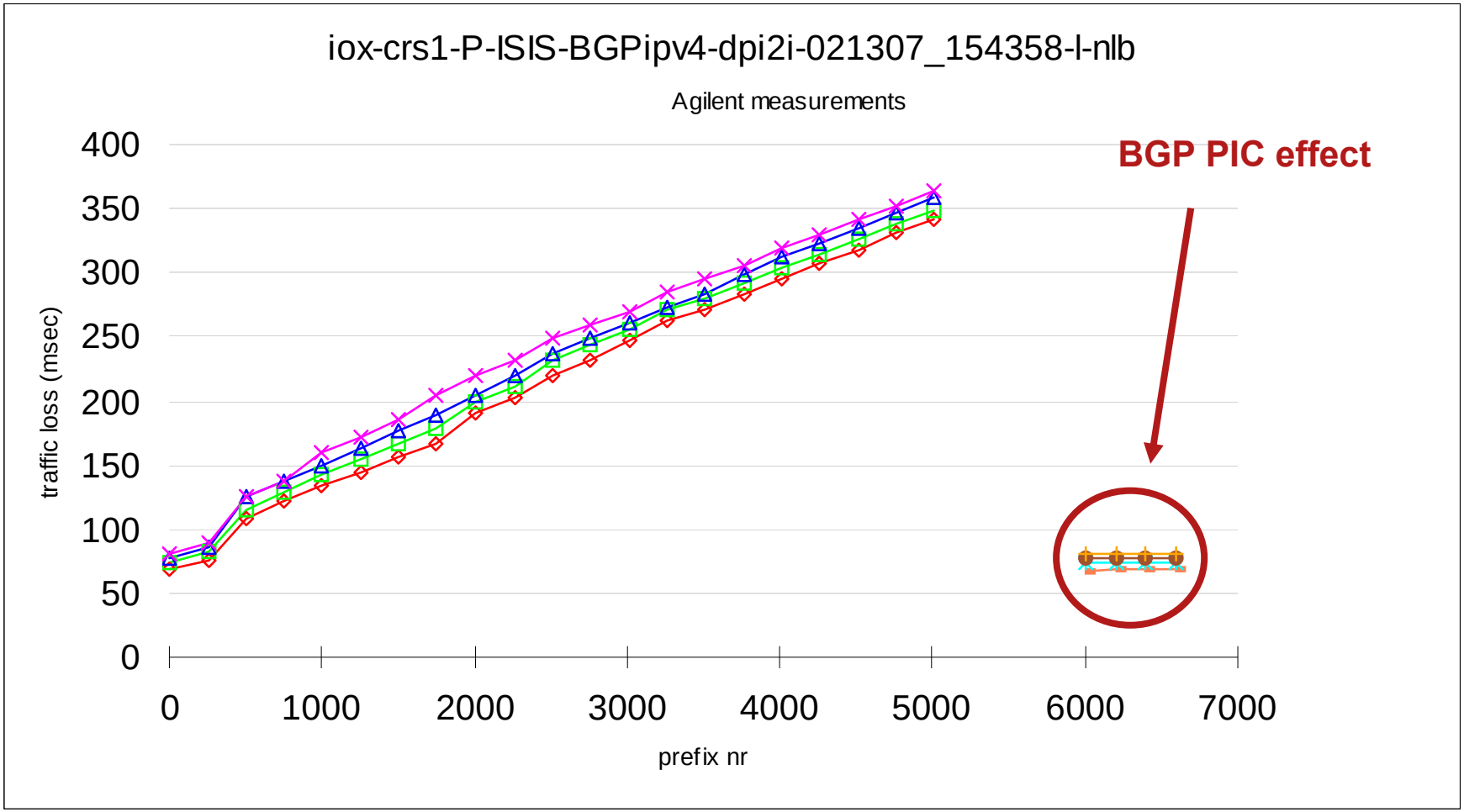
- ISIS priority:
  - 10 /24 in critical
  - 500 /32 in high
- BGP-NH is first of high

ISIS FC in large-scale T1 network



Testbed: Tier1 ISP topology (1000nodes), CRS1, IOX3.5, 5000 ISIS prefixes, 350k IPv4 BGP dependents to impacted BGP nhop

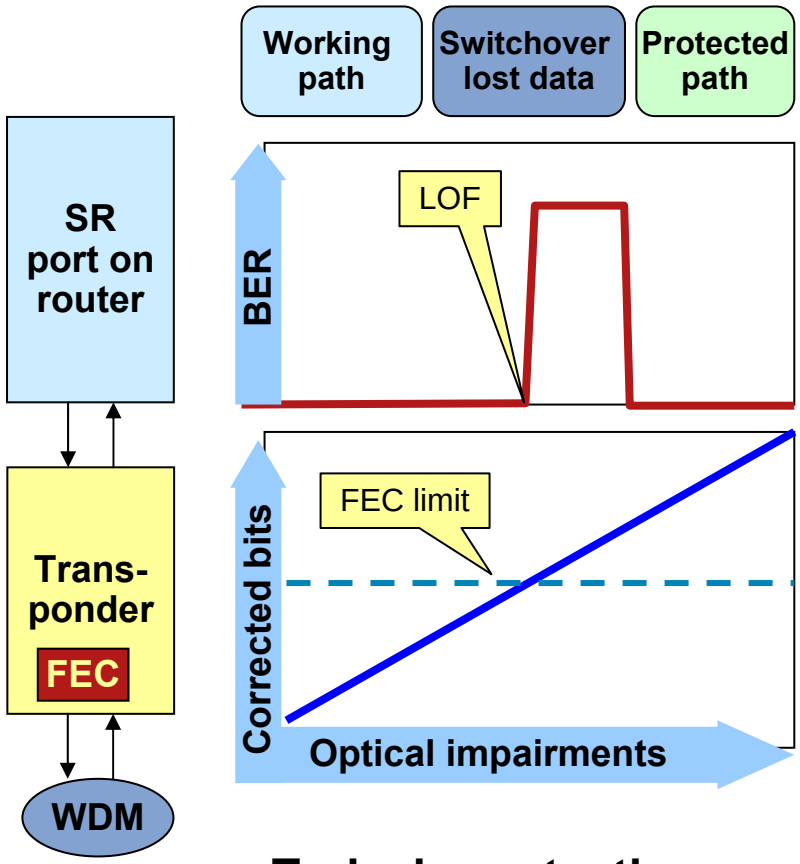
ISIS and BGP PIC core in a T1 ISP



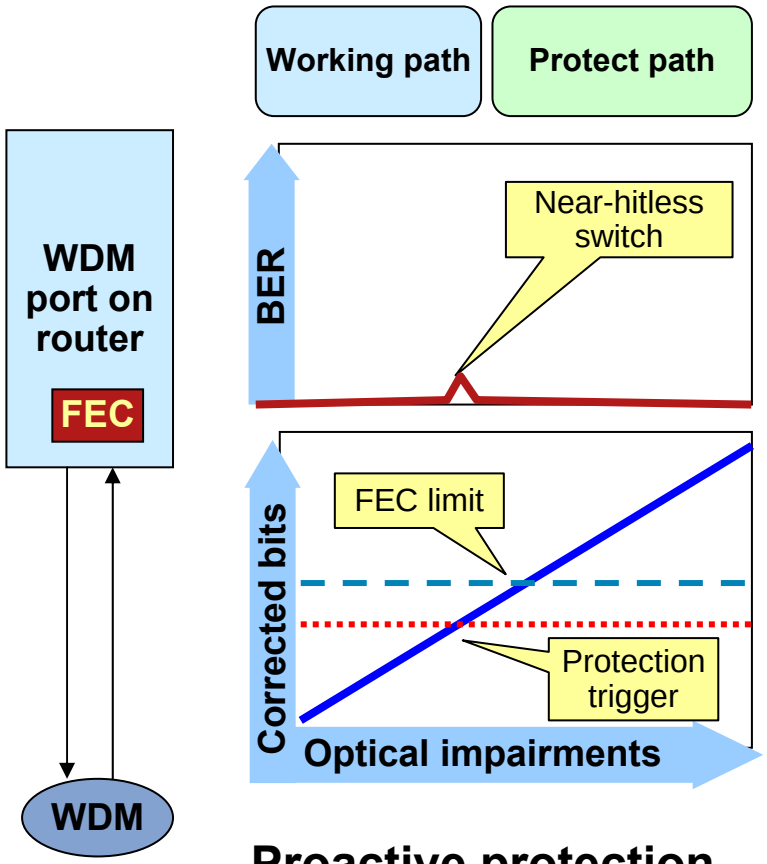
Testbed: Tier1 ISP topology, CRS1, IOX3.5, 5000 ISIS prefixes, 350k IPv4 BGP dependents to impacted BGP nhop

IOS-XR IPoDWDM / Routing Integration

IP-over-DWDM Advanced Protection feature



Today's protection



Proactive protection

Superior Protection Compared to Transponder-Based Networks

Test Matrix - MPLS FRR

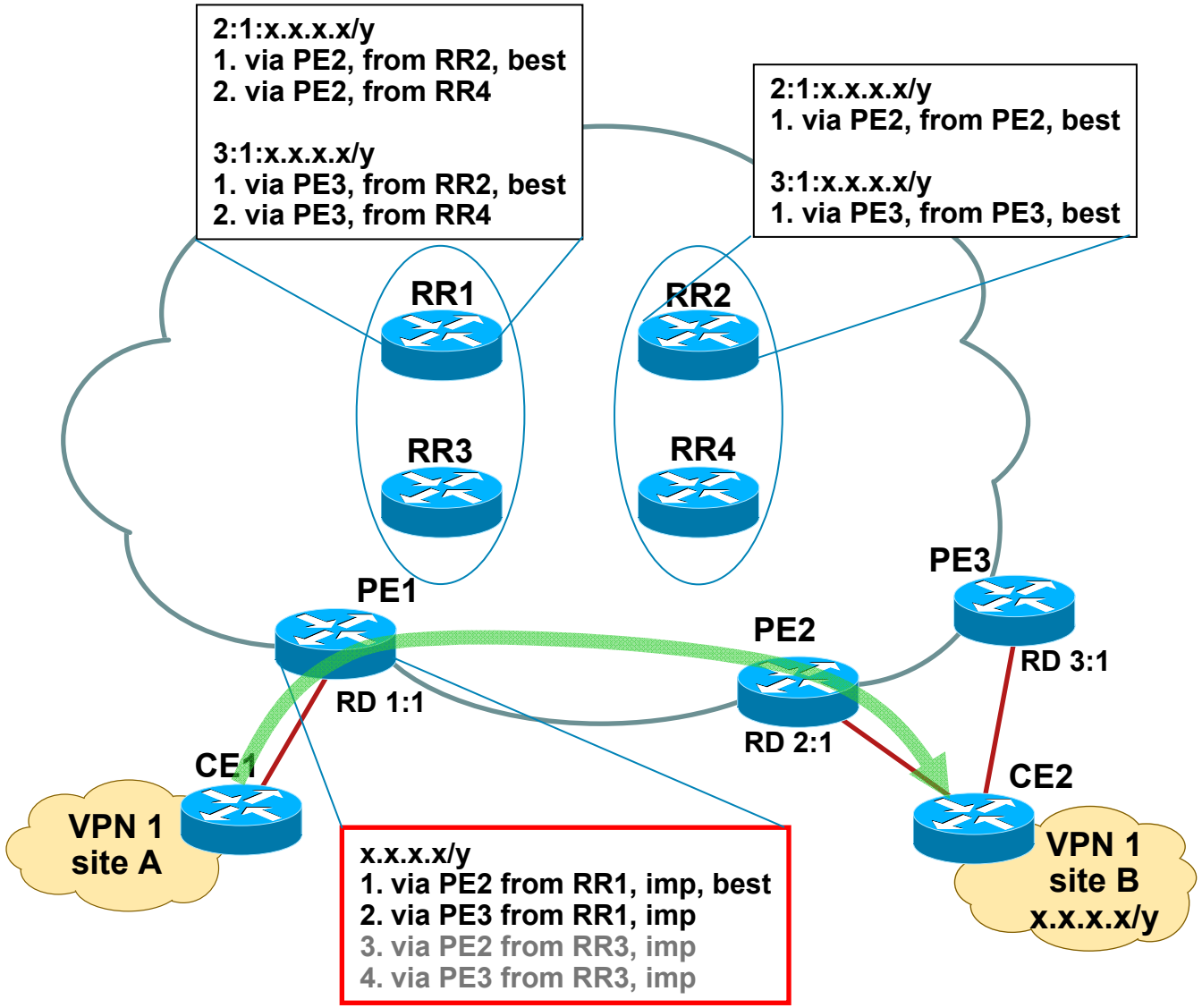
pre-FEC FRR	Fault	Packet Loss (ms)		
		Highest	Lowest	Average
Y	Optical-switch (25ms)	11.48	10.99	11.24
Y	Noise-injection	0.12	0	0.05
Y	Fibre-pull	14.97	0	4.97
N	Optical-switch (25ms)	11.61	11.16	11.32
N	Noise-injection	2852	2602	2727
N	Fibre-pull	83.43	13.49	37.63

Pre-FEC Results for ISIS Fast Convergence

Pre FEC	Fault	Max. Packet Loss (ms)		
		C(500)	C(1000)	C(1)
Y	Optical-switch	170	220	163
Y	Slow noise-injection (0.1dB/1000ms)	3	12	0
Y	Fast noise-injection (0.5dB/500ms)	3	9	0
N	Optical-switch	180	205	159
N	Slow noise-injection (0.1dB/1000ms)	2990	3035	2880
N	Fast noise-injection (0.5dB/500ms)	596	620	540

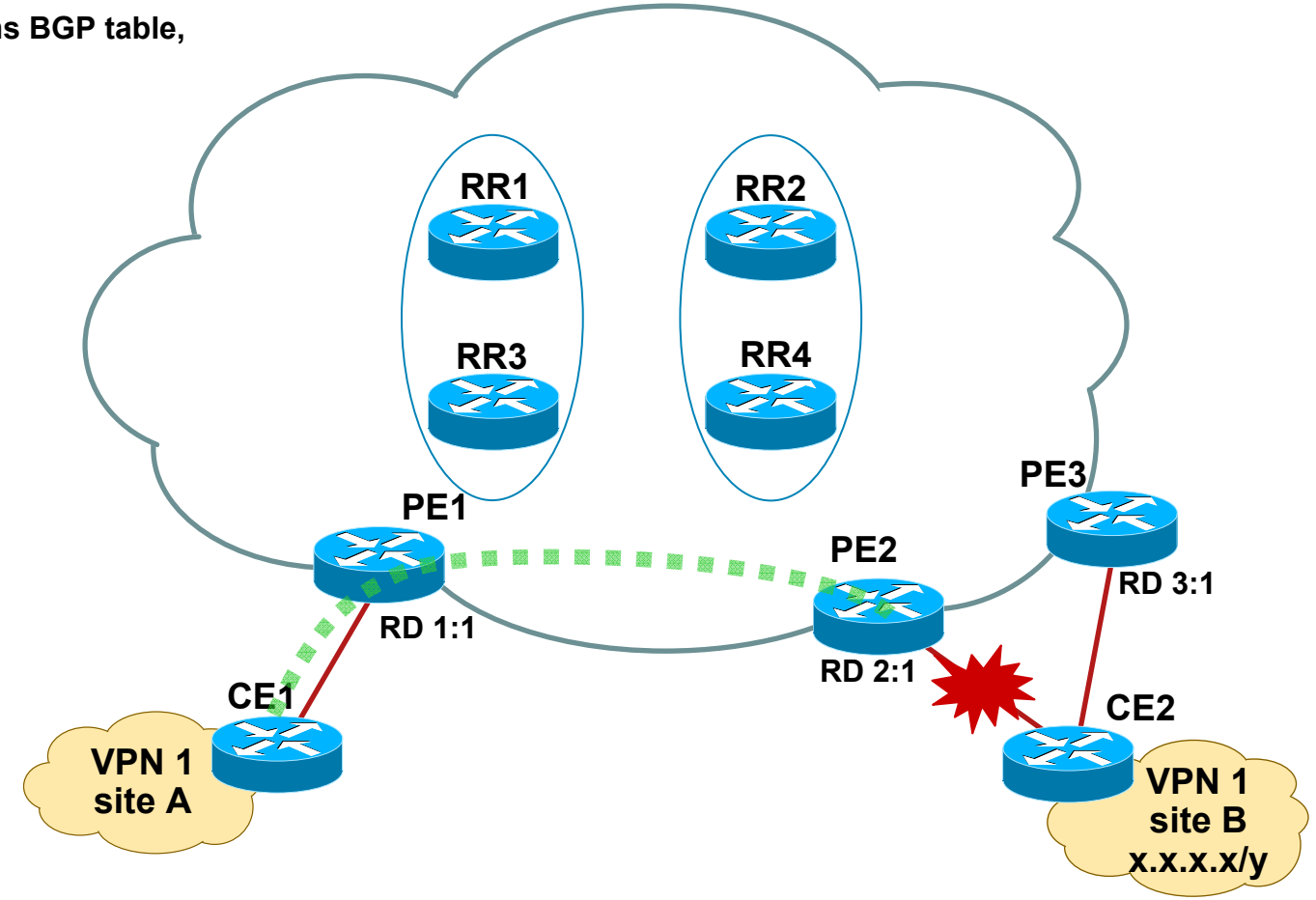
IOS-XR BGP Fast Convergence

PE-CE link failure



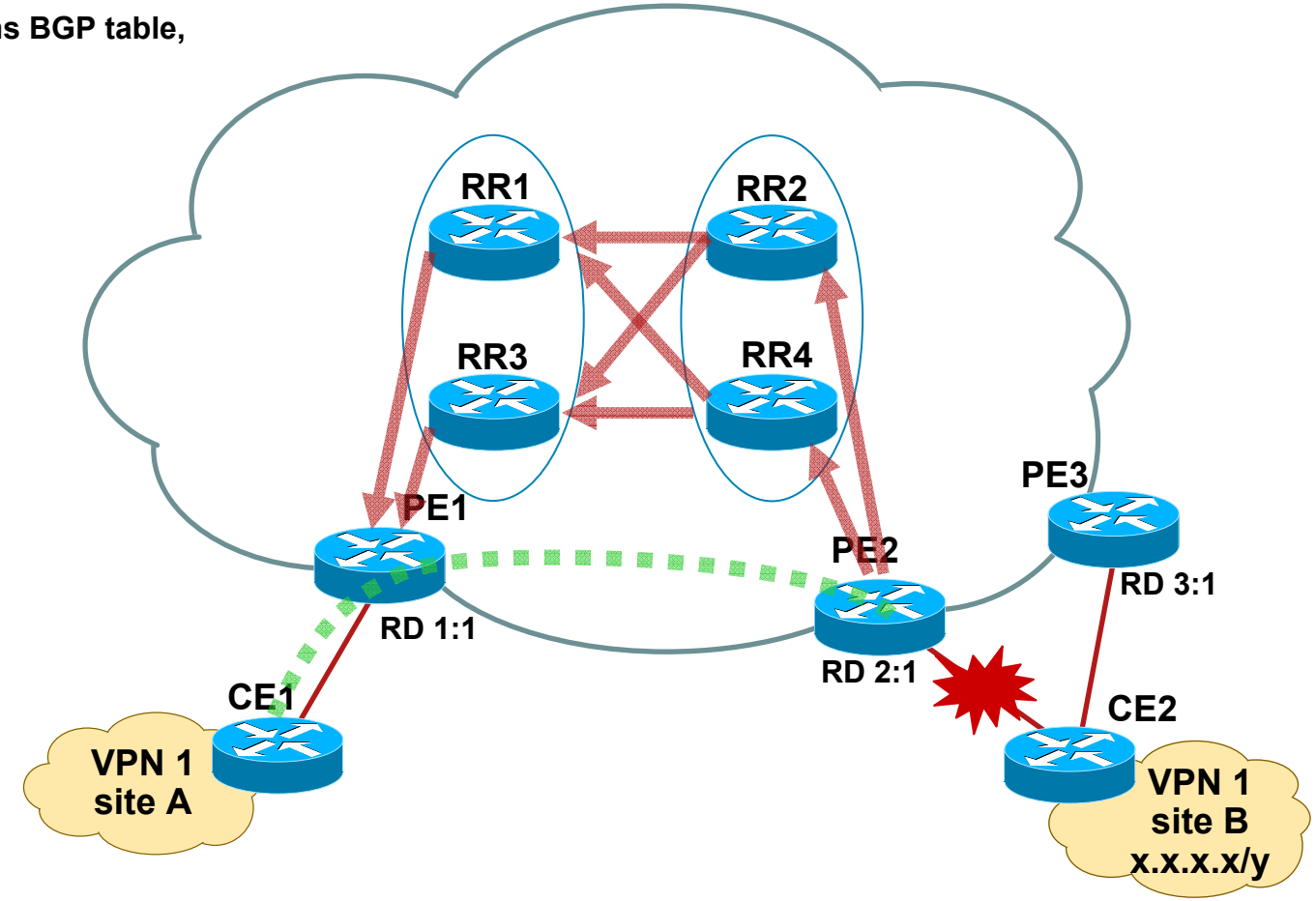
PE-CE link failure

1. link PE2-CE2 fails
2. Fast External Fallover scans BGP table, calculating new bestpaths



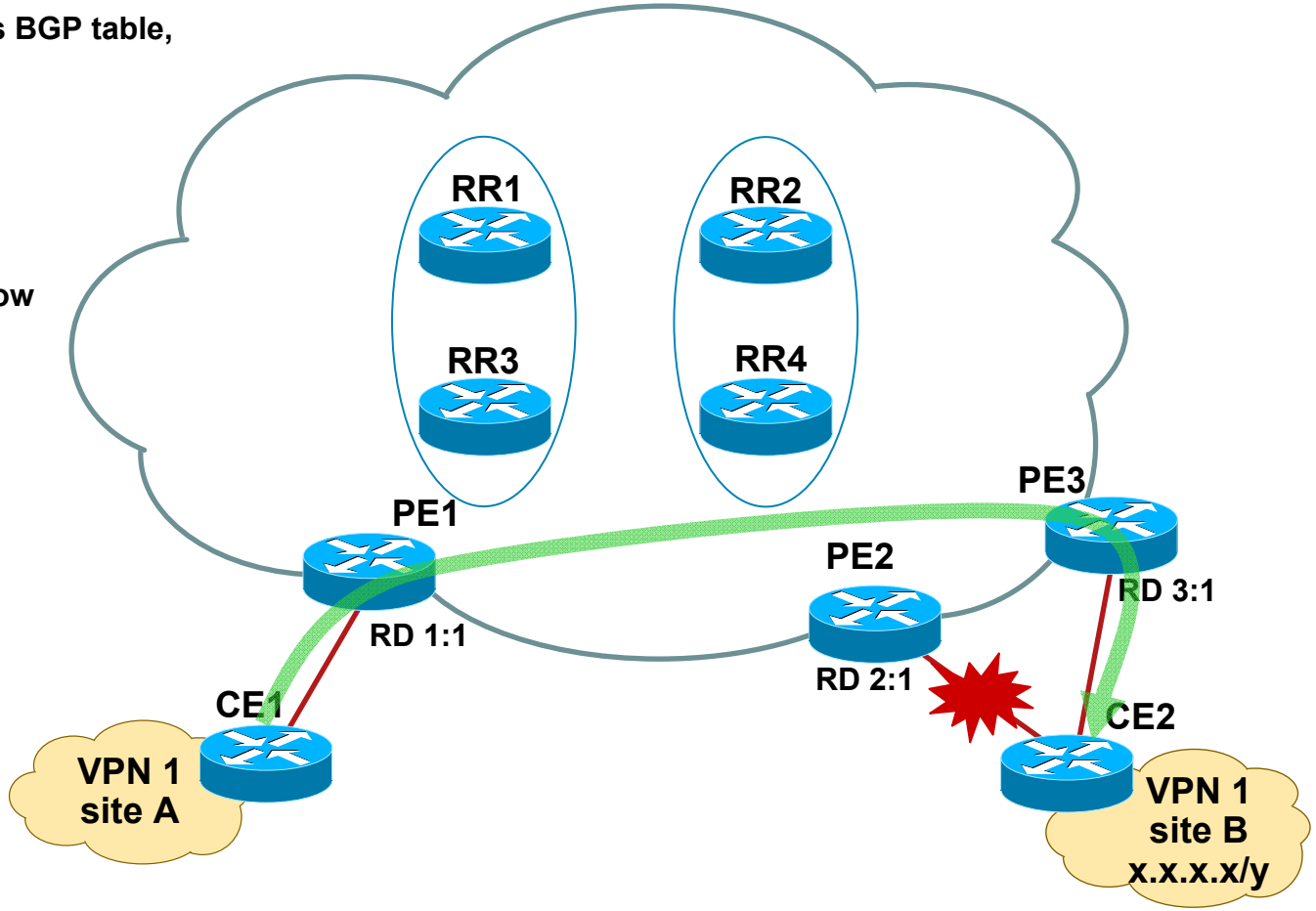
PE-CE link failure

1. link PE2-CE2 fails
2. Fast External Fallover scans BGP table, calculating new bestpaths
3. PE2 withdraws paths
4. RR2 and RR4 propagate withdraws
5. RR1 and RR3 propagate withdraws



PE-CE link failure

1. link PE2-CE2 fails
2. Fast External Fallover scans BGP table, calculating new bestpaths
3. PE2 withdraws paths
4. RR2 and RR4 propagate withdraws
5. RR1 and RR3 propagate withdraws
6. PE1 deletes path via PE2, now going via PE3

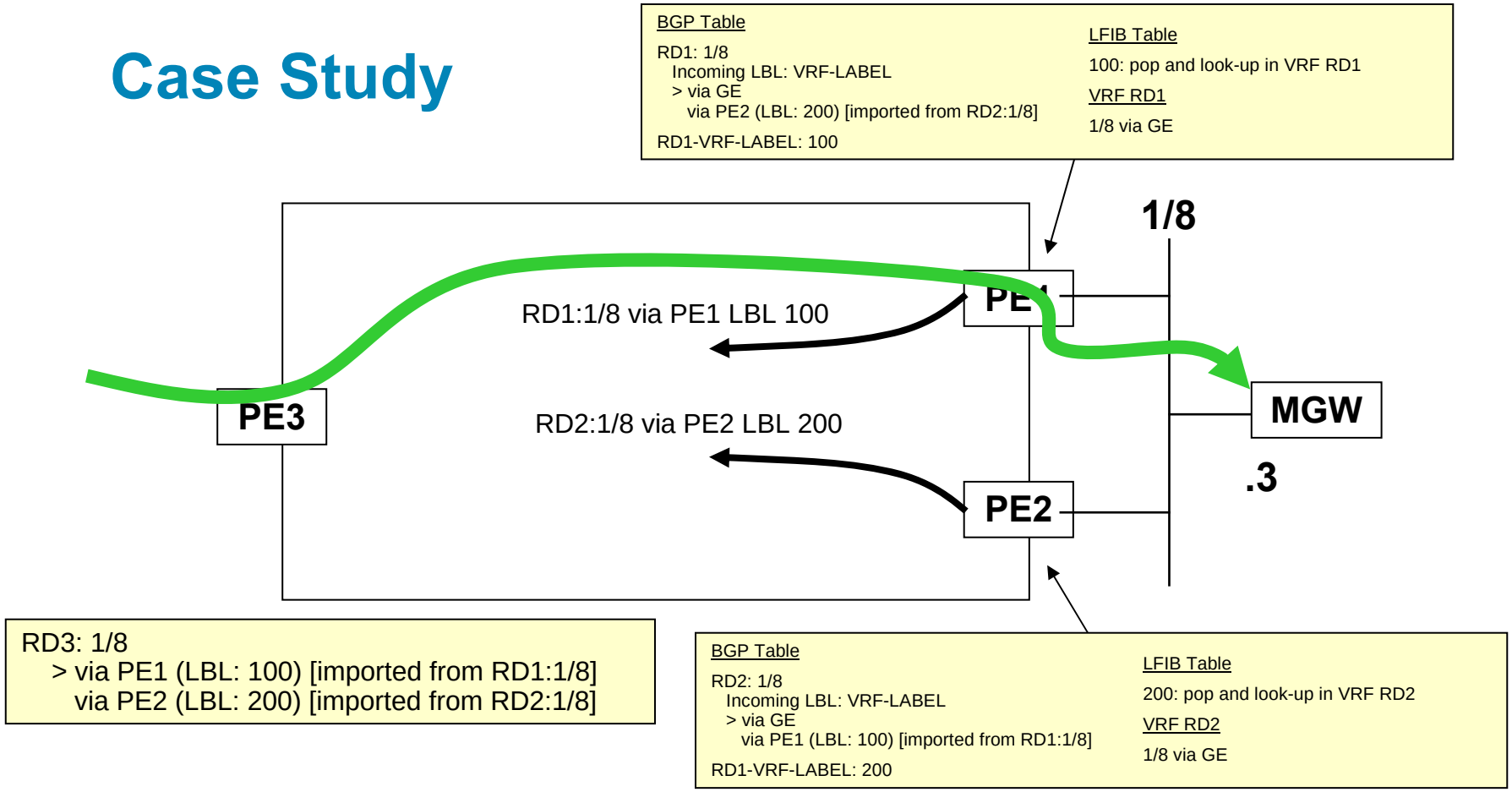


BGP Local Convergence Upon PE-CE Link Failure

Case Study

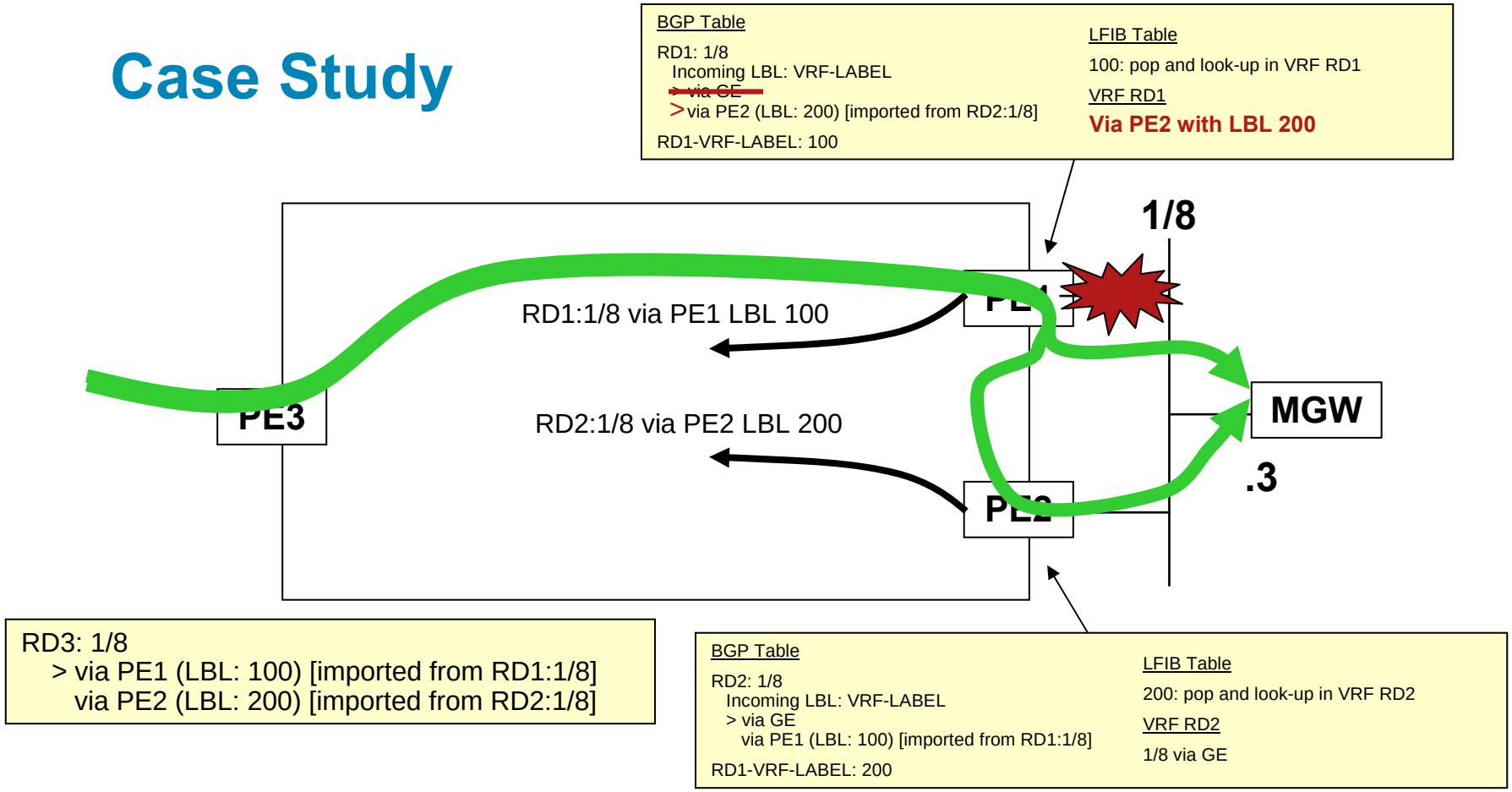
- Premium VPN with very tight convergence requirement (eg. VoIP MGW…)
- Connectivity to VPN customer can either be as a simple GE host, or via static route or via eBGP over any type of link
- SP requires a solution which allows for a multiservice PE (Internet + VPN's)
- Failure scenario: PE-CE Link Down or PE-CE Link Up
- Assumption: unique RD's are used

Case Study



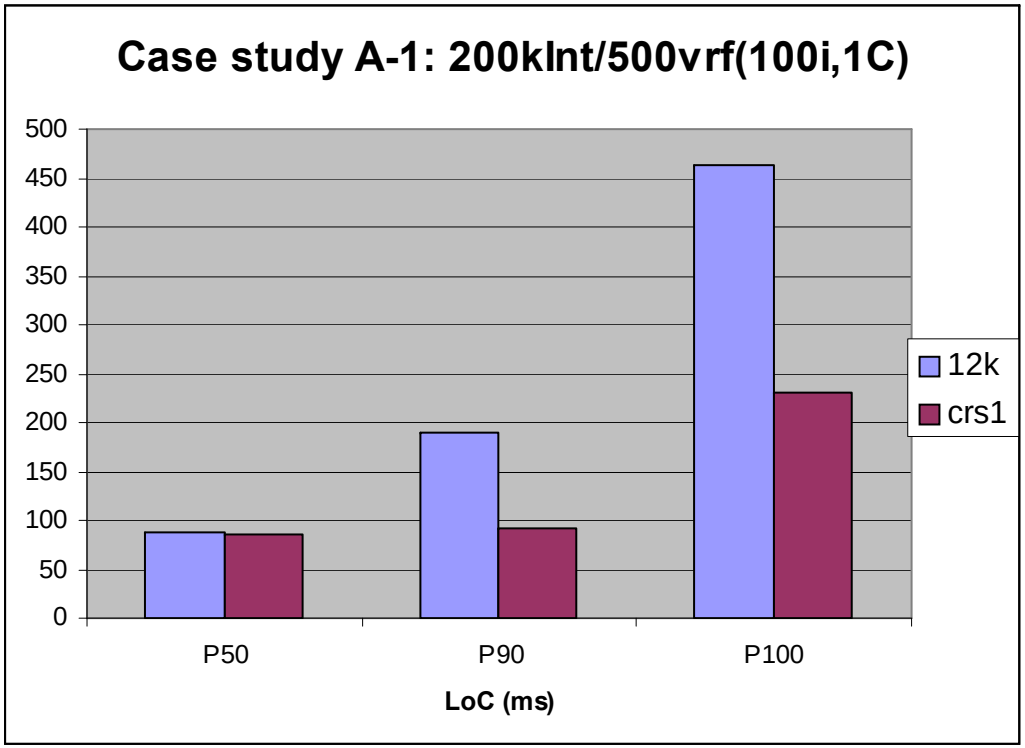
- An aggregate label is dedicated to a VRF. It is available as long as the VRF exists.
- Upon matching such a label, a receiving PE pops that label and look-up the resulting IP packet in the VRF table pointed out by the VRF Label

Case Study



- Upon local interface failure (same for BFD failure detection if MGW supports BFD), PE1's RIB immediately notifies BGP which immediately finds and updates the impacted route (no scan). PE1's BGP selects an alternate (already-known) path and inserts it in the VRF.
- Even if PE3 still routes through PE1, any packet received at PE1 with LBL=100 and IP DA=1.*.*.* is redirected via PE2
- **BGP Local Convergence**

Very Conservative Connected Case Study



200k internet routes

50k VPNv4 routes

500 VRF's, each with 100 ibgp prefixes and 1 connected prefix

Event: Interface to VoIP VRF goes down

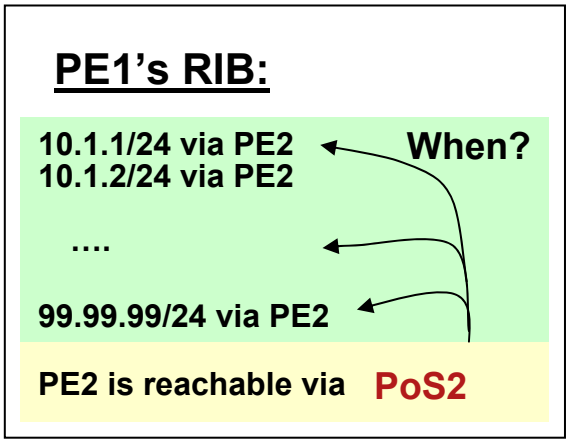
Results obtained with IOX 3.3

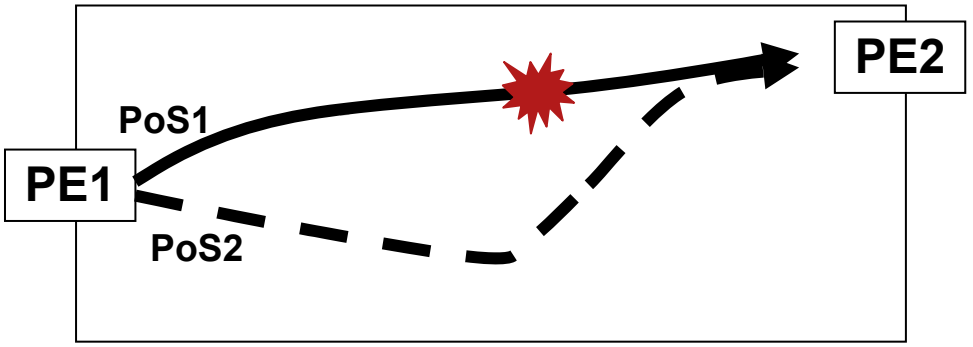
A multiservice PE with 200k IPv4 routes, IPv4 peerings, 500 VPN's, 50k VPNv4 routes detects and reroutes around the failure to a critical VPN site in less than 250ms with CRS/IOX3.3 (the median is < 100ms!)

IOS-XR BGP Prefix Independent Convergence

CORE Scenario

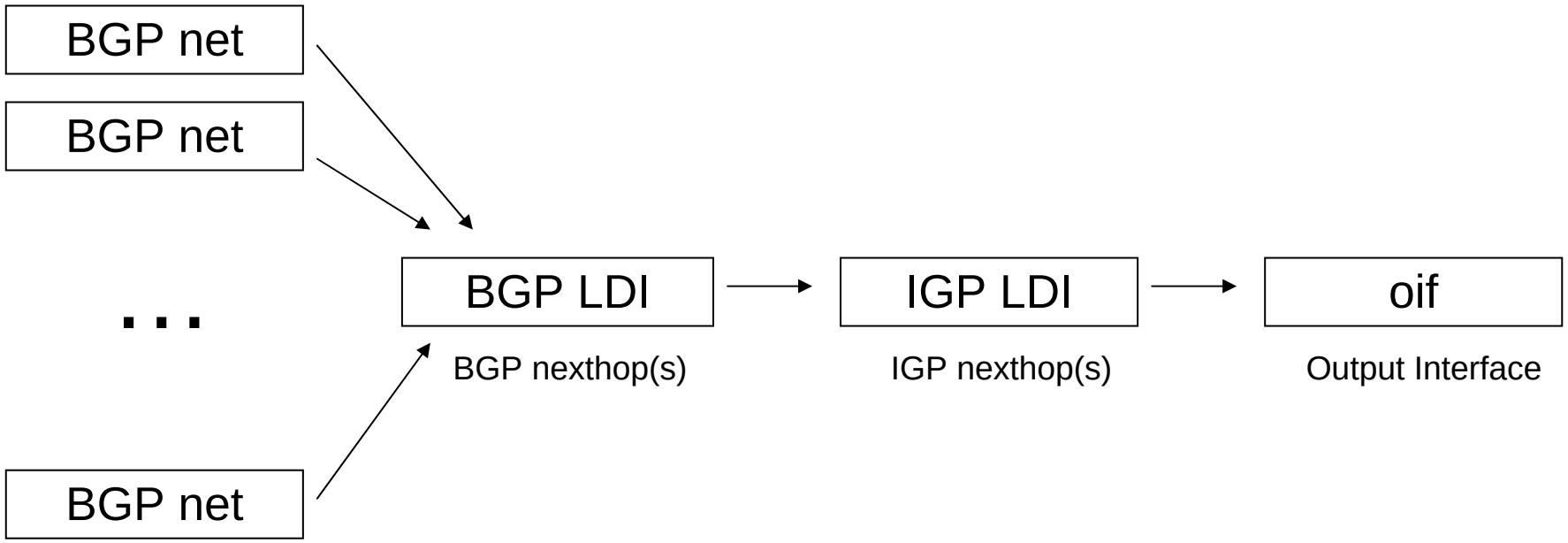
BGP Prefix-Independent Convergence Core failure





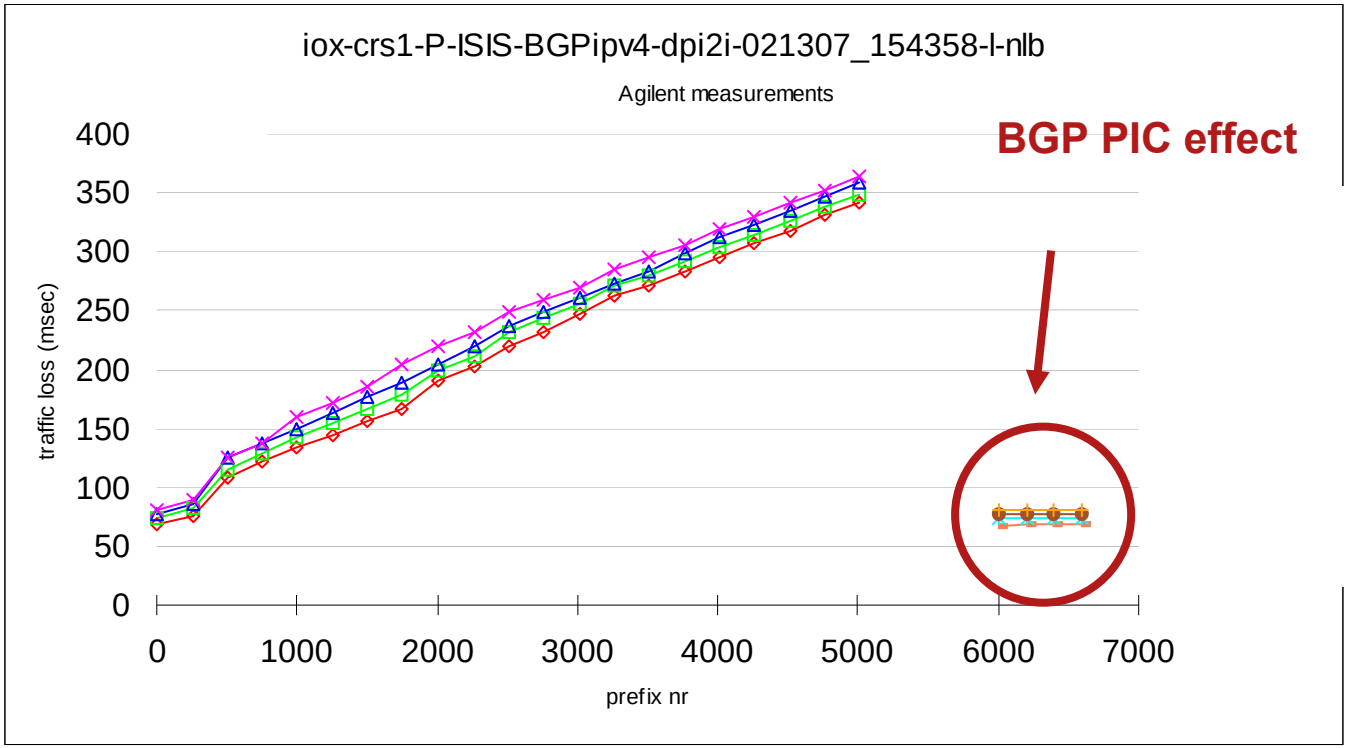
- Upon a core failure, the IGP finds an alternate path to the BGP next-hop PE2 in a few 100's of msec
- Requirement: the BGP prefixes depending on reachability to PE2 must leverage the new ISIS path as soon as it is updated in the FIB

The right architecture: hierarchical FIB



- Pointer Indirection between BGP and IGP entries allow for immediate leveraging of the IGP convergence

The right behavior is “BGP PIC/core”



- Testbed: Tier1 ISP topology, CRS1, IOX3.5, 5000 ISIS prefixes, 350k IPv4 BGP dependents to impacted BGP nhop
  - same behavior for VPNv4 as of 3.5
- When ISIS converges, all the BGP dependents immediately leverage the ISIS convergence

Conclusion

Simplicity

- IGP, BGP and PIM are meant to route around failure
- Goal: meet the restoration targets without externalizing any complexity to the designer/operator
  - the complexity/intelligence is in the SW/HW implementation
  - all behaviors are tuned by default
  - no complex design to conceive or operate
- Simplicity = Cost Savings for the designer/operator

Reality

IGP sub-second convergence is very conservative

- Tier1 lead customer achieving mythical sub-200msec

- BGP convergence for VoIP VPN < 500 msec
- IPTV SSM convergence for 400 channels in < 300msec
- Border router failure with 300k BGP prefixes in < 1sec
- Failure Frequency
  - An european backbone (2003): Mean Time Between Link Failure > 230 hours

Dziękuję za uwagę, komentarze i pytania

kmazepa@cisco.com

