



Cisco Expo  
2010

## Advanced SP Technologies



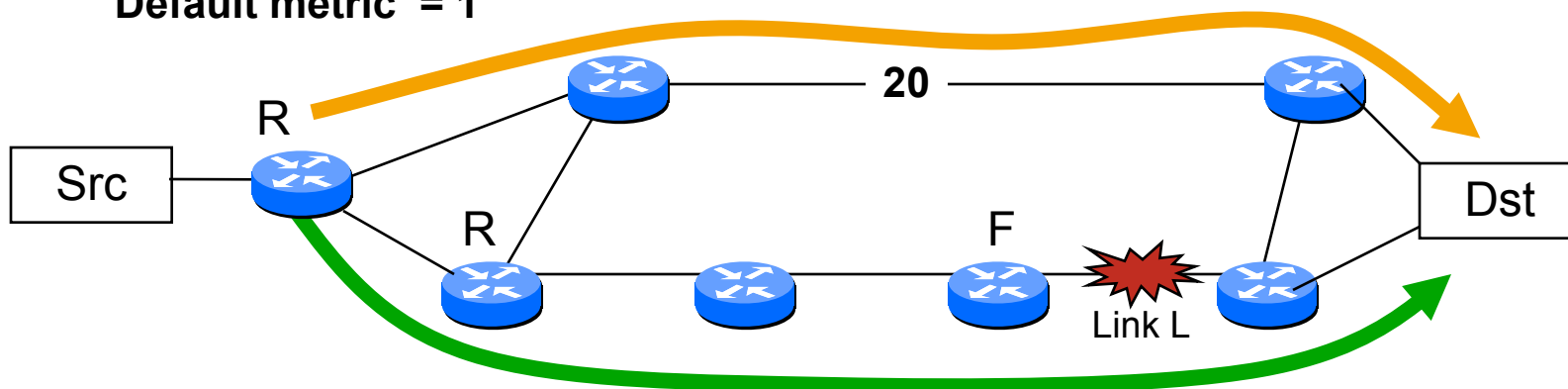
**Fast Convergence on  
CRS, 12k, IOS XR**

**Ovidiu Neghina**

# Convergence



Default metric = 1



1. Assume a flow from Src to Dest
2. T1: when L dies, the best path is impacted  
loss of traffic
3. T2: when the network converges, a next best path is computed  
traffic reaches the destination again
4. Loss of Connectivity: T2 – T1, called “Convergence” hereafter
5. Analyzed for streams going to IGP and BGP learned prefixes

# Fast convergence (FC)

1. The routing protocol detects the failure and computes an alternate path around the failure
2. 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

# The value is in the implementation

1. The ability to detect, compute, update or enable prefix-independent precomputed alternate paths is essentially dependant on the excellence of
  - the hardware
  - the platform independent software (BGP, IGP, PIM, LDP)
  - the platform dependent software (FIB)
2. Any externalization of a tuning is a sign of lack of excellence

# The value is in the implementation

## 1. Hardware

CRS and 12k are the first platform in the industry to support a two-level hierarchical FIB dataplane. This is a key innovation with fundamental consequences for BGP services. The loss of availability upon BGP failure goes from several minutes to 100's of msec

## 2. PI Software

We rewrote the entire ISIS code to optimize for prioritized fast convergence. From 3.2 to 3.3, factor 12 gains were obtained

## 3. PD Software

The 12k FIB update is improved by a factor 9 from 3.3 to 3.8

The CRS FIB update is improved by a factor 3 from 3.3 to 3.8

# The value is in the implementation

## 1. No tuning is required

not in OS, not in ISIS, not in OSPF, not in RIB, not in FIB, not in LDP, not in BGP, not in PIM

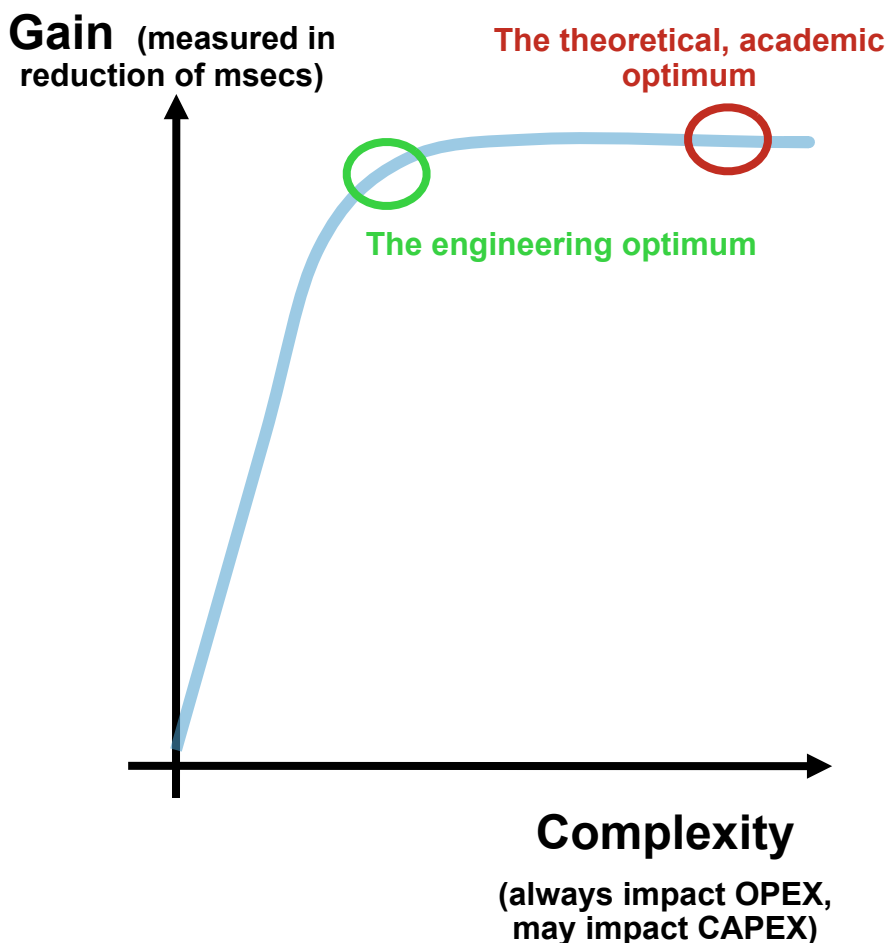
the goal of the program is for the router to deal with any failure (within the AS, Outside the AS, Link, node, SRLG) without any tuning or specific design.

## 2. For IOS XR FC, simplicity is more important than speed

but we want to achieve both

# The Simplicity Principle

1. Simplicity is the guiding rule of all the IOS XR FC technology
2. Note: it sounds obvious but there are numerous examples where the optimum could easily be missed



# Agenda

1. IOS-XR IGP Fast Convergence
2. IOS-XR BGP Fast Convergence
3. IOS-XR BGP Prefix Independent Convergence
4. Conclusion

# IOS-XR IGP Fast Convergence



# ISIS

1. LSP Generation is optimized by default

```
lsp-gen-interval maximum-wait <M>  
                    initial-wait <I> secondary-wait <E>
```

Default value of I = 50 msec

2. Flooding & Pacing is optimized by default
3. 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)

4. Incremental SPT

```
ispf [level-1 | level-2]
```

# ISIS

## 1. 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

## 2. 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

## 1. Prefix Prioritization is THE key behavior

CRITICAL: IPTV SSM sources

HIGH: Most Important PE's

MEDIUM: All other PE's

LOW: All other prefixes

## 2. Prefix prioritization customization is required for CRITICAL and HIGH

# ISIS

1. Prefix Prioritization Customization can be driven
  - at the converging node: ACL/RPL
  - by the origin: Prefix tagging

# ISIS: Prefix Priority Customization

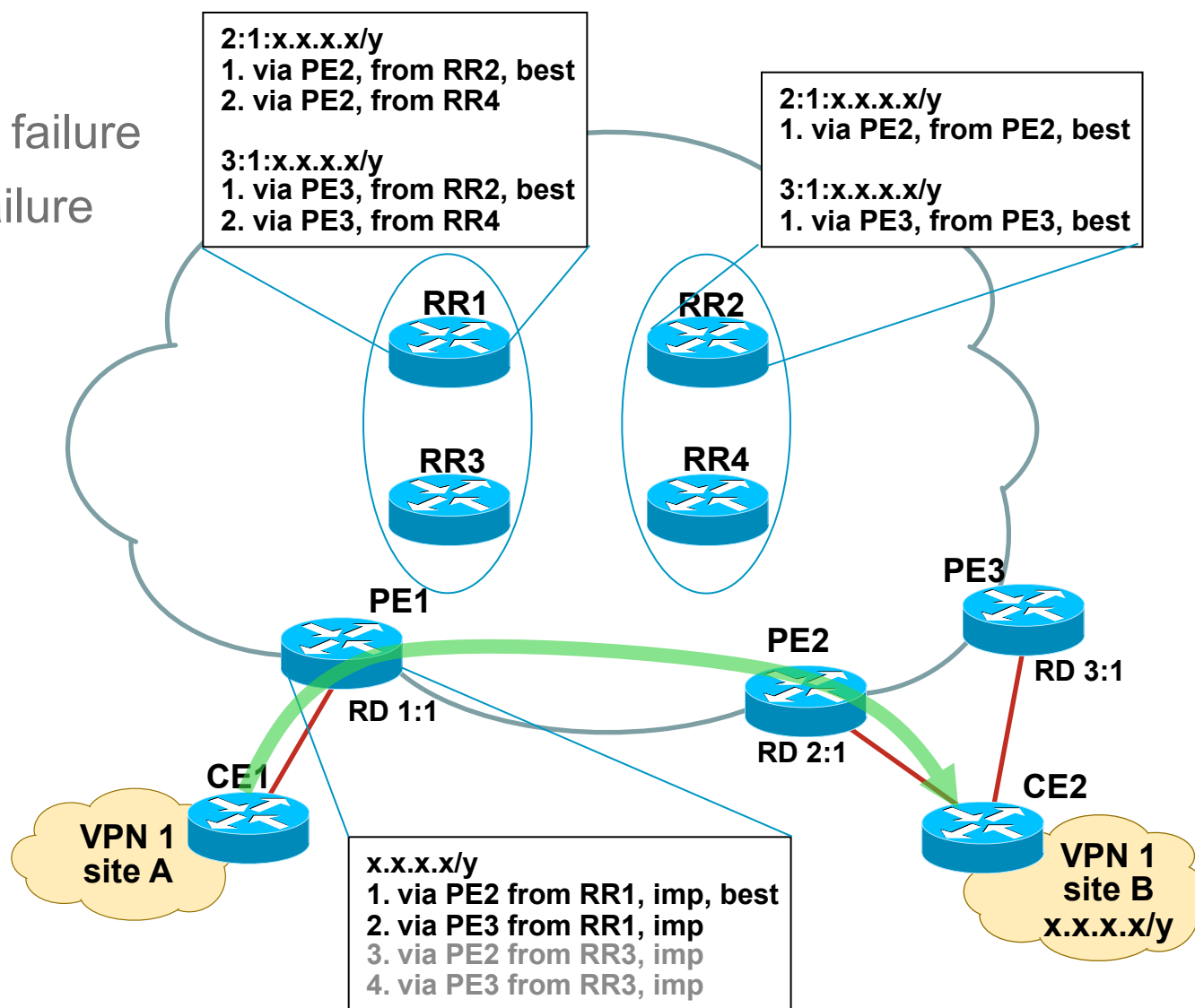
```
ipv4 prefix-list isis-critical-acl
10 permit 0.0.0.0/0 eq 32
ipv4 prefix-list isis-high-acl
10 permit 0.0.0.0/0 eq 30
ipv4 prefix-list isis-med-acl
10 permit 0.0.0.0/0 eq 29
router isis 1
  address-family ipv4 unicast
    spf prefix-priority critical isis-critical-acl
    spf prefix-priority high isis-high-acl
    spf prefix-priority medium isis-med-acl
```

# IOS-XR BGP Fast Convergence



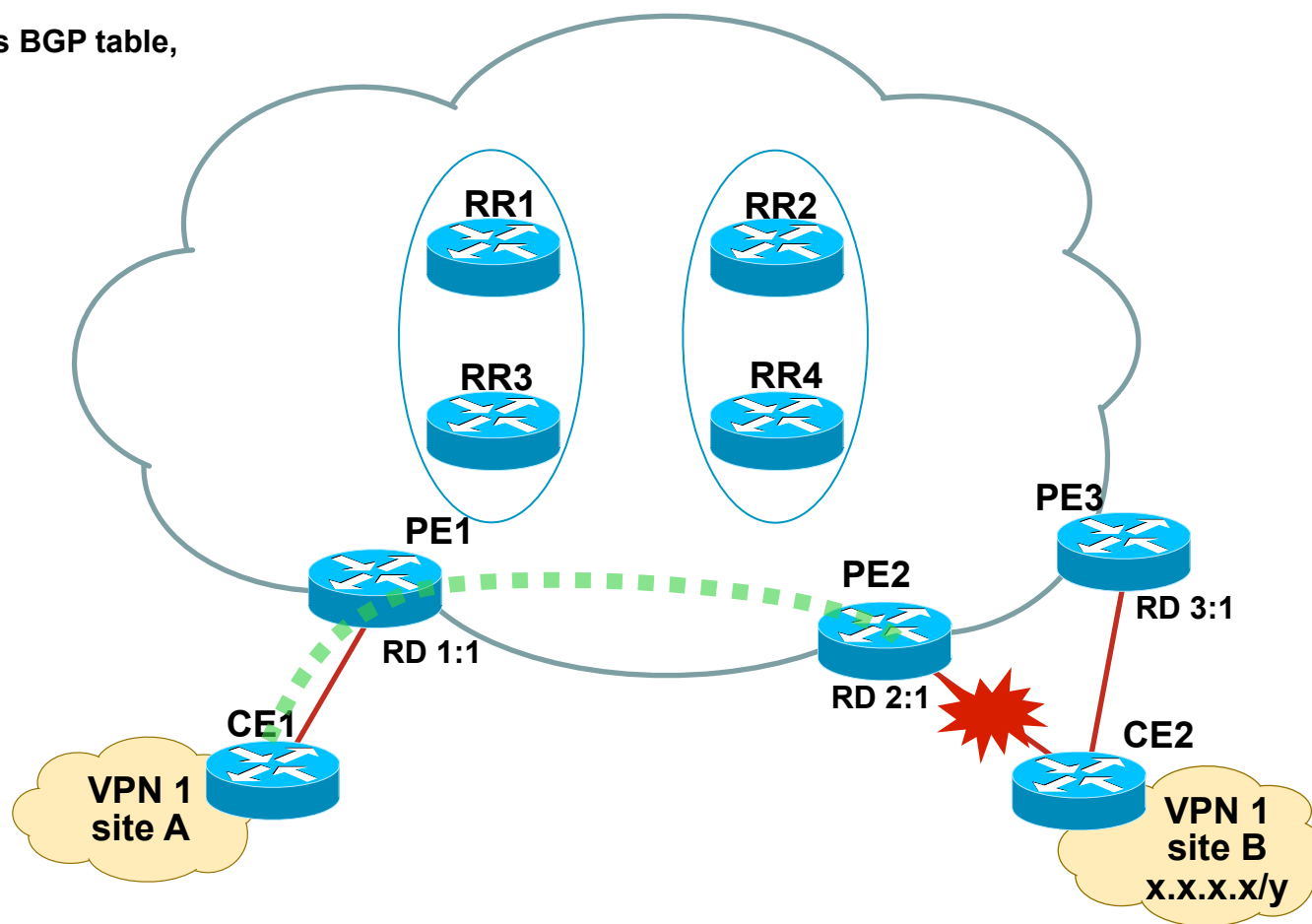
# Failures

1. PE-CE link failure
2. PE node failure



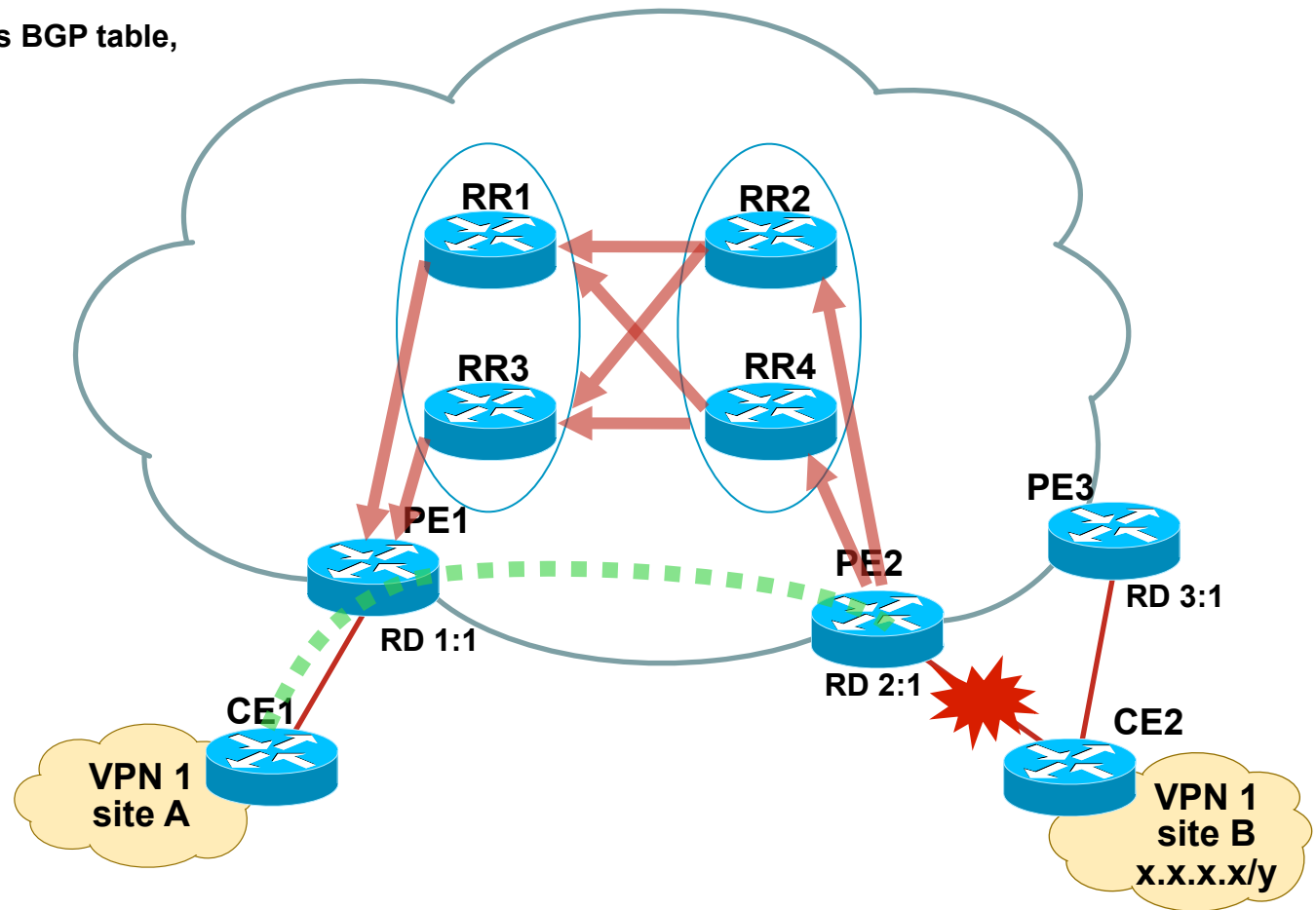
# PE-CE link failure

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



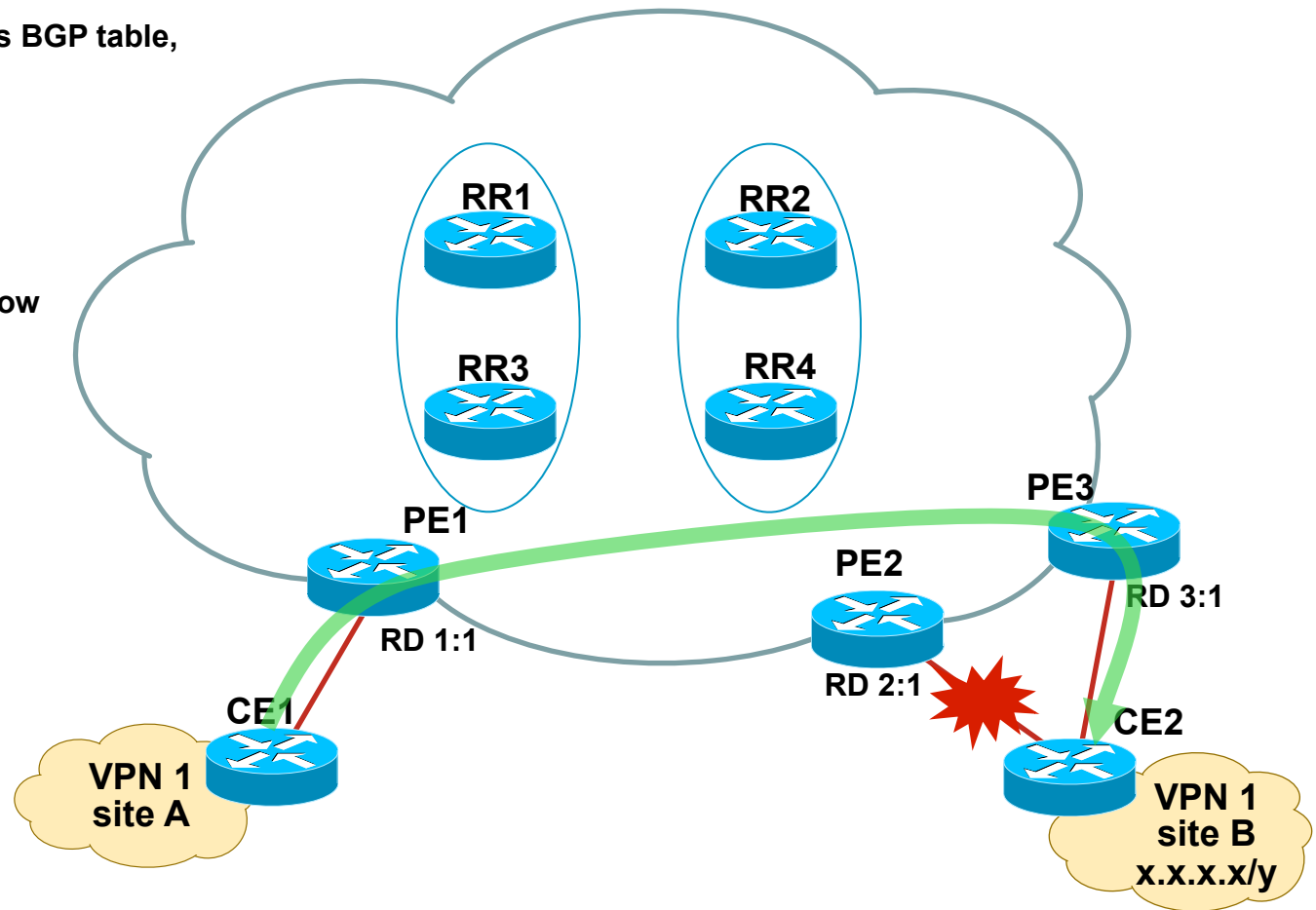
# PE-CE link failure

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



# PE-CE link failure

Convergence depends on

D: time to detect failure

S(p): time to scan BGP table

Per-RD walk for VPNv4 and then IPv4

B(p): time to compute bestpath for impacted routes and update FIB

Wtx(p): time to generate/propagate all withdraws

RR(p): time for the RR reflection

Wrx(p): time to receive and process all withdraws

B(p): time to compute bestpath for impacted routes and update FIB

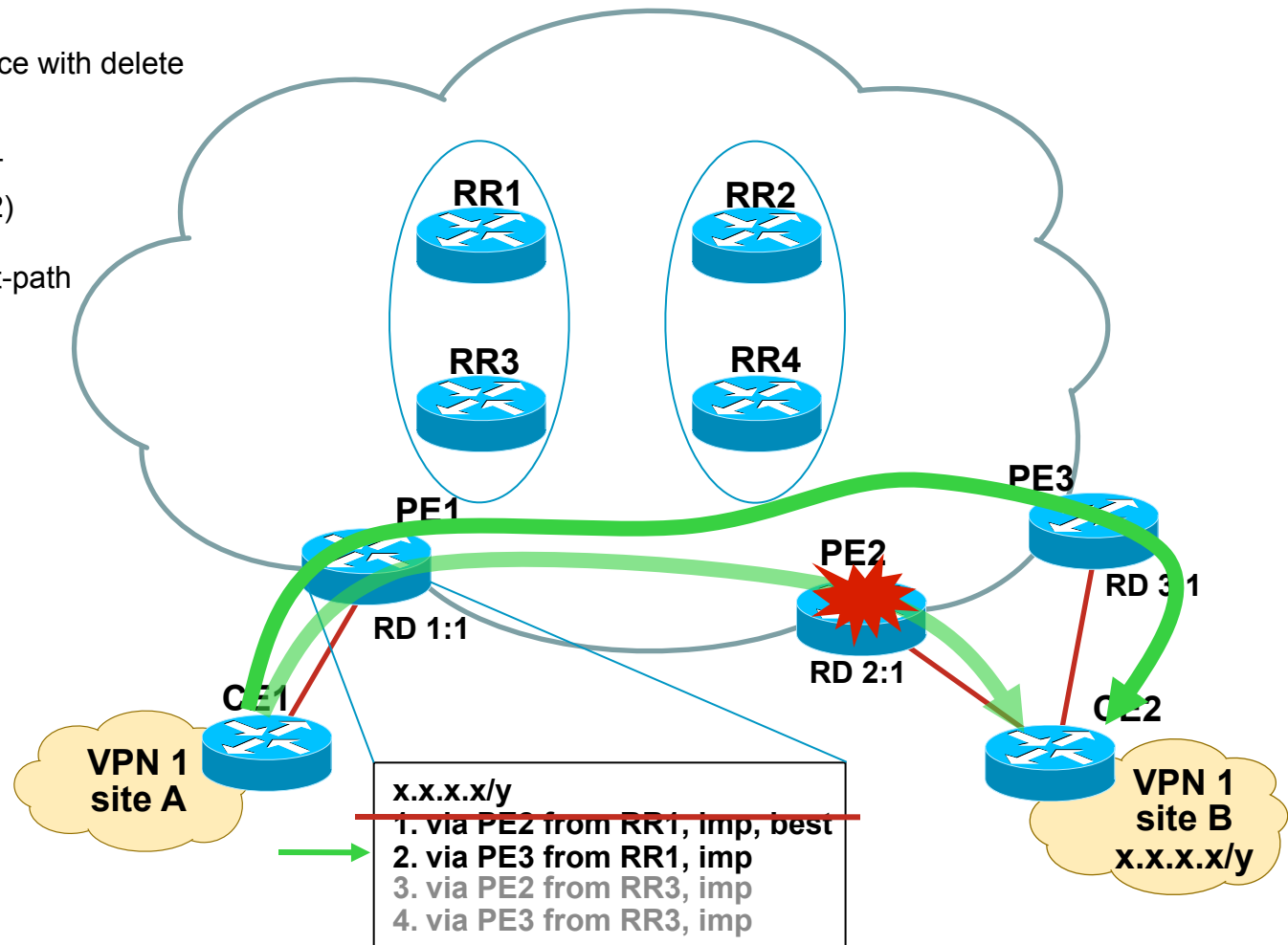
Where X(p) means that this component scales with the table size

**Eliminated  
with  
BGP PIC**

**Eliminated  
With  
BGP Local  
Convergence  
and  
BGP PIC**

# PE Node Failure

1. IGP neighbors detect the loss of PE2, re-originate and flood their LSP
2. PE1 completes IGP convergence with delete (PE2)
3. On PE1, RIB notifies BGP NHT of the loss of a BGP nhop (PE2)
4. On PE1, BGP recomputes best-path and selects an alternate nhop



# PE Node Failure

Convergence depends on

D: time to detect failure

IGP: time to complete a simple IGP convergence

simple: leaf node deletion leads to maximum iSPFgain and very few prefixes deletions

**Eliminated  
with  
BGP PIC Edge**

S(p): time to scan BGP table

Full VPNv4 walk(\*) and then IPv4 walk

B(p): time to compute bestpath for impacted routes and update FIB

(\*) CSCsm80316: 0ms for default VPNv4 NHT timer

(\*) Future: RD-based scoped walk upon NHT trigger for iBGP nhop invalidation

# IOS-XR BGP Prefix Independent Convergence



# Hierarchical Dataplane



# Hierarchical Dataplane FIB

## 1. FIB Leaf:

Z1/z1 (a BGP route)

## 2. BGP Path-List

list of best ECMP BGP nhops

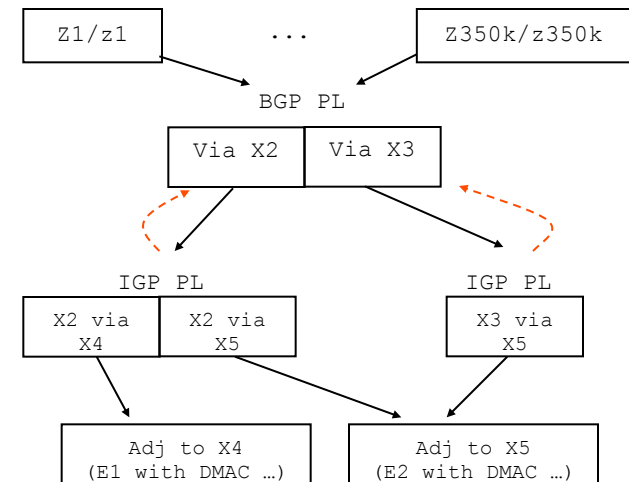
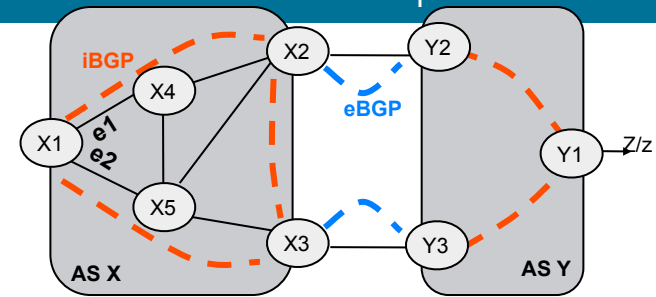
list of alternate BGP nhops

## 3. IGP Path-List

list of ECMP IGP paths

## 4. Adjacency

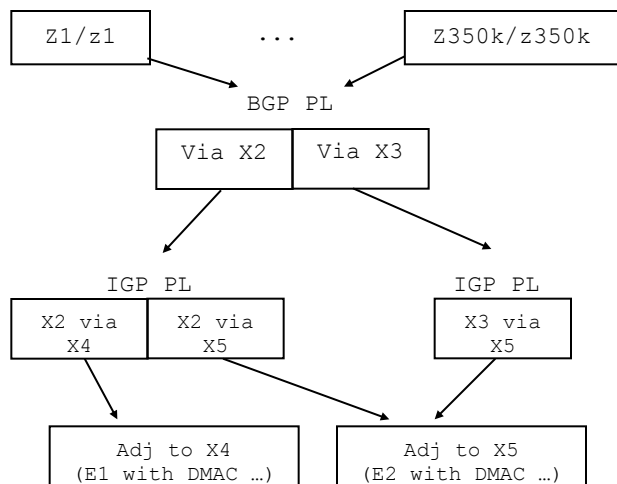
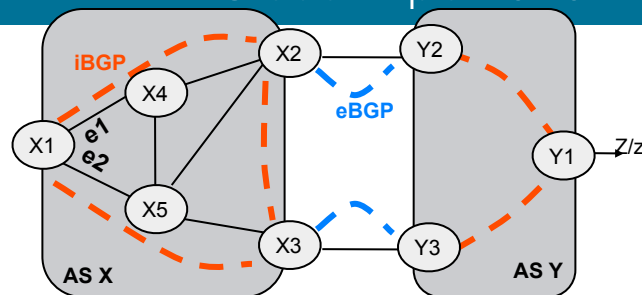
OIF and immediate nhop



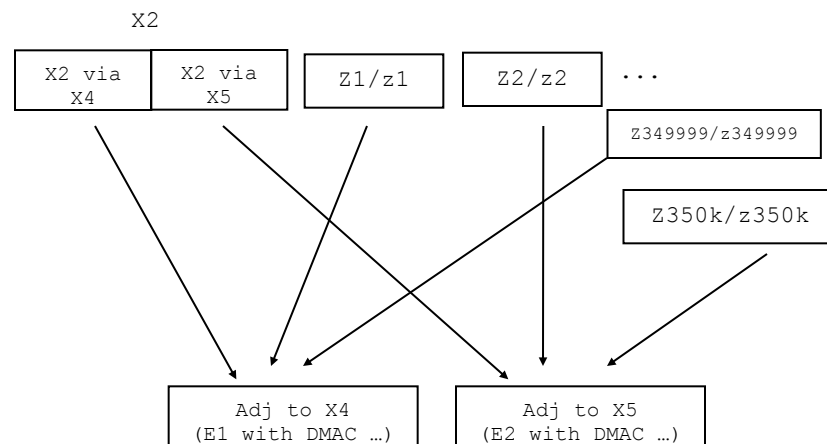
Example: X1 is configured for BGP Multipath. X1 has two ECMP paths to X2 and 1! IGP path to X3. The alternate list is empty.

Note: SW FIB maintains a linked list of the children BGP PL's on a per IGP PL (red dotted arrow)

# Hierarchical vs Flattened Dataplane

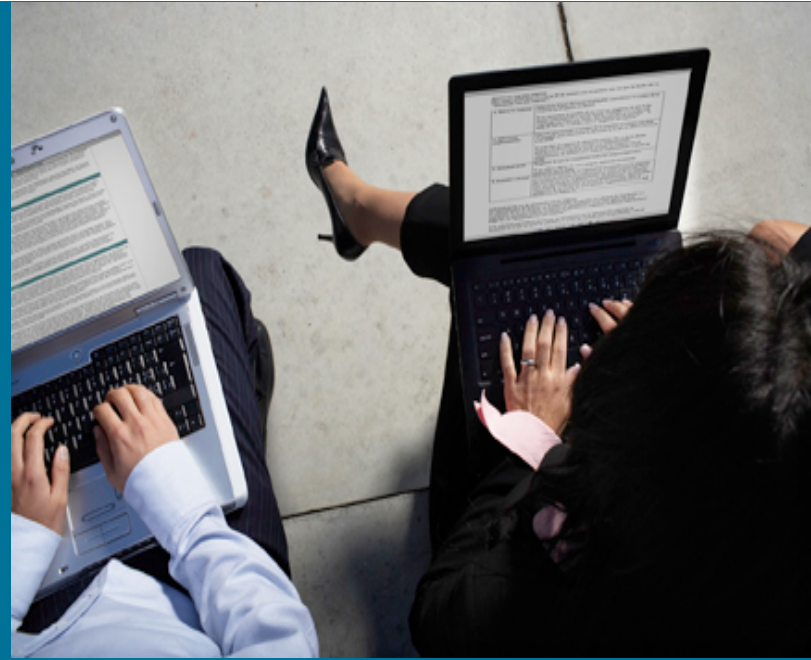


## Hierarchical Dataplane

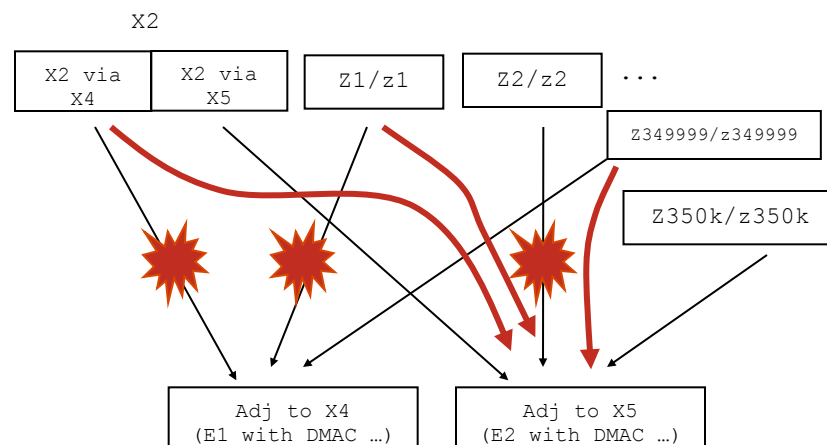
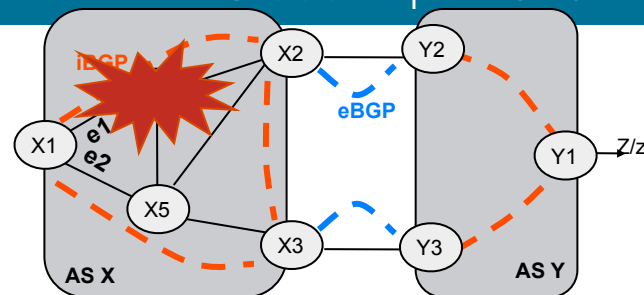
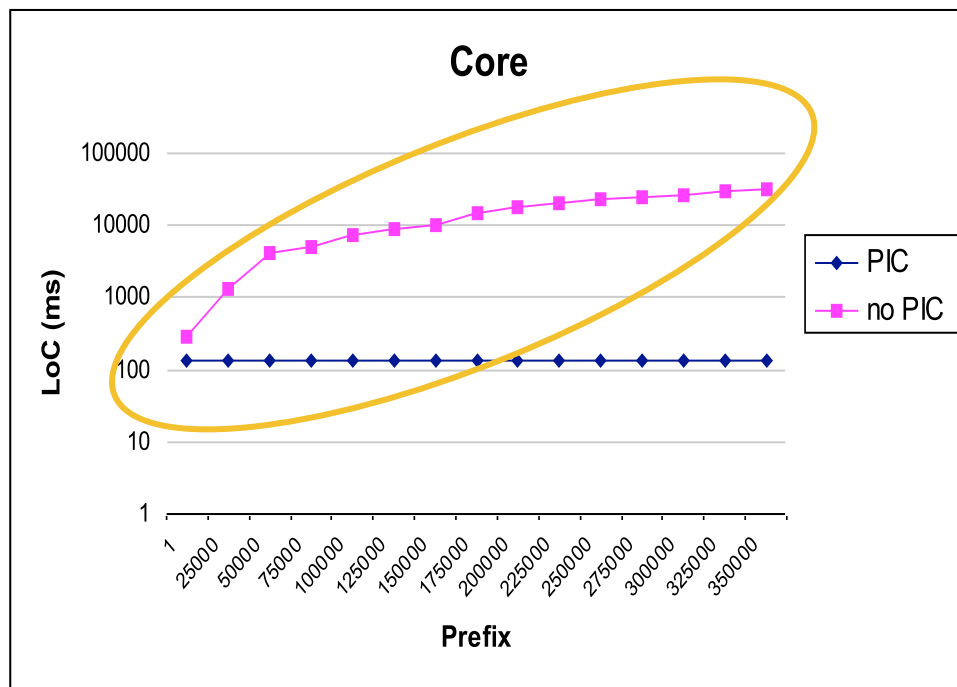


## Flattened Dataplane

# BGP PIC Core

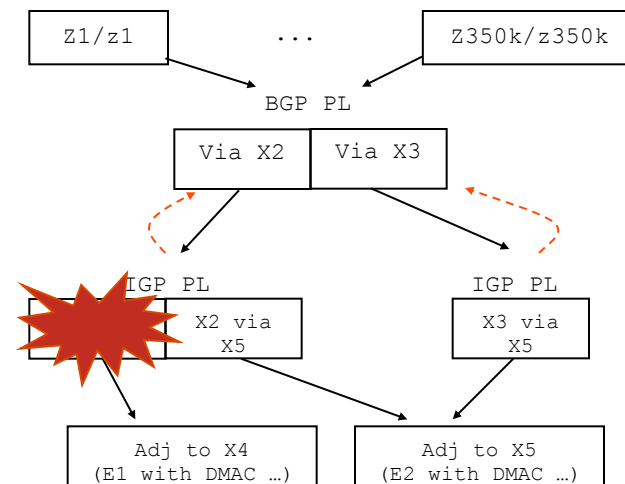
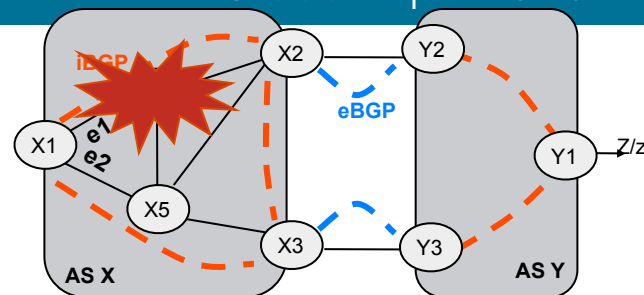
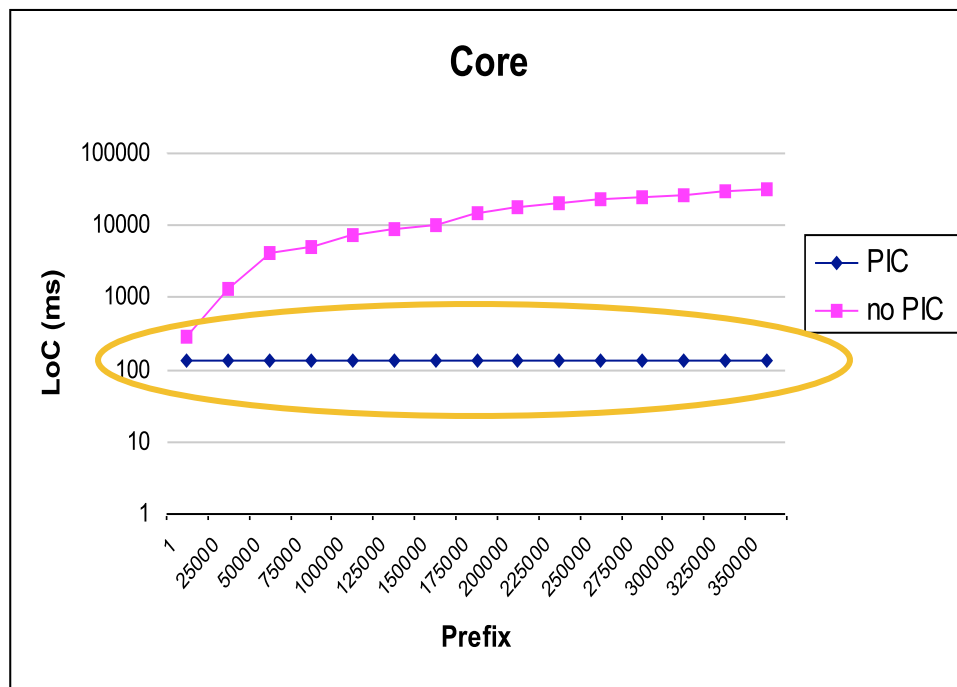


# Characterization Without BGP PIC Core



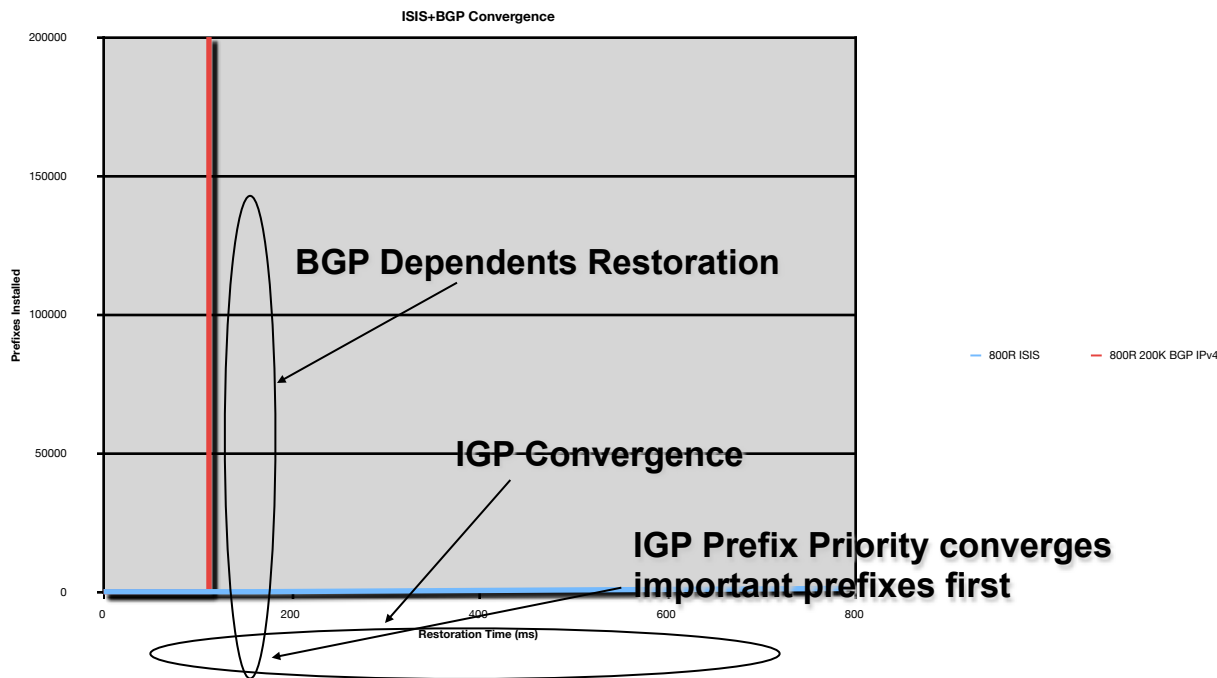
1. Without BGP PIC Core, the dataplane FIB is flattened. Upon IGP path change, ALL the BGP children entries need to be individually updated. The Loss of Connectivity scales with the BGP table size. This leads to slower convergence, non-optimum loadbalancing and worse robustness

# Characterization BGP PIC Core Analysis



1. As soon as IGP converges 0(200msec), the IGP PL memory is updated and hence all children BGP PL's leverage the new path immediately
2. Optimum convergence, Optimum Load-Balancing, Excellent Robustness

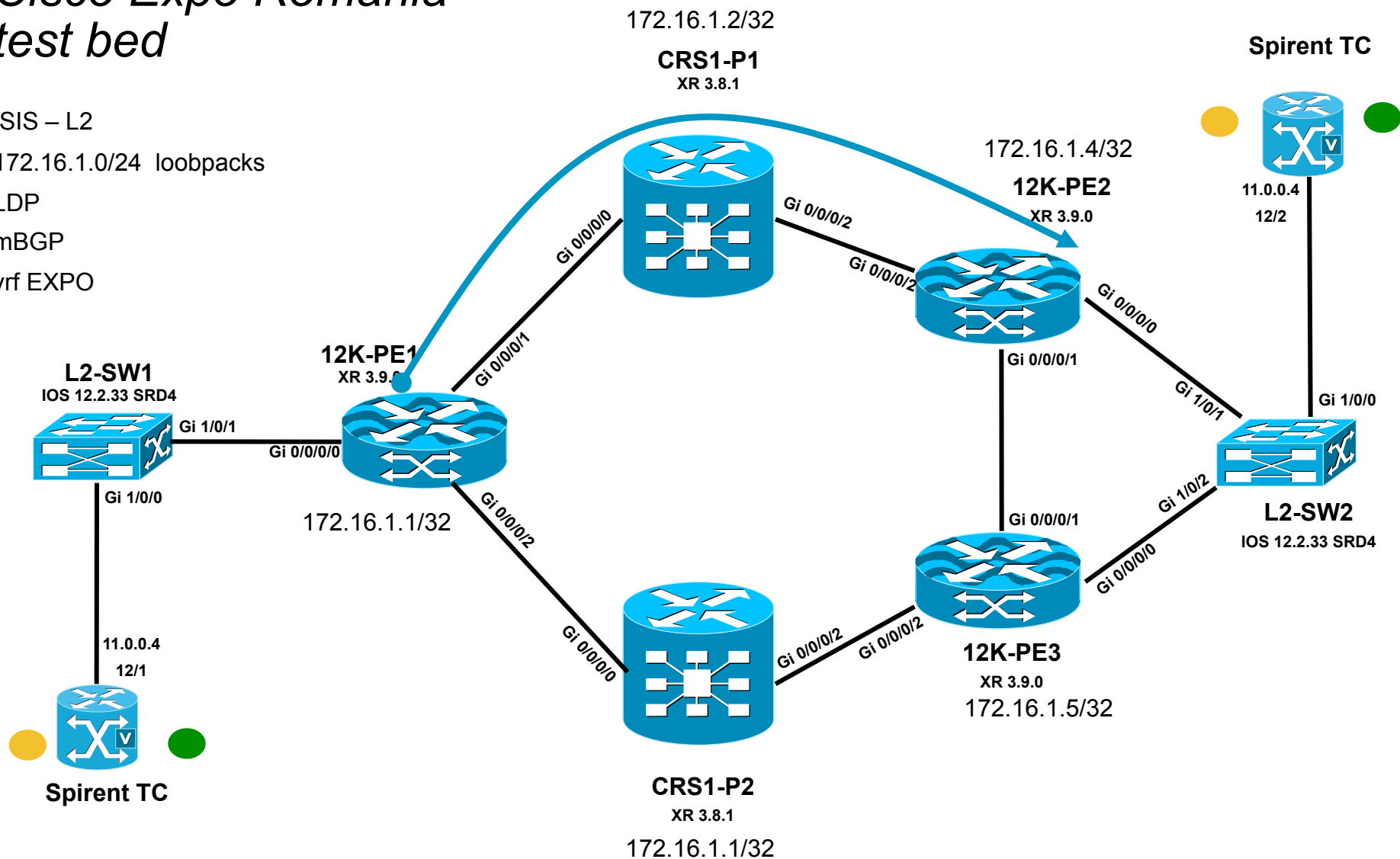
# Target Behavior for all network services



1. Testbed: 809 Nodes, 2350 ISIS prefixes, 200k BGP dependents to impacted BGP nhop
2. When ISIS converges, all the BGP dependents immediately leverage the ISIS convergence (Prefix Priority for BGP NH)

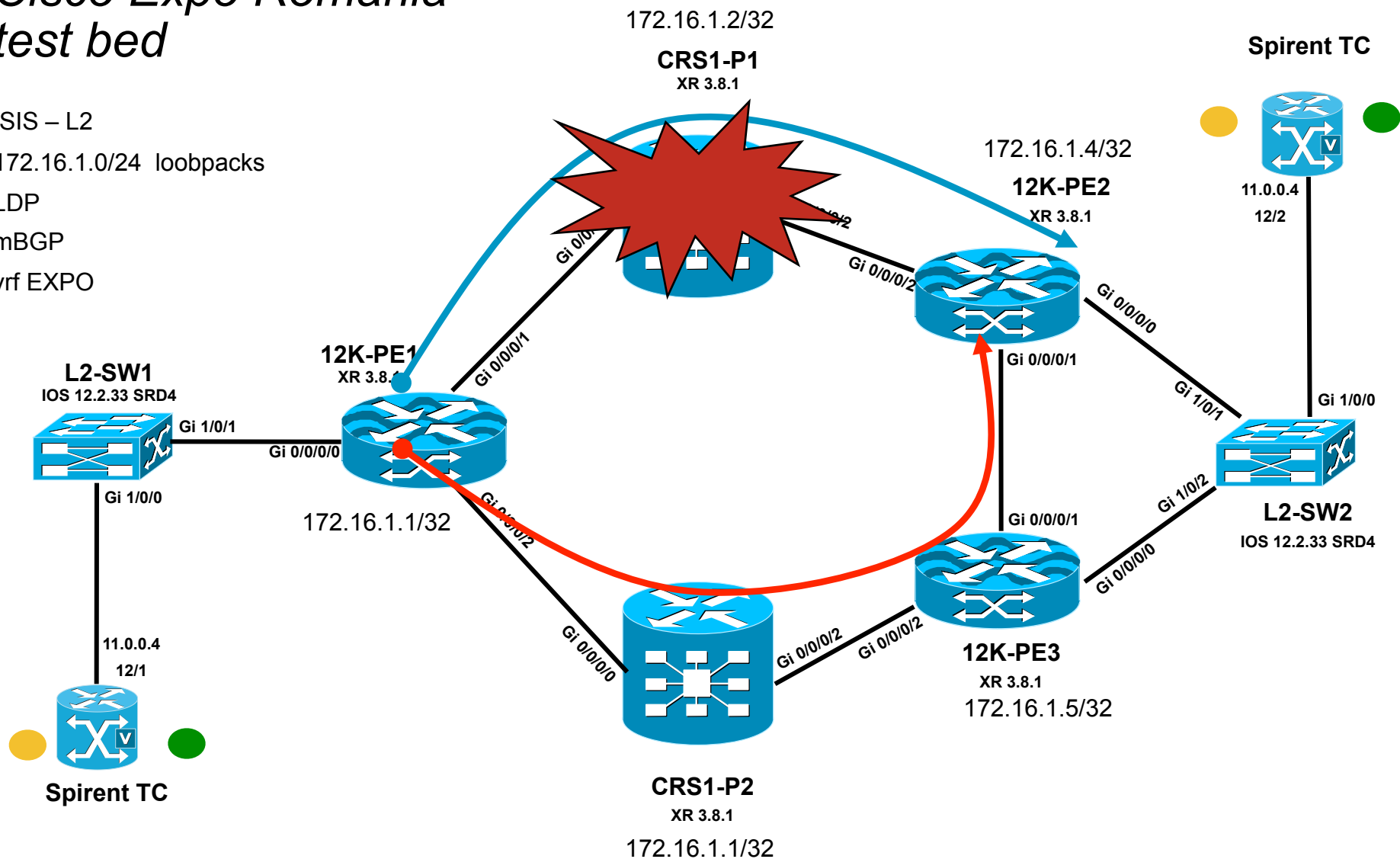
# Cisco Expo Romania test bed

ISIS – L2  
172.16.1.0/24 loopbacks  
LDP  
mBGP  
vrf EXPO



# Cisco Expo Romania test bed

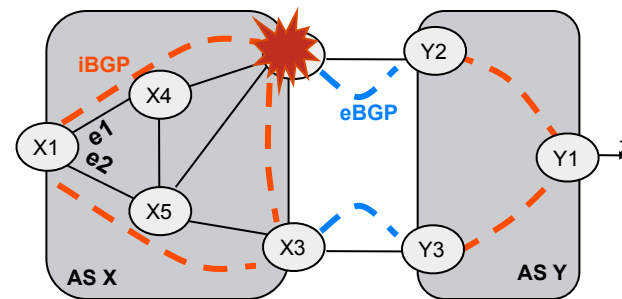
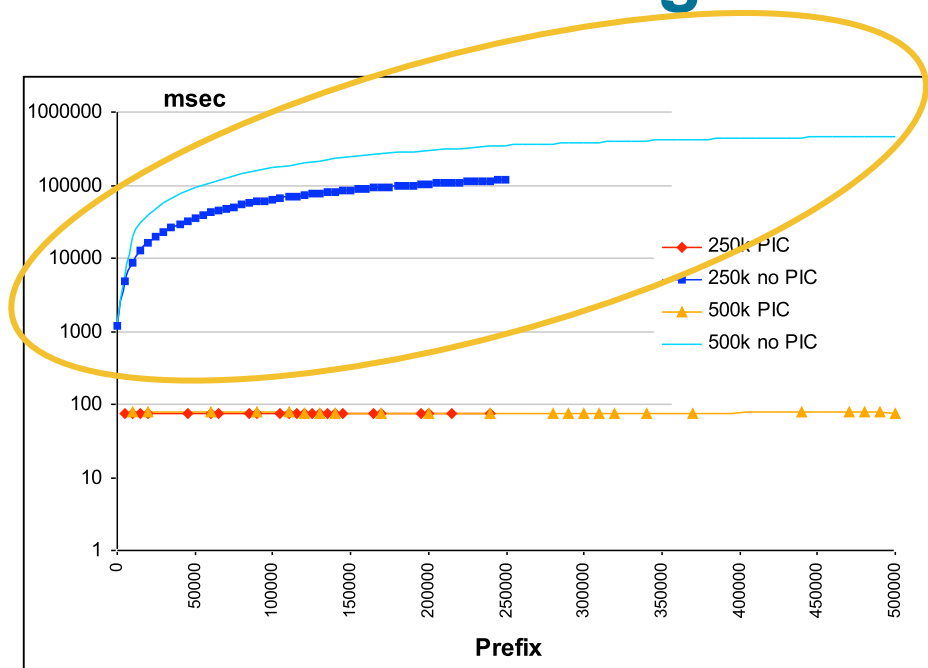
ISIS – L2  
172.16.1.0/24 loopbacks  
LDP  
mBGP  
vrf EXPO



# BGP PIC Edge

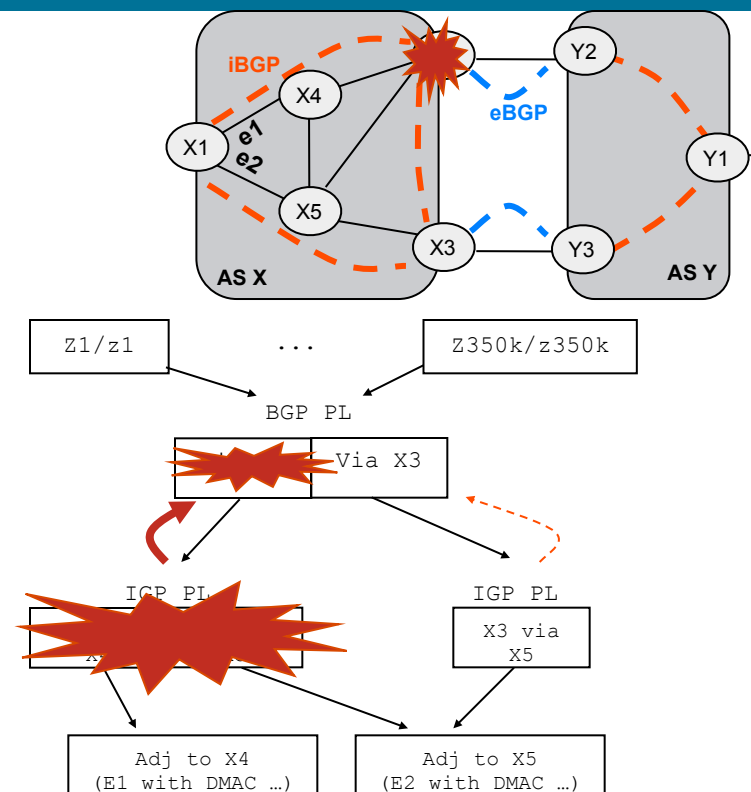
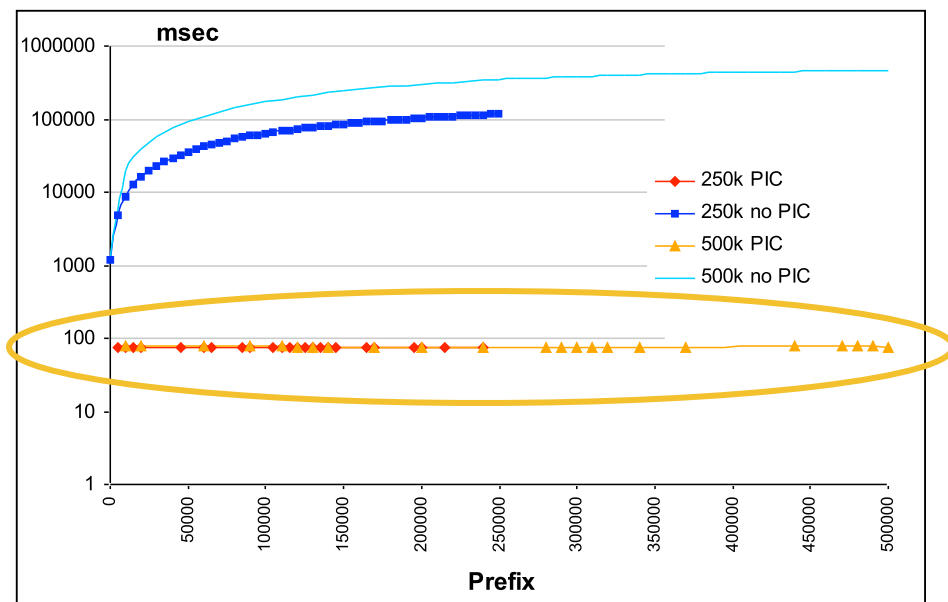


# Characterization Without BGP PIC Edge



1. At IGP Convergence time, in a flattened dataplane FIB, all the BGP entries recursing via X2 point to an invalid path. No dataplane protection is possible.
2. The control plane convergence is now required to move each BGP entry onto an alternate next-hop and then update the flattened dataplane FIB accordingly. This may take minutes.

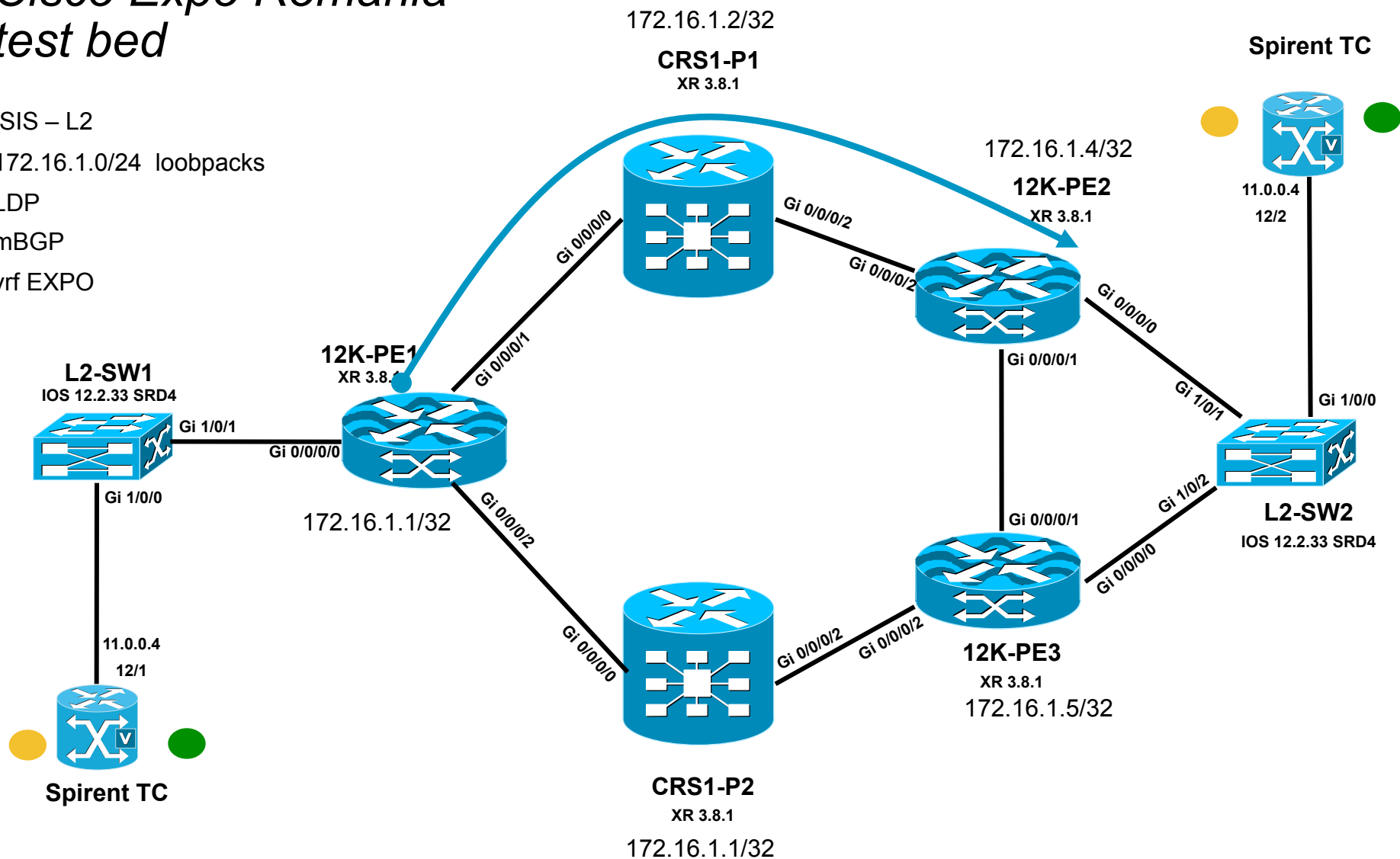
# Characterization BGP PIC Edge



1. At IGP Convergence time, the complete IGP PL to X2 is deleted. SW FIB walks the linked list of parent BGP PL and in-place modify them to use alternate ECMP best nhops or enable alternate next-best nhops. This is quick because the BGP PL sharing is efficient.
2. The control plane convergence still occurs in the background (blue curve) but its slowness does not impact dataplane connectivity and hence the T-SLA experience

# Cisco Expo Romania test bed

ISIS – L2  
172.16.1.0/24 loopbacks  
LDP  
mBGP  
vrf EXPO



# Cisco Expo Romania test bed

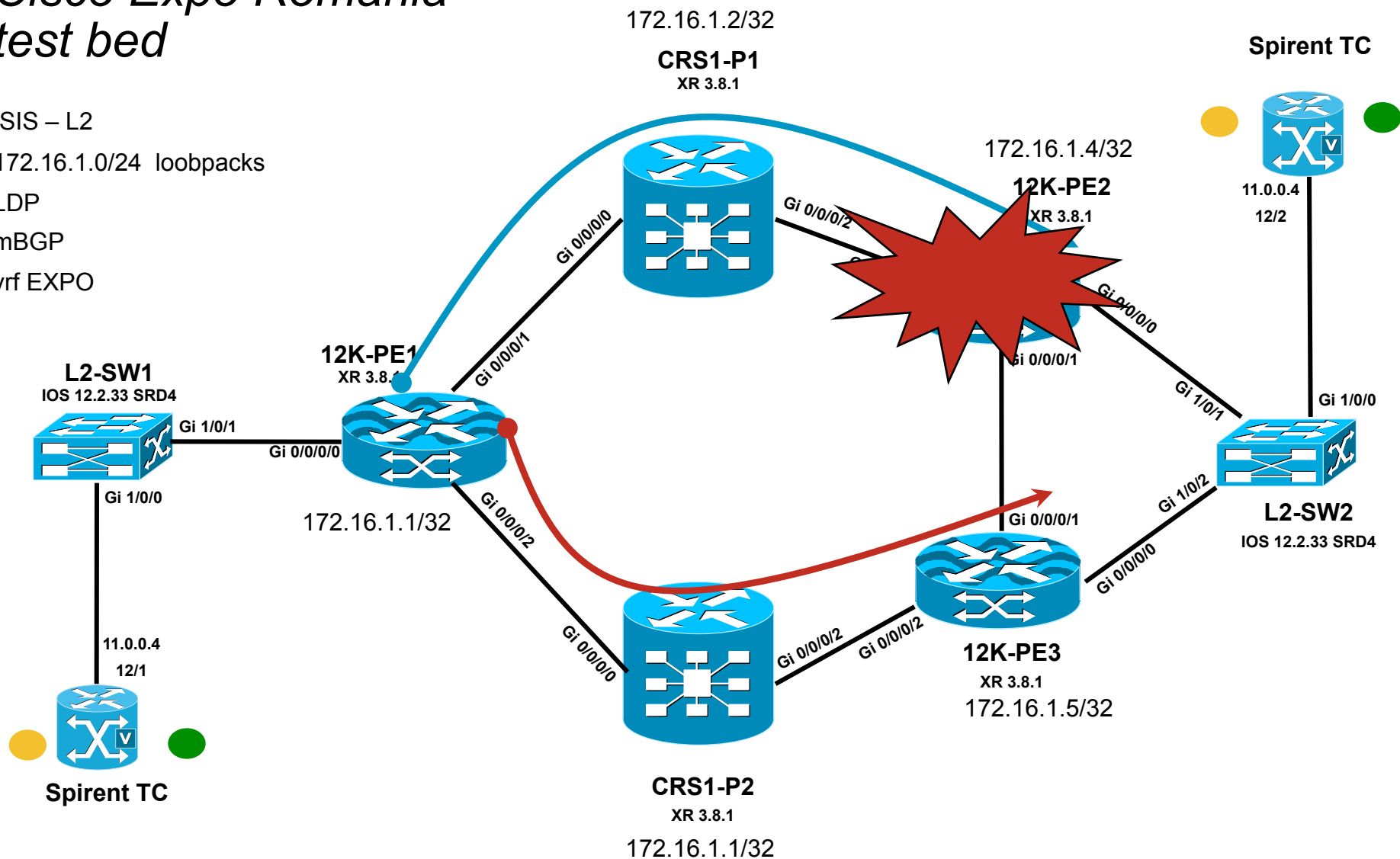
ISIS – L2

172.16.1.0/24 loopbacks

LDP

mBGP

vrf EXPO



# Conclusion



# Simplicity

1. IGP, BGP are meant to route around failure
2. 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
3. Simplicity = Cost Savings for the designer/operator