Architectural Best Practices for Ensuring High Availability

In IOS XR IP/MPLS Backbone Networks

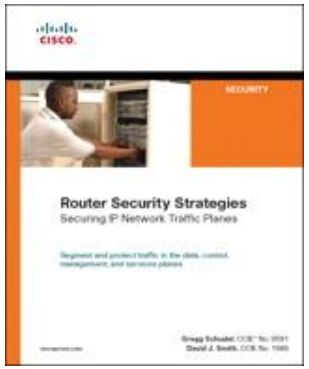
David J. Smith
Distinguished Solutions Engineer

Lokesh Khanna
Solutions Engineer

BRKSPG-2080

About Us

- David J. Smith
  - Live in the New York City area
  - Joined Cisco in August 1995
  - SE supporting Service Providers in the Americas
  - Contact: djsmith@cisco.com
- Lokesh Khanna
  - Live in the Denver, Colorado area
  - Joined Cisco in 2015
  - SE supporting Service Providers & Hyperscalers
  - Contact: lokhanna@cisco.com



Cisco Webex App

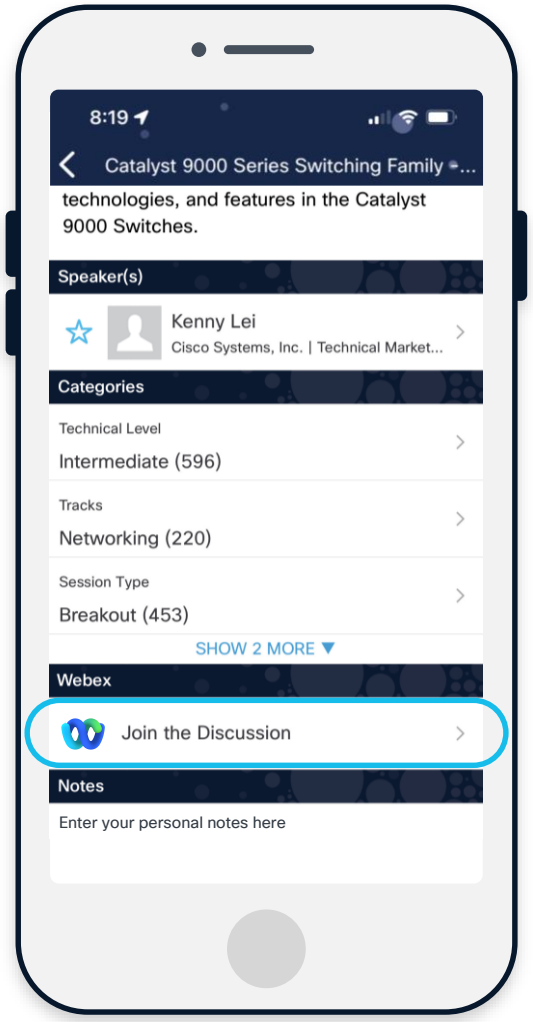
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 13, 2025.



Agenda

01 Network Redundancy

02 Control Plane Best Practices

03 Forwarding Plane Best Practices

04 Network Simplification

Out of Scope

- Operational best practices

Refer to BRKSPG-2695:
Resilient Networks: From Prevention to Recovery

Network Redundancy

Role of Network Redundancy

- To eliminate single points of failure
- Multiple layers of redundancy are often implemented:
  - Redundant links, nodes, paths and facilities
  - Redundant power and cooling systems
  - Redundancy in each network layer: IP and Optical
  - Redundancy in both the forwarding and control planes
  - Redundant links and paths should use separate conduits
  - If some fibers must use the same conduit, SRLGs may help

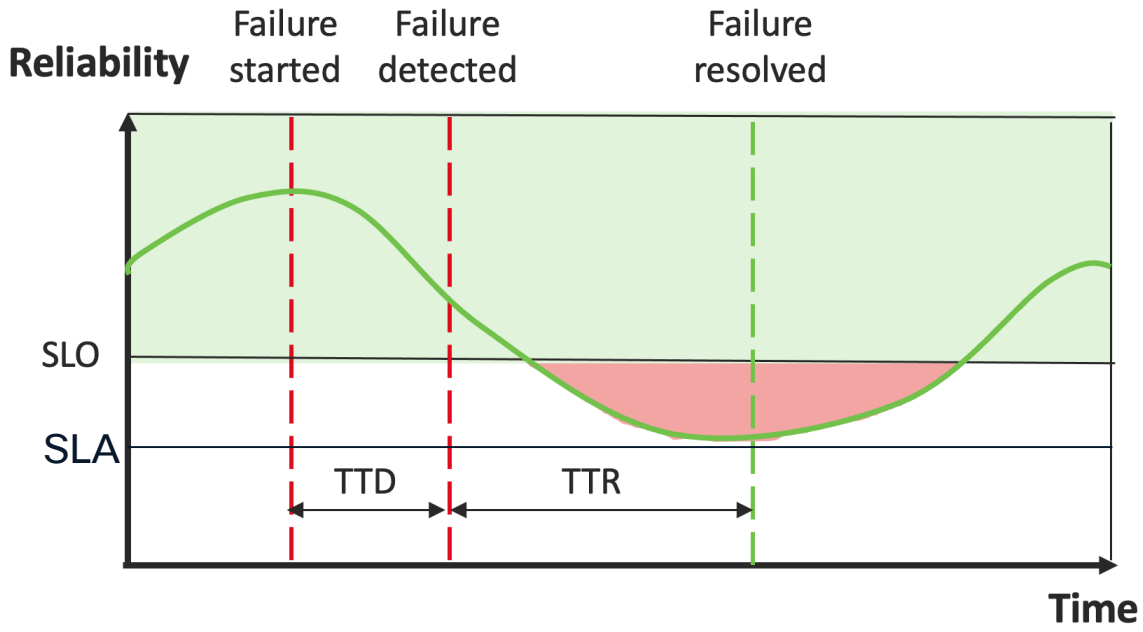
Why do network outages happen despite redundancy?

Network Redundancy Considerations

- Redundancy is highly effective when components fail hard and fast
- However, when a component malfunctions without completely failing, the consequences can be severe, even with redundancy

Goals for Network Redundancy

- Ensure the network consistently operates with reliability that exceeds the specified SLOs and SLAs
- Mitigate the effects of failures and malfunctions due to:
  - Hardware
  - Software
  - Protocols
  - Out of Resource (OOR) conditions
  - Configuration errors
  - Security attacks
  - Maintenance activities
  - Environmental factors



Why do relatively minor issues cause major outages?

Concept of the Critical State

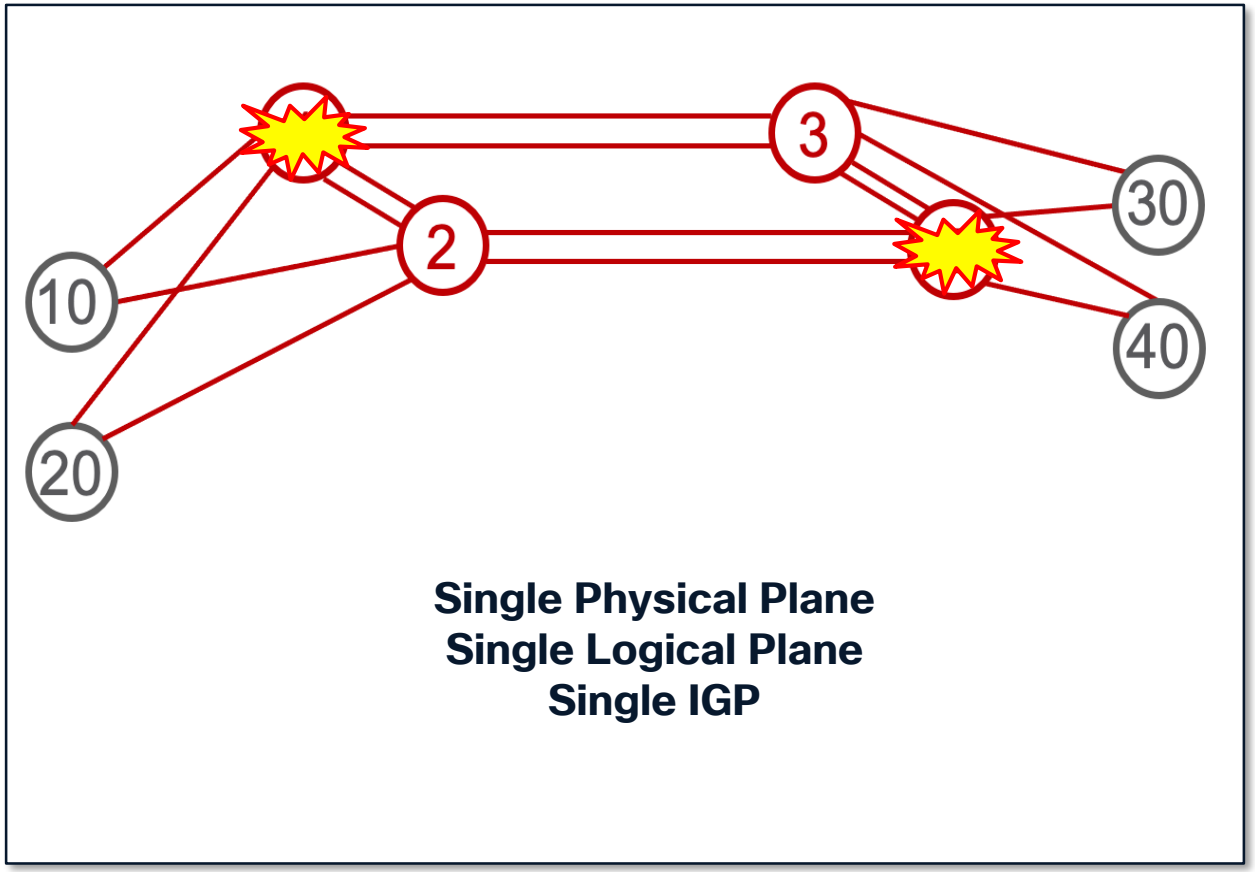
- A complex system is a sum of its parts and their interactions
- Applies the sand pile model, where a single dropped grain of sand can be harmless, or trigger an avalanche
- Failure results from the critical state of the sand pile that was built, not from a single grain of sand
- Considers the culmination of steps that led to the failure event
- Asymmetric relationship between cause and effect:
  - Relatively small causes can produce very large effects

Avoid the Critical State

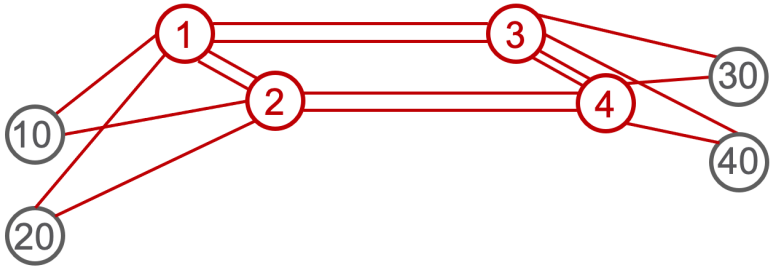
Architectural Principles for Highly Available IP/MPLS Networks

- Network redundancy
- Control plane best practices
- Forwarding plane best practices
- Network simplification

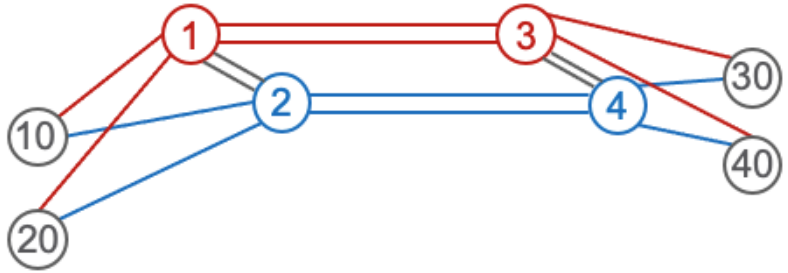
Network Redundancy Considerations



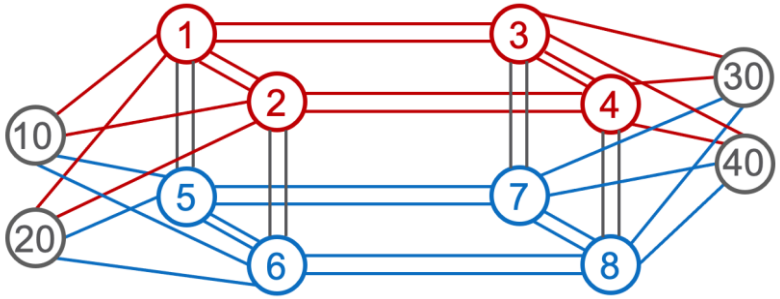
Multi-Plane Network Redundancy



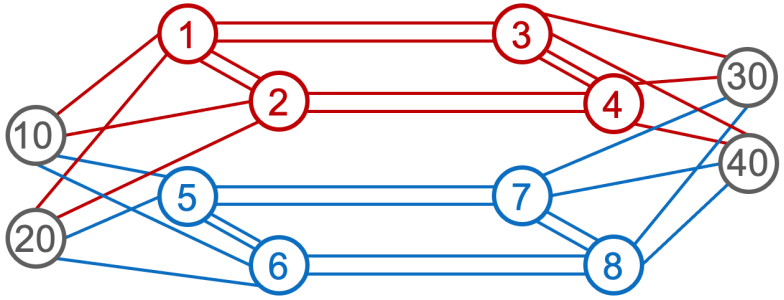
1 **Single Physical Plane**
Single Logical Plane
Single IGP



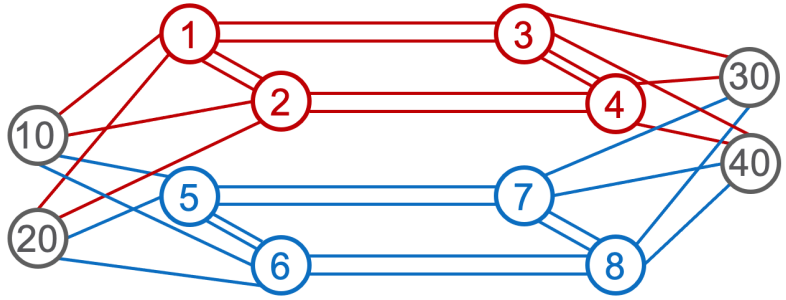
2 **Single Physical Plane**
Two Logical Planes
Single IGP



3 **Two Physical Planes**
Two Logical Planes
Single IGP; Possible fallback



4 **Two Physical Planes**
Two Logical Planes
Single IGP; Full isolation

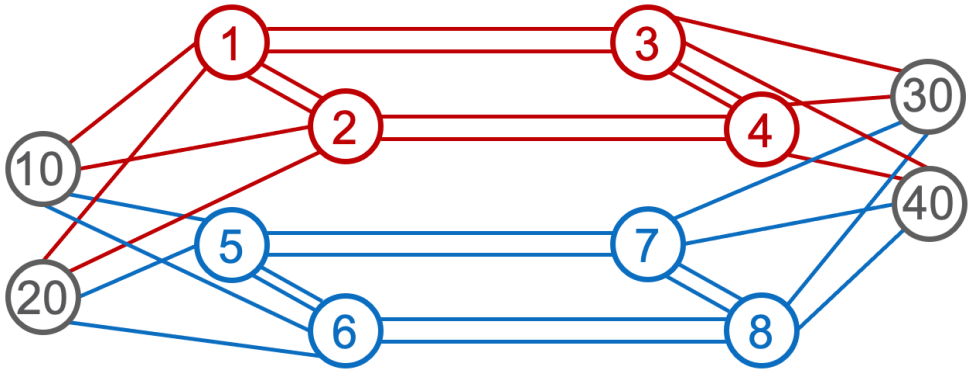


5 **Two Physical Planes**
Two Logical Planes
Two IGPs (one per plane)

Multi-Plane Traffic Steering

- Deployment models for multi-plane architectures:
  - Active / Backup
  - Active / Active load balancing
  - Service-based routing (e.g., VPN vs. Internet, Mobility vs. Wireline)
  - Secure (e.g., MACsec encrypted) versus Unsecured circuits
- Traffic steering options:
  - IGP costs/metrics
  - BGP policies
  - MPLS/RSVP-TE tunnels
  - SR-TE policies
  - SR Flexible Algorithms

} Recommended



Node Redundancy Considerations

	Distributed	Centralized	Fixed
HW redundancy	Full redundancy (RP, fabric, power, fan)	Full redundancy (RP, fabric, power, fan)	Partial redundancy (Power and fan only)
Scaling	Scales vertically; Facilitates BW scaling	Scales horizontally; Facilitates service scaling	Scales horizontally; Facilitates service scaling
Blast radius	Large	Small	Small

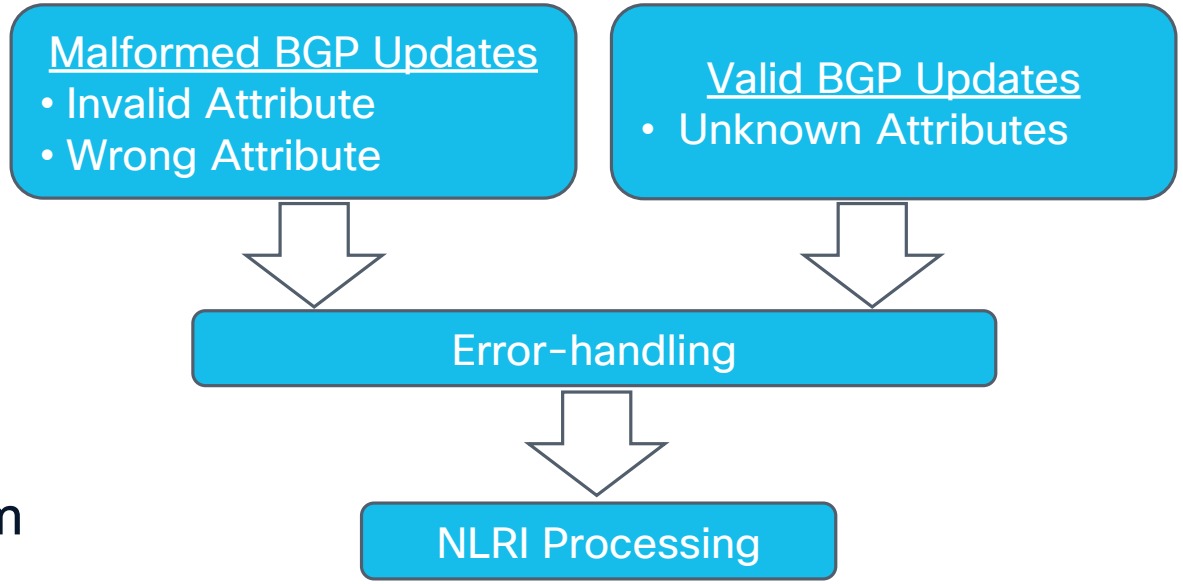
- Each support IP control, forwarding, and management plane best practices for network resiliency, except no NSR on fixed routers
- Hybrid deployment may be the most optimal in terms of availability, scaling and cost

IP Control Plane Best Practices

BGP Considerations

- Avoid BGP redistribution into IGP
  - If not done correctly, it may cause IGP failure or routing loops
- IP/MPLS core routers should not carry MP-BGP service routes
  - No advantage participating in the Internet BGP control plane
  - BGP-free core recommended
  - Service traffic should be forwarded via classic MPLS, SR-MPLS or SRv6 (latter recommended)

BGP Error Handling



- Basic error handling is enabled by default, and it prevents BGP session reset if a malformed update is received
- Treats a malformed update as withdraw, which removes the prefixes with malformed attribute from BGP table
- Cisco XR supports attribute filtering which allows removing the malformed attribute instead of withdrawing the prefixes

```
router bgp 65530
  attribute-filter group BGP-ATTRIBUTE-FILTERING
  attribute range 28 discard
  !
  neighbor 10.101.203.203
    update in filtering attribute-filter BGP-ATTRIBUTE-FILTERING
!
```

Extended BGP Error Handling

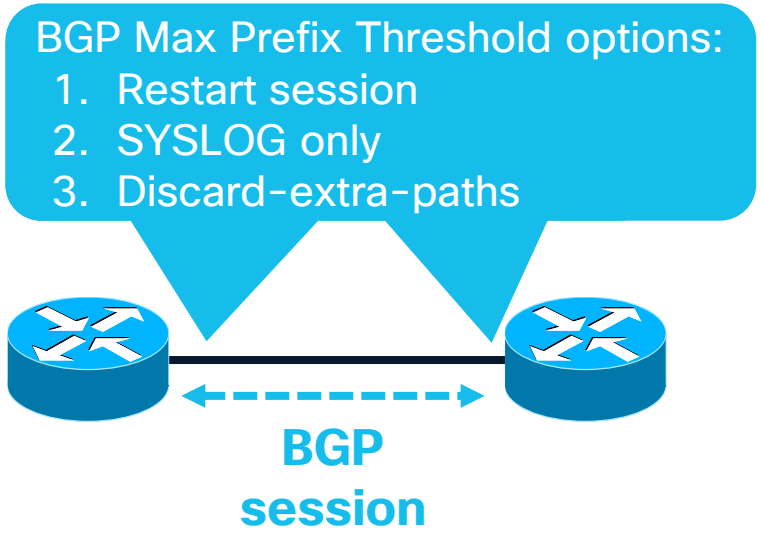
- Cisco added extended error handling capabilities to avoid session reset under following conditions (as per RFC 7606)
  - NEXT_HOP length is not 4
  - MP_REACH Nexthop length is 0
  - MP_REACH Nexthop length is invalid for the MP AFI/SAFI
  - Duplicate Non-Optional Transitive attributes
- Recommend: “Extended” revised error handling in IOS XR requires the following:

```
(config-bgp)# update in error-handling extended ebgp
(config-bgp)# update in error-handling extended ibgp
```

- Modern BGP routers should comply with RFC 7606

BGP Peer Maximum Prefixes

- Controls how many BGP prefixes can be received from a neighbor
  - Protects against a customer or ISP/CDN peer leaking full Internet routes back to the SP
- Prior to IOS XR 7.3.1 max peer prefixes per address family were enabled by default for both E-BGP and I-BGP and, if any limits were hit, the session would be taken down
- As of IOS XR 7.3.1 the default max limits were removed
- Recommend: Configure E-BGP max peer limits per address family based on the expected scale, but discard any extra paths instead of terminating the session
  - Also configure a SYSLOG warning threshold per address family



BGP Session Authentication

TCP Authentication Option (AO)

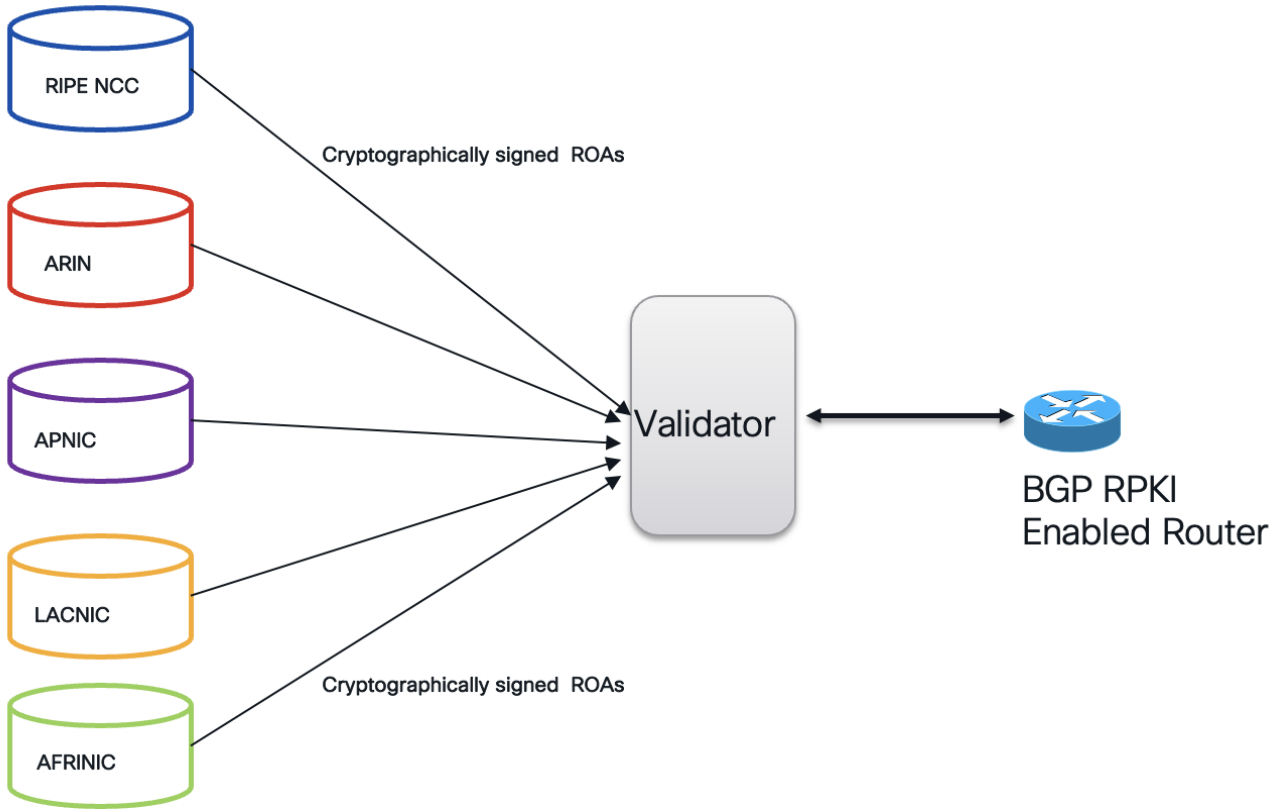
- TCP Authentication Option (TCP-AO), as defined in RFC 5925, is the proposed replacement for TCP MD5 Signature Option (TCP MD5) (RFC 2385)

Features	TCP MD5	TCP-AO
Authentication	MD5 Hash	MACs (HMAC-SHA, AES-CMAC)
Hash Algorithm	Fixed (MD5)	Flexible (multiple algorithms)
Key Management	Requires session reset	Dynamic key changes
Key Change	Disconnects connection	Maintains connection

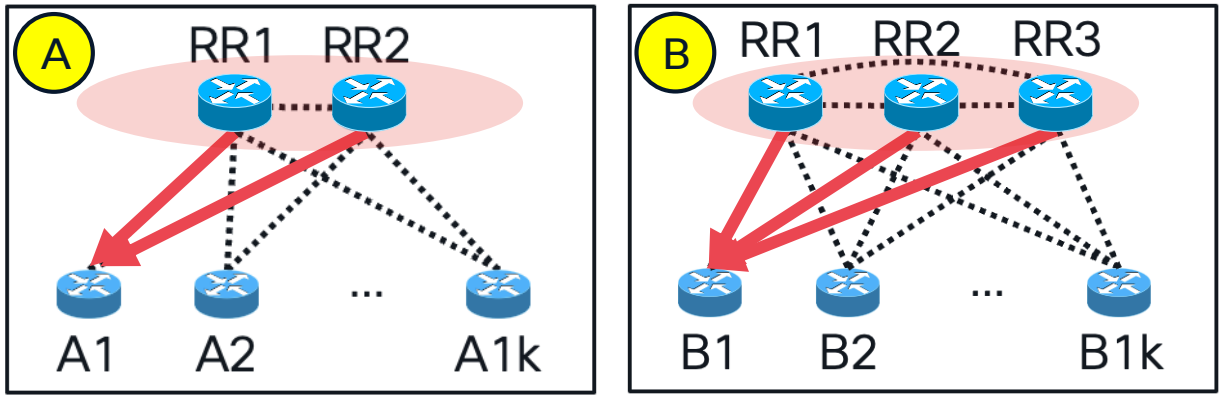
- Recommend: TCP-AO is a more secure and flexible alternative to TCP MD5, addressing the limitations of the older protocol and providing better support for modern security requirements and dynamic key management

BGP Prefix Origin Validation Based on RPKI

- RPKI is a certificate-based, global database that maps BGP prefixes to their authorized origin-AS numbers
- BGP routers connect to an RPKI validator to verify the origin-AS of BGP prefix advertisements received
- Routes considered 'Invalid' are not considered for BGP best path (default)
- Recommended to reduce the risk of:
  - BGP prefix hijacking
  - BGP prefix mis-announcements



BGP Route Reflectors



- RRs simplify BGP control plane provisioning and scaling in large-scale BGP networks
- RR redundancy eliminates single points of failure, however, excessive RR redundancy can be detrimental
  - Increases # of I-BGP sessions, I/O packet processing and the # of BGP paths
  - May delay BGP convergence given RR clients cannot remove an I-BGP route until a WITHDRAWN is received from each RR
- Recommendation: avoid more than three (3) RRs in a cluster
- Recommendation: maintain smaller RR cluster sizes to reduce the blast radius of cluster failures

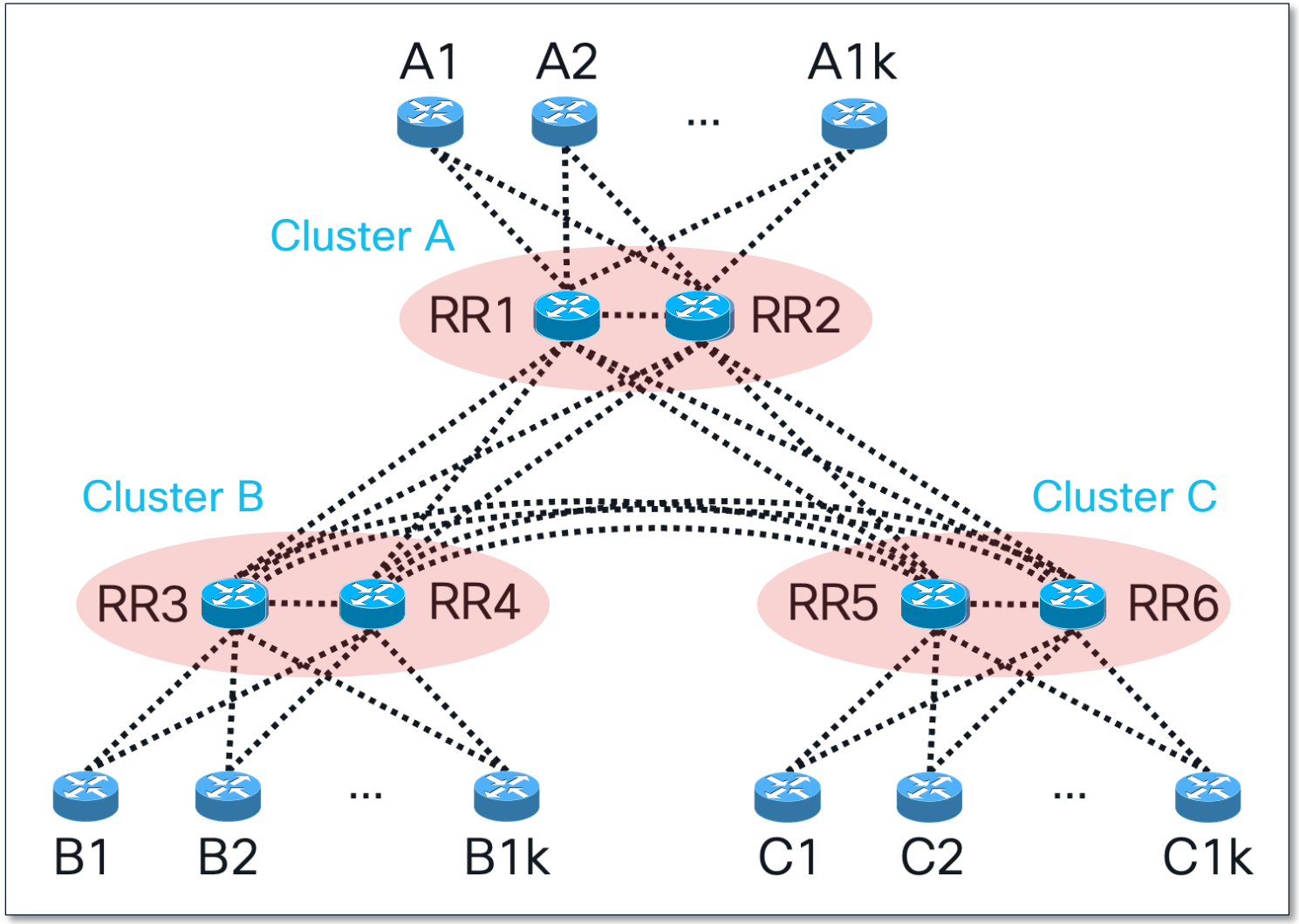
BGP RR Designs with Multiple Clusters

- Large-scale BGP networks may require multiple RR clusters
- In general, there are three (3) main variations of multi-cluster BGP RR designs:
  - Full-mesh
  - Multi-tier
  - Multi-plane

BGP RR Design with Multiple Clusters

Full Mesh

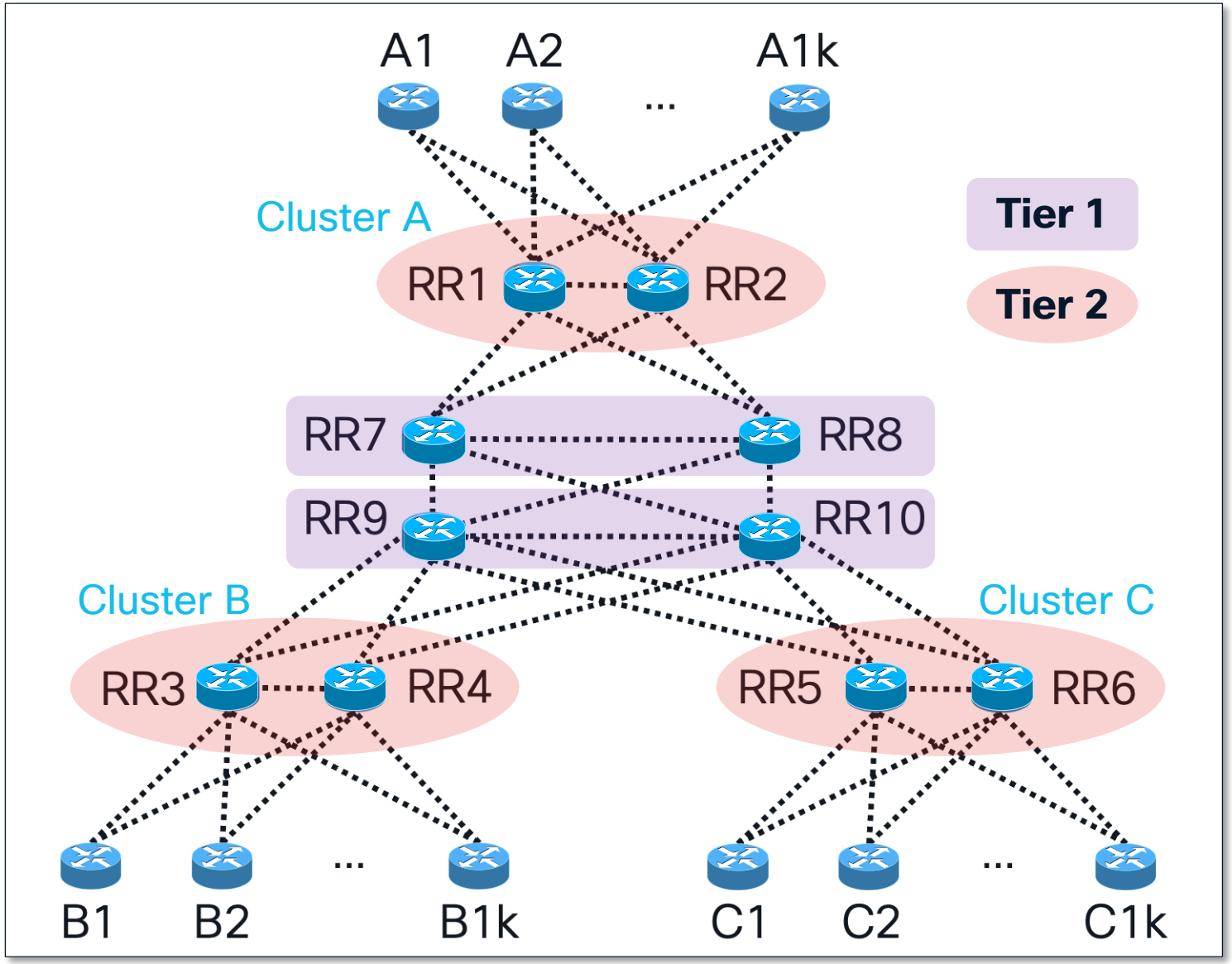
····· I-BGP



- Limiting factor is BGP RR I-BGP provisioning and scale
- Adding a new RR requires adding I-BGP sessions on all existing RRs

BGP RR Design with Multiple Clusters

Multi-Tier



····· I-BGP

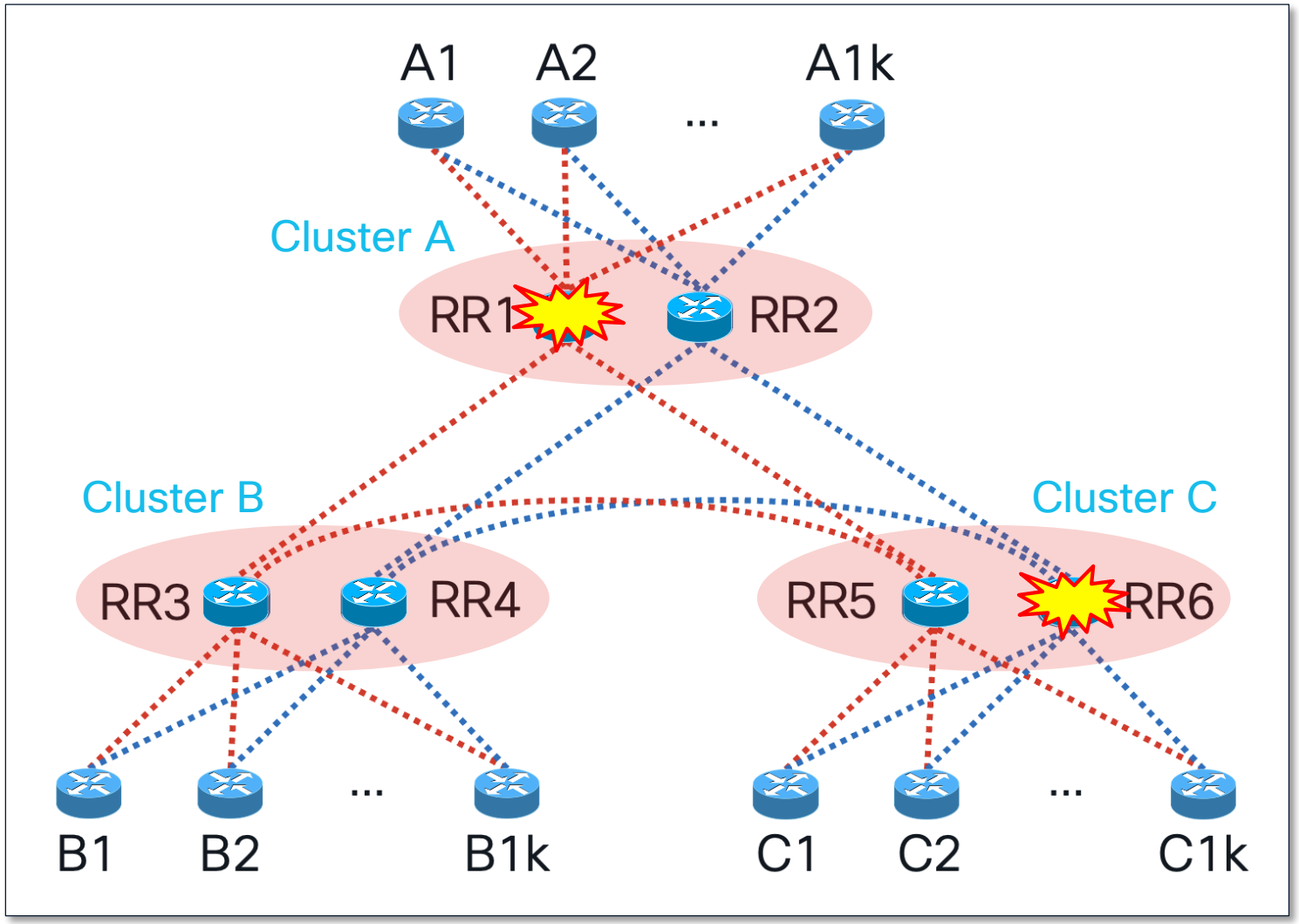
- Simplifies RR I-BGP provisioning
- Additional RR tier may degrade BGP convergence
- Tier 1 cluster failures have a large blast radius

BGP RR Design with Multiple Clusters

Multi-Plane

I-BGP (red)

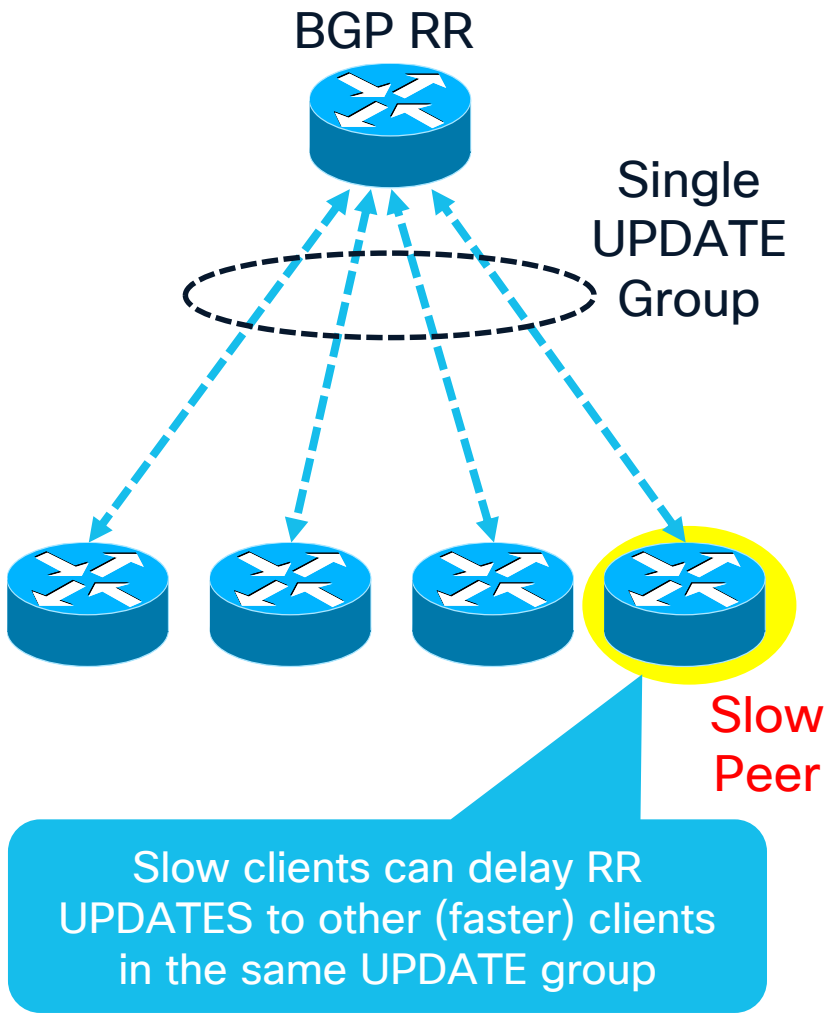
I-BGP (blue)



- Reduces I-BGP provisioning per RR
- Eases transition during RR software upgrades
- 3rd plane recommended to protect against double failure scenarios

BGP RR Slow Peers

- BGP RR clients that are slow may adversely affect other clients within the same UPDATE group
  - May delay BGP RR UPDATES to faster clients
- IOS XR 7.3.1 introduced BGP slow peer automatic isolation
  - Enabled by default, however, use of this version is not recommended
- IOS XR 7.9.1 introduced an updated version of slow peer automatic isolation
  - Use of this capability and version is recommended
  - Disabled by default
- Alternatively, explicitly configure BGP RR clients that are 'permanently slow' in their own UPDATE group separate from fast clients




```
%BGP-5-SLOWPEER_DETECT: Neighbor IPv4 Unicast 10.1.6.7 has been detected as a slow peer
```

Other BGP RR Considerations

- Recommend the use of separate BGP RRs per address family for increased fault isolation and scaling:
  - BGP Internet service routes (IPv4 unicast, IPv6 unicast, 6PE)
  - BGP VPN service routes (VPNv4, VPNv6, EVPN)
  - BGP transport routes (i.e., BGP-LU aka IPv4 labelled unicast)
  - BGP-LS for IGP topology export
- BGP RRs often hide paths, which can lead to suboptimal routing and, in specific configurations, route oscillations per RFC 3345
  - BGP Add Paths group best recommended to prevent this
  - Alternatively, BGP ORR (Optimal Route Reflection)
- BGP Route Target Constrain (RTC) can dramatically reduce the number of L3VPN and/or EVPN route updates to PE nodes

Other BGP Best Practices (1)

Reference

- Static configuration of router ID using loopback address to prevent changes to the router ID and consequent flapping of BGP sessions
- Enable TCP Path MTU discovery to enable use of the largest packet size that does not require fragmentation anywhere along the path between two BGP peers
- Recommend configuring a BGP table-policy on BGP IPv4/IPv6 RRs deployed outside of the forwarding plane to avoid unnecessary installation of BGP Internet routes in the FIB
- E-BGP route policies to restrict routes accepted from and advertised to E-BGP neighbors (e.g., bogons, more specifics, infrastructure routes)
- Delete inbound communities and extended communities, especially if doing VRF peering; some vendors may accept routes with an RT set from an eBGP neighbor

Other BGP Best Practices (2)

Reference

- Limit E-BGP peering to only explicitly configured neighbors
  - Restrict the number of dynamic BGP sessions on routers under your administration
- E-BGP TTL security (i.e., RFC 3682 GTSM) to help protect against remote BGP attacks
- AS-PATH limits to filter prefixes with an AS-PATH length greater than a specific value (e.g., 50)
- IETF BCP 194 (RFC 7454: BGP Operations and Security) if providing Internet service
- eBGP Route Flap Damping (RFD) can be considered to suppress Internet BGP churn (see http://rfd.rg.net/)
- BGP Best External to advertise the best-external path to I-BGP peers, when a locally selected best path is from an I-BGP peer – may enable faster restoration of connectivity (i.e., BGP PIC Edge)

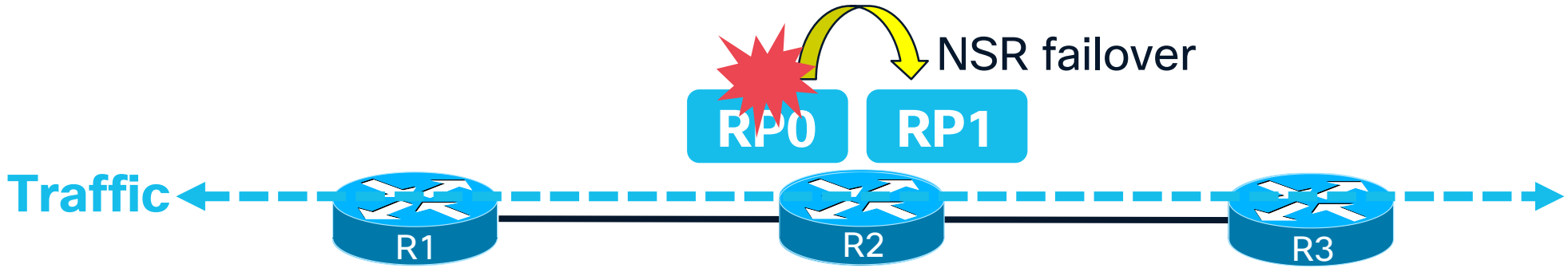
Other BGP Best Practices (3)

Reference

- BGP Flowspec for rapid, intra-domain, distributed attack mitigation
  - Take caution not to inadvertently filter eBGP sessions with flow specifications
- Be aware that different vendors use different default RIB administrative distances and, therefore, have different preferences for IGP routes versus eBGP routes
  - In multi-vendor environments, RIB admin distances should align to avoid routing loops
- BGP update wait-install to postpone advertising UPDATES until the RIB confirms that BGP routes have been installed

Non-Stop Routing (NSR)

- Enables lossless traffic forwarding during an RP failover
- Backup RP synchronizes and preserves the routing state, which includes protocol sessions and routing process data (e.g., IGP LSDB, BGP table)
- During RP failover, the backup RP is used to maintain control plane sessions and traffic forwarding without interruption
- Peer routers are unaware of such events – no protocol signaling required with peers, however, backup RP is required
- Standby RP must be in “Ready” state for RP failover to work

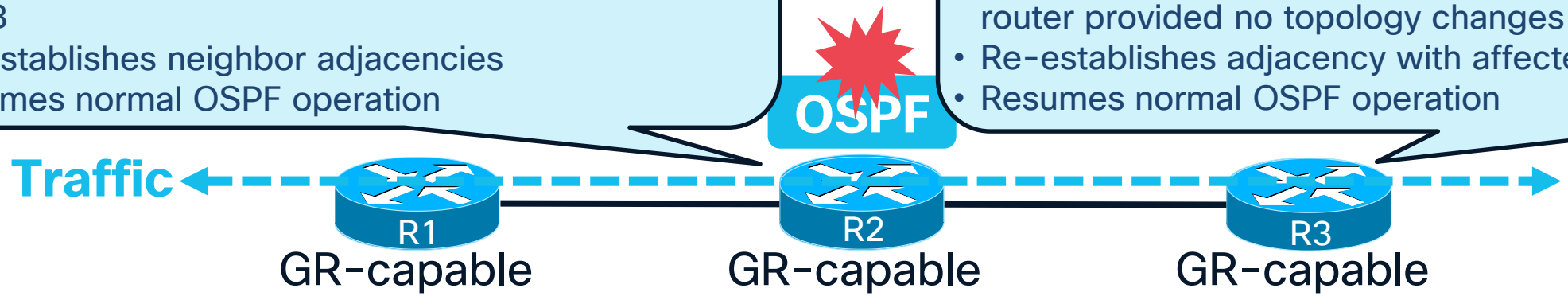


Non-Stop Forwarding (NSF) with Graceful Restart

- OSPF (RFC 3623), IS-IS (RFC 8706), LDP (RFC 3478), BGP (RFC 4724)
- Enables a routing process to restart without traffic loss
- Redundant RP is not required. However, neighbors must be GR-enabled

Graceful Restart procedures on affected router:
- Originates grace-LSAs (prior to restart)
- Starts grace period (lifetime) timer
- Restarts OSPF process while preserving OSPF routes in FIB
- Re-establishes neighbor adjacencies
- Resumes normal OSPF operation

Graceful Restart procedures on neighbors:
- Receives grace-LSAs from affected router
- Starts grace period (lifetime) timer
- Preserves OSPF routes and forwarding via affected router provided no topology changes
- Re-establishes adjacency with affected router
- Resumes normal OSPF operation



- If the topology changes or the NSF/GR timer expires before OSPF functions return to normal, the NSF/GR routes will be purged, potentially impacting traffic forwarding

IGP Best Practices

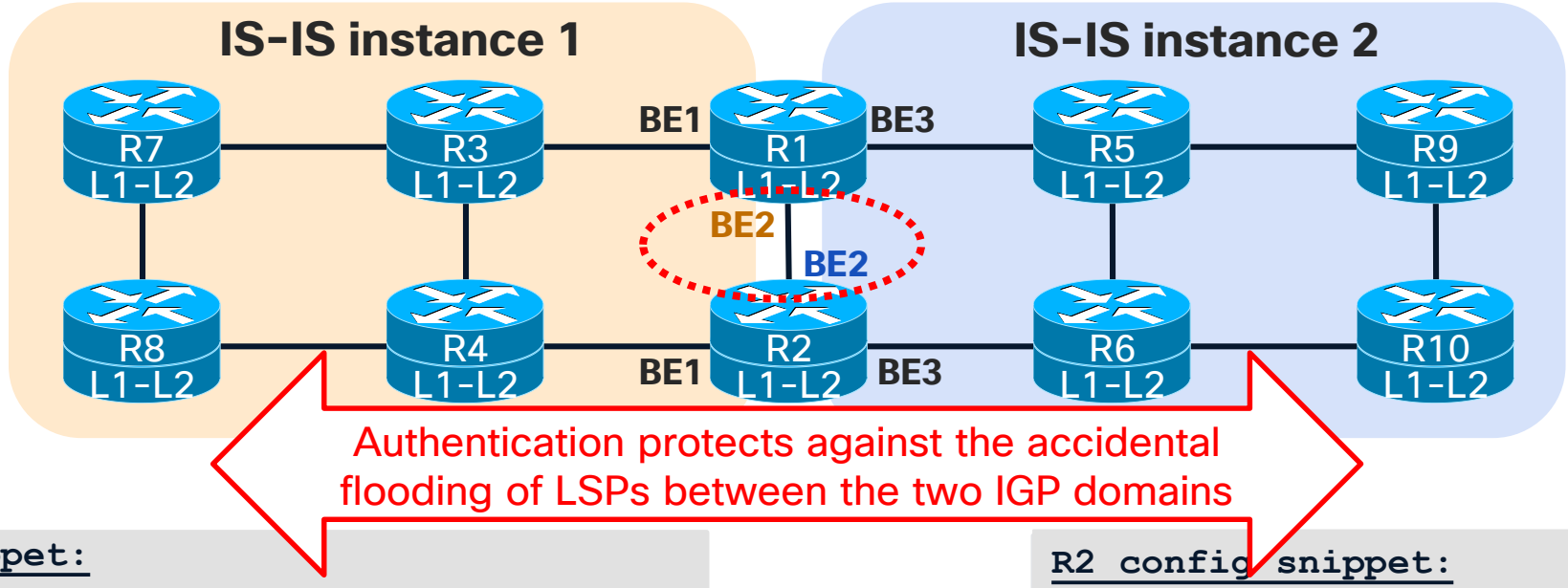
- Again, avoid BGP redistribution into IGP; If not done correctly, may cause IGP failure or routing loops
- Ensure an upper limit on the number of prefixes that can be redistributed into the IGP to protect against a misconfiguration

```
maximum redistributed-prefixes 10000 75 /* IOS XR default */
```

- Configure OSPF max-lsa commands to limit the number of non-self-generated LSAs kept in the LSDB
  - Prior to IOS XR 7.9.1 ‘max-lsa’ was disabled by default
  - IOS XR 7.9.1 added ‘max-lsa’ default of 500K
  - IOS XR 7.10.1 added ‘max-external-lsa’
  - Not applicable to IS-IS; However, IS-IS restricts the max number of LSPs an IS-IS node can originate to 256

IGP Best Practices

- Configure IGP cryptographic authentication to:
  1. Mitigate the risk of malicious IGP attacks
  2. Protect against the accidental collapsing of two IGP domains



* Example scenario applies equally to OSPF

```
R1 config snippet:
router is-is 1
  net 49.0001.0001.0001.0001.00
  interface Bundle-Ether 2
    address-family ipv4 unicast
    hello-password hmac-md5 encrypted [pwd1]
```

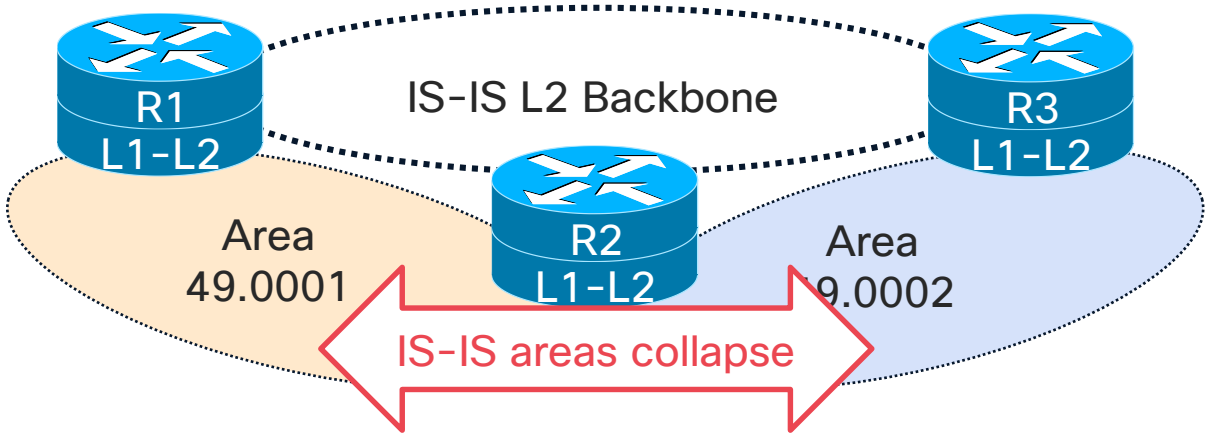
```
R2 config snippet:
router is-is 2
  net 49.0002.0001.0001.0001.00
  interface Bundle-Ether 2
    address-family ipv4 unicast
    hello-password hmac-md5 encrypted [pwd2]
```

IGP Best Practices

- Configure the IS-IS routing process type along with proper area addresses (or NETs) to ensure the proper level of adjacencies
  - is-type level-1-2 specifies that a router act as a gateway (e.g., ABR) to connect different areas (IOS XR default)
  - is-type level-1 specifies that a router only establish adjacencies with other routers in the same area
  - is-type level-2-only specifies that a backbone router cannot communicate with level-1 only routers
- Recommendation: ONLY use is-type level-1-2 configuration on ABRs

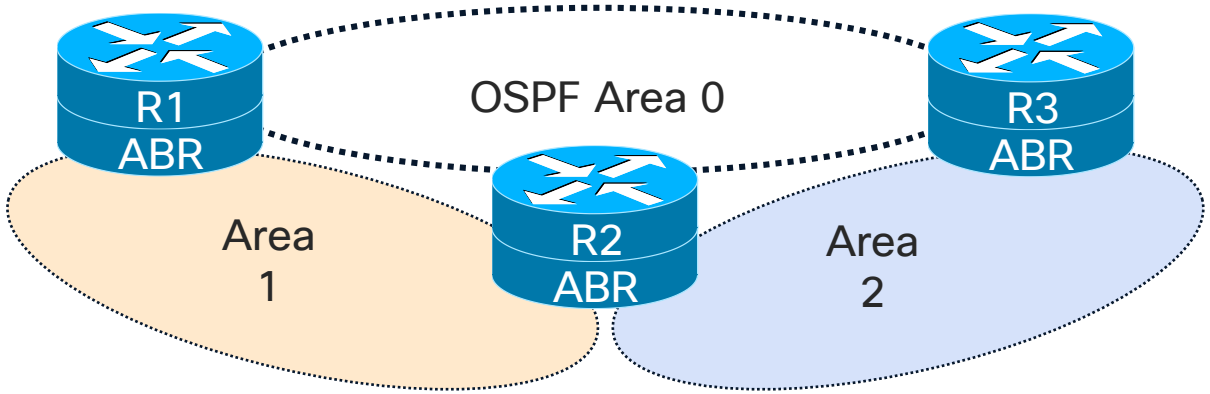
IGP Best Practices

- Recommend: Not to configure multiple area addresses for a single IS-IS instance
  - Only useful (temporarily) to merge multiple IS-IS areas or split up an area – use with caution!
  - Otherwise, risk of accidental collapse of / LSP flooding between IS-IS areas




```
R2 config snippet:
router is-is 1
  net 49.0001.0001.0001.0001.00
  net 49.0002.0001.0001.0001.00
```

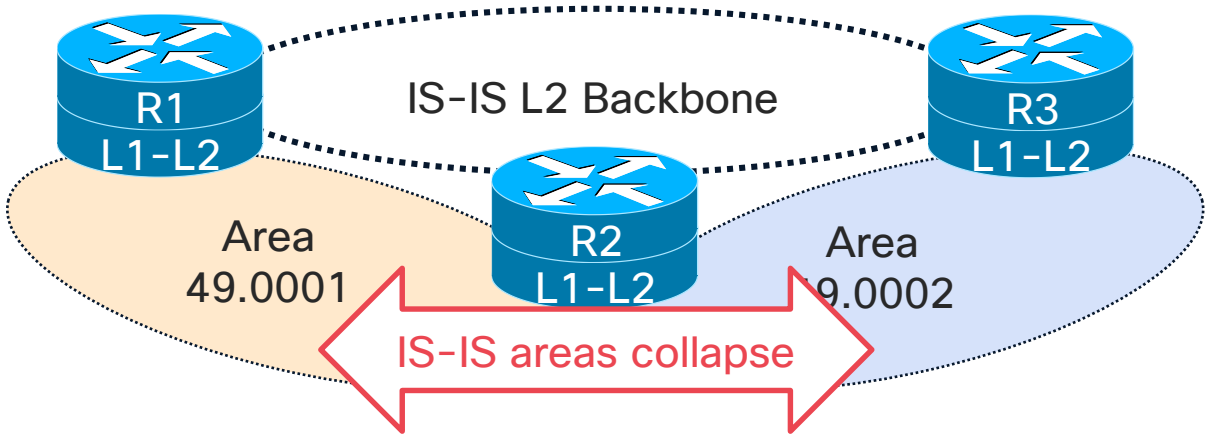
- IS-IS associates areas to nodes (not interfaces)




```
R2 config snippet:
router ospf 1
  area 0
    interface GigabitEthernet 0/0/0/0
  area 1
    interface GigabitEthernet 0/0/0/1
  area 2
    interface GigabitEthernet 0/0/0/2
```

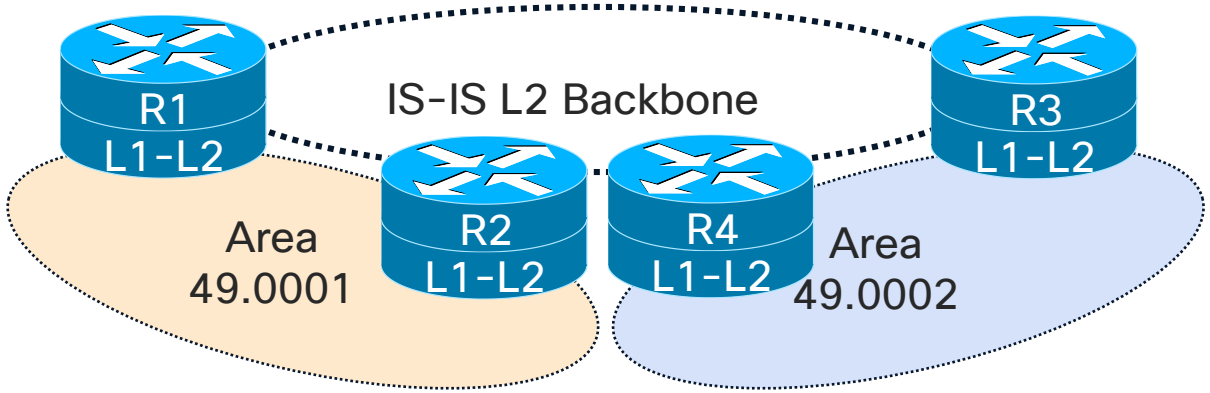
IGP Best Practices

- Recommend: Not to configure multiple area addresses for a single IS-IS instance
  - Only useful (temporarily) to merge multiple IS-IS areas or split up an area – use with caution!
  - Otherwise, risk of accidental collapse of / LSP flooding between IS-IS areas




```
R2 config snippet:
router is-is 1
  net 49.0001.0001.0001.0001.00
  net 49.0002.0001.0001.0001.00
```

- IS-IS associates areas to nodes (not interfaces)



- IS-IS requires separate ABRs per area

IGP Best Practices – Scaling / Fault Containment

- Managing IGP scale is critical to IGP and wider network stability
- Networks deployed with modern routers and a proper design can accommodate 1000+ routers in a single IGP area
- Multiple well-known techniques are available (if necessary) to manage IGP scale and blast radius:
  - Hierarchical IGP (i.e., using multiple areas)
  - Multi-instance IGP with BGP-LU (Unified MPLS / Seamless MPLS)
  - Multi-instance IGP with SR-PCE (Converged SDN Transport)
  - Multi-instance IGP with IPv6 summary routes (SRv6)
  - Multi-plane architecture with each plane having its own distinct IGP

Other IGP Best Practices (1)

Reference

- Optimal IGP exponential backoff algorithm timers to rate limit LSA/LSP generation and SPF computation during network instability (IOS XR default)
- IGP prefix prioritization of IPv4 /32 and IPv6 /128 host prefixes (i.e., I-BGP next hops) during SPF run to minimize convergence for transit traffic (IOS XR default)
- Configure "network point-to-point" for all router-to-router IGP links, otherwise they become LAN which is complicated (DR/BR, LSAs, features)
- OSPF TTL security (RFC 3682 GTSM) to filter remote attacks against OSPF
  - IS-IS runs directly on Layer 2 so it is not exposed to remote IP attacks
- Consider IGP prefix-suppression to avoid carrying the P2P prefixes of transit links in the LSDB, thereby, reducing IGP scale & convergence time

Other IGP Best Practices (2)

Reference

- Control plane policing (e.g., LPTS) to protect router CPU and ensure control plane stability (IOS XR default) – applies to all control protocols; Adjust default policing rates if necessary
- Enable LDP/IGP synchronization to prevent MPLS LSP forwarding on a link when the associated LDP session is down
- Configure LDP label allocate and LDP label advertise to permit I-BGP next hops only (e.g., PE loopbacks) so that (i) only transit traffic is MPLS LSP forwarded and (ii) to reduce MPLS forwarding (LFIB) resource consumption
- LDP session protection minimizes traffic loss and provides faster network convergence during link DOWN➔UP events
- In multi-plane architectures with multiple IGPs, avoid redistributing IGP routes between planes

IP Multicast Considerations

- NSR and NSF are also applicable and supported for PIM and IGMP
- RP (Rendezvous Point) redundancy is fundamental to high availability for IP multicast
  - With PIM-SM, RPs serve as the root node of shared trees
  - PIM-SSM does not require an RP to operate
- For large-scale IP multicast networks with numerous shared trees and significant control plane activity, it is recommended to use dedicated routers as RPs
- IP multicast trees impacted by failure events rely on unicast routing protocols to converge before they can be rebuilt
  - Consequently, IP multicast convergence benefits from fast IGP convergence
- Alternatively, technologies are available that enable FRR protection for multicast:
  - MoFRR, mLDP, P2MP RSVP-TE, IR and SR-based multicast (Tree SID)

Other IP Multicast Considerations

Reference

- Prevent unauthorized sources from registering with the PIM-SM RP (i.e., **PIM accept-register**)
- Filter PIM messages based on IP source addresses (i.e., **PIM neighbor-filter**)
- Filter PIM join and prune messages received from non-PIM neighbors (i.e., **PIM neighbor-check-on-recv enable**)
- Restrict the sending of PIM join and prune messages to non-PIM neighbors (i.e., **PIM neighbor-check-on-send enable**)
- Set upper limits for PIM register states, route interfaces, and routes (i.e., **PIM global maximum**)
- Restrict IGMP join requests to specific multicast groups (i.e., **IGMP access-group**)
- Restrict the max number of multicast groups per IGMP node (i.e., **IGMP maximum groups**)
- Restrict the max number of multicast groups per IGMP interface (i.e., **IGMP maximum groups-per-interface**)
- Configure MSDP password authentication to protect sessions against TCP attacks (i.e., **MSDP password**)

IP Forwarding Plane Best Practices

Interface Fault Handling

- Recommend using Carrier Delay and Event Dampening to avoid churn in the control plane caused by unstable interfaces or WAN circuits

```
R1(config-if)# carrier-delay down {msecs} up {msecs}
```

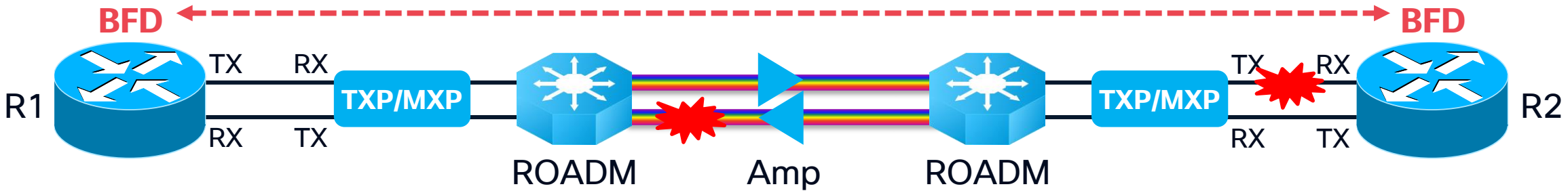
- Carrier-delay down delays processing hardware link down events (default in IOS XR = 0 msecs)
- Carrier-delay up delays processing hardware link up events (default in IOS XR = 200 msecs)
- The optimal carrier-delay down value depends on the protection and/or restoration mechanisms, if available, along with the recovery times offered by each network layer: Optical, Ethernet, and IP
- Event Dampening helps to limit the propagation of frequent interface state changes, thereby, improving network control plane stability

```
R1(config-if)# dampening /* disabled by default in IOS XR */
R1# show dampening interface - Displays dampened interfaces.
```

Bidirectional Forwarding Detection (BFD)

- Recommended for fast failure detection and, in turn, to rapidly trigger IGP/BGP control plane convergence, IP/MPLS FRR protection and BGP PIC Edge
- Also provides end-to-end L1 optical path verification and advanced capabilities

Failure Detection Method	Detection Time	Applicability
Optical LoS/LoF	~10 msecs.	• Relies on optical network to propagate remote faults
Routing protocol timers	>=30 secs.	• Lower timers affect control plane scale
Ethernet CFM with EFD	>=12 msecs.	• HW offload enables fast timers
BFD	>=12 msecs.	• HW offload enables fast timers • Supports advanced capabilities: BFD strict mode, BFD dampening, BFD over Bundles, BFD multi-hop

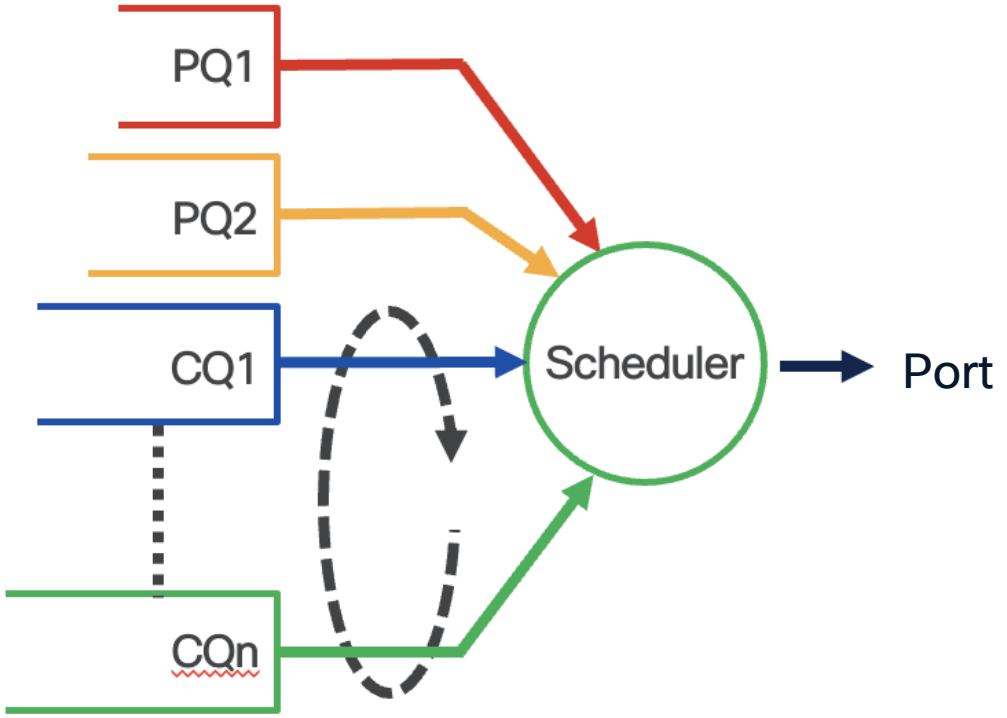


Bidirectional Forwarding Detection (BFD)

- BFD for Ethernet Bundles includes two (2) deployment modes:
  - BFD over Bundles (aka "BoB") - defined in RFC 7130. "BoB" is where a (micro) BFD session is run on each member link within the bundle. Cisco supports two "BoB" modes: Cisco mode (pre-standard) and IETF mode. IETF mode is recommended and required for multi-vendor BoB interoperability
  - BFD over Link Bundles (aka "BLB") - defined in RFC 5880. "BLB" is where a single BFD session is run for the entire bundle
- BoB is recommended and provides faster detection versus BLB

QoS Considerations

- DiffServ QoS plays a vital role in ensuring high network availability
  - Isolates different traffic classes and guarantees priority traffic during congestion
- Recommendation: Designate a traffic class queue (e.g., CQ1) with ample bandwidth exclusively for control and management plane traffic to avoid drops
  - If control plane sessions / adjacencies go down, traffic to affected prefixes may be disrupted
- Recommendation: At the network edge (i.e., PE), accurately mark (or color) traffic to ensure proper packet classification and queuing downstream
  - IP control plane traffic (e.g., BGP, OSPF) is by default marked as high priority (i.e., IP DSCP value cs6)



QoS Buffer Sizing Considerations

- Take note of buffer sizing variations on high-speed interfaces of modern routers

Platform NPU	NPU Bandwidth	NPU HBM	~Buffering
ASR 9000 3rd Gen	240 Gbps	6 GB	<200 msec. @ 100GE
ASR 9000 5th Gen	400 Gbps	3 GB	<100 msec. @ 100GE
NCS 5700	10 Tbps	8 GB	<50 msec. @ 400GE
8000 Q200	12.8 Tbps	8 GB	<50 msec. @ 400GE

- Proper buffer configurations are vital during platform migrations to prevent indiscriminate 'no buffer' packet drops during congestion – which may lead to traffic disruption
- Additionally, configurations should account for the number of NPU ports that may experience simultaneous congestion

MPLS Label Scale Considerations

Configuration	Applicability	ASR 9000	NCS 5500/5700	8000
LDP label allocation for host-routes only	P + PE	Recommended	Recommended	Recommended
MP-BGP label allocation for L3VPNs and 6PE address families	PE	Per-Prefix* Per-VRF Per-CE	Per-VRF; Required	Per-VRF; Required
Inter-AS L3VPN Option B MP-BGP sessions	ASBR	Per-Prefix	Per-NextHop-Received-Label; Recommended	Per-NextHop-Received-Label; Recommended
I-BGP next hop on PEs for IPv4/IPv6 address families	PE	Next-hop-unchanged* Next-hop-self	Next-hop-self; Recommended	Next-hop-self; Recommended

- Recommend settings above to help avoid Out-of-Resource (OoR) label forwarding conditions as well as improve routing convergence

* Default IOS XR configuration

Internet VRF vs. GRT

- Benefit of Internet VRF:
  - Automatically blocks external IP access to core routers
- Challenges associated with Internet VRF:
  - Migrating a brownfield network from the GRT to a VRF is complex
    - For example, PE routers may have to carry the BGP Internet table twice during the migration:
      1. Once as IPv4/IPv6 unicast routes
      2. Again, as VPNv4/VPNv6 unicast routes
    - This doubles RIB/FIB scale during migration
  - VPN prefixes consume more scarce router resources (e.g., MPLS labels)
  - Makes intra-domain Internet routing more complex (e.g., RDs, VRFs, RTs, VRF import/export)
  - If applicable, makes Internet multicast more complex (i.e., MVPN)
- Operators must carefully balance the risks, complexity, and costs

Fast Re-Route (FRR) Protection

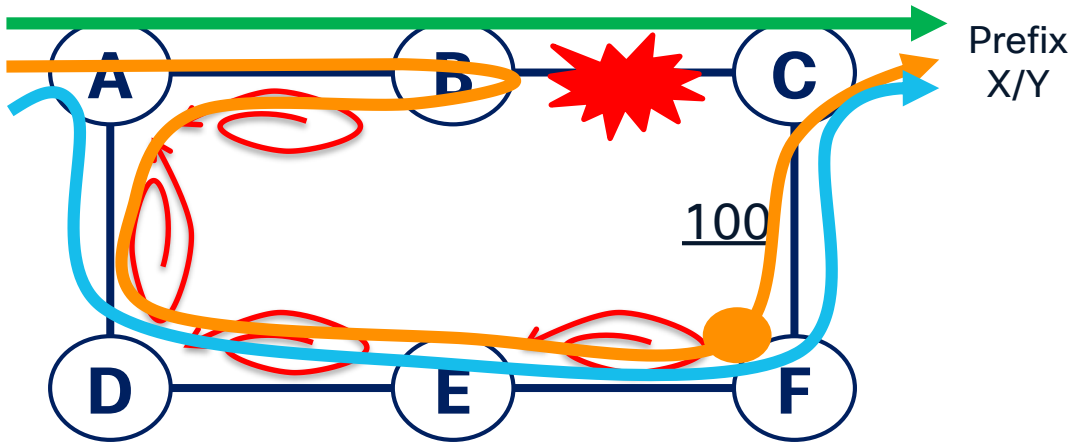
- IGP convergence times vary, typically falling under 2-3 seconds
  - Sub 1 second convergence is achievable using modern hardware and the optimized (IOS XR default) timers
- To minimize traffic disruption further, requires FRR protection
- Different techniques available to attain 50 msec. FRR protection
  - MPLS/RSVP-TE – requires MPLS and many stateful core tunnels
  - Per-Prefix LFA – cannot guarantee FRR coverage (e.g., box topology)
  - Remote LFA – requires targeted LDP sessions be established
  - TI-LFA – topology independent, no stateful tunnels, no targeted LDP sessions (enabled by SR)

Recommended

* LFA (Loop Free Alternate)

TI-LFA FRR Protection

```
router ospf 100 area 0 fast-reroute per-prefix
router ospf 100 area 0 fast-reroute per-prefix ti-lfa enable
```



Default metrics: 10

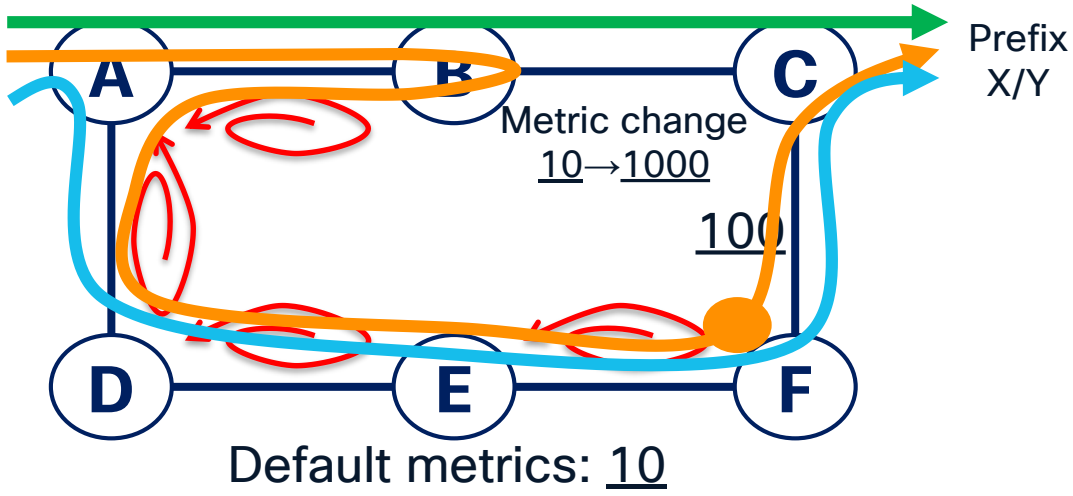
- IGP path before link failure
- Traffic (micro) loop in the absence of TI-LFA
- TI-LFA protect path (SID list @ B = [F, F→C])
- Post IGP convergence path

- <50 ms protection: link, node, SRLG failures
- Simple to operate and understand given it's automatically computed by the IGP:
  - Node B computes shortest path to Prefix X/Y via Node C (active path)
  - Node B also computes TI-LFA path to prefix X/Y via Node F (TI-LFA backup path)
  - When Node B detects link failure to Node C, it FRR switches traffic onto TI-LFA backup path
  - TI-LFA backup path is used until IGP reconverges, thereby, minimizing traffic disruption
- No stateful core tunnels required
- 100% topology coverage / independent

SR Microloop Avoidance

- Hop-by-hop IP routing may induce transient microloops during convergence events
  - E.g., link up, interface shutdown, metric change
- Microloops can lead to increased packet loss which is obviously undesirable
- SR microloop avoidance prevents microloops for isolated convergence events
- When a node learns of a topology change and then computes new paths for its destinations:
  - If the node sees that transient microloops are possible for a destination, then it constructs a SID-list to steer traffic microloop-free
  - SID-list @ A, B, D, E for Prefix X/Y = [F, F→C]

```
ipv4 unnumbered mpls traffic-eng Loopback0
router ospf 100 microloop avoidance segment-routing
```



IGP path before link metric change

Traffic loop (microloop) due to link metric change and the absence of SR microloop avoidance

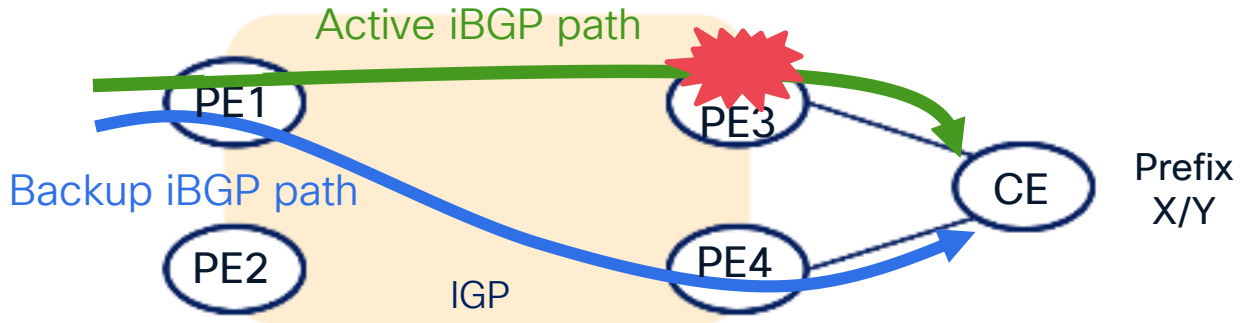
SR microloop avoidance

Post IGP convergence path

BGP Prefix Independent Convergence (BGP PIC)

BGP PIC Edge Node Protection

- PE1 has active path(s) to prefix X/Y
  - I-BGP multipath
  - I-BGP active and backup paths
- PE1 has alternate path via PE4 pre-programmed in its FIB in advance of PE3 node failure

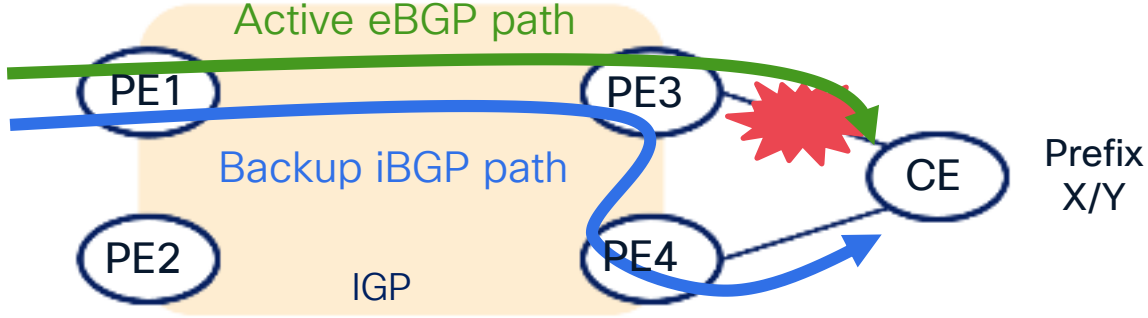


- Egress PE node failure (PE3) triggers BGP PIC Edge Node Protection on ingress PE1
  - Triggered by the removal of PE3 from PE1's IGP database
- PE1's convergence onto its alternate PE4 path is BGP prefix independent and pre-programmed, making it fast (~sub-second)
  - Again, it depends on IGP convergence and the removal of PE3

BGP Prefix Independent Convergence (BGP PIC)

BGP PIC Edge Link Protection

- PE3 has active and backup paths to prefix X/Y
  - E-BGP path (active)
  - I-BGP paths (backup)
- PE3 has alternate path via PE4 pre-programmed in its FIB in advance of PE3-CE link failure



- Egress PE-CE link failure (PE3-CE) triggers BGP PIC Edge Link Protection on egress PE3
- PE3's convergence onto its PE4 backup path is BGP prefix independent and pre-programmed, making it fast (~sub-second)
- BGP PIC Edge Link Protection requires Per-Prefix or 'Resilient' Per-CE label allocation

Other IP Forwarding Plane Best Practices

Reference

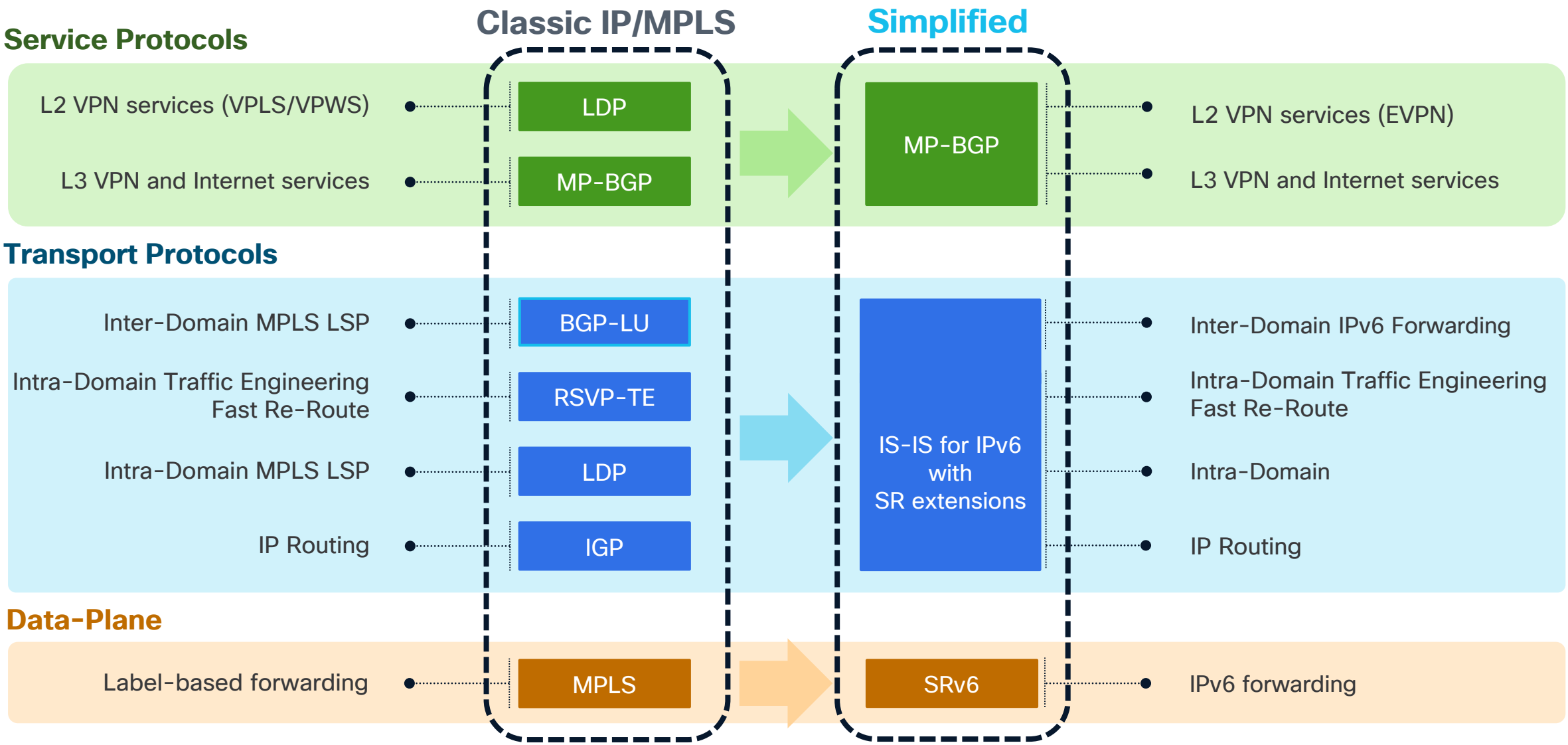
- Interface ACLs / packet filtering at edge – automation is key in maintaining accuracy
- If providing Internet services, Unicast RPF or ACLs to mitigate IP source address spoofing (IETF BCP 38 and BCP 84) as well as to facilitate traceback of security attacks
- ICMP best practices – no ip unreachables, no ip redirects, no IP→MPLS TTL propagation to prevent TTL expiry attacks
- Network-wide MTU sizing to avoid IP fragmentation and/or ICMP 'Fragmentation Needed and Don't Fragment was Set'

Network Simplification

Network Simplification

- Reducing potential failure scenarios makes achieving high availability easier
- Makes automation easier and more sustainable long-term
- Recommended technologies for a simplified IP/MPLS network
  - MP-BGP
  - Segment Routing
  - SR-TE
  - Centralized SDN Controller (PCE)
  - BGP-LS
  - BFD
  - DiffServ QoS
  - YANG model driven manageability
  - Routed Optical Networking

Network Evolution



Segment Routing (SR)

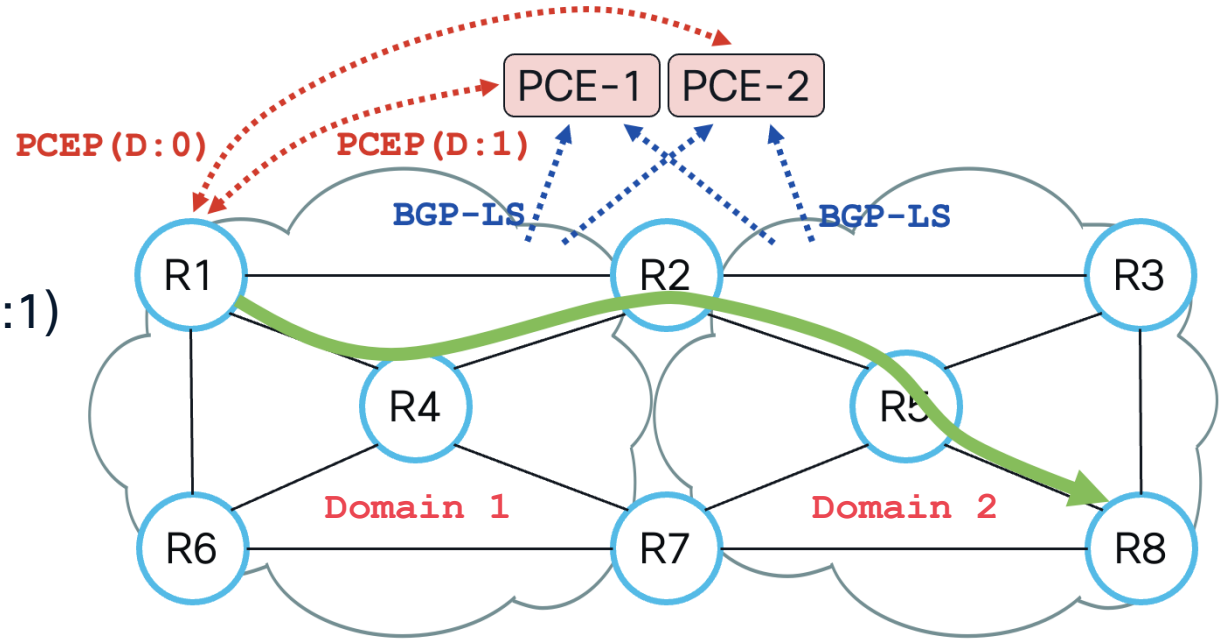
- A programmatic IP source-routing architecture that provides the optimal balance between distributed intelligence and centralized control
- Mass network simplification
  - Reduces control plane protocols (LDP, RSVP-TE, BGP-LU, MPLS OAM, IGP/LDP sync)
  - Unified forwarding plane for all services (IP, MPLS VPN, Ethernet, Private Line, Wave)
  - Automatic topology independent 50 msec FRR protection
- Mass network scaling
  - No stateful TE tunnels throughout the infrastructure, On-Demand path instantiation
  - Transport route summarization between network domains (SRv6)
- Advanced network capabilities
  - Advanced TE: e.g., intent-based, ECMP-aware, multi-domain, circuit-style, on-demand SR path instantiation (ODN), automated traffic steering, network slicing, service chaining, and integrated performance measurements

* Note, SRv6 provides maximum simplicity, scale and capabilities

SR PCE High Availability

Reference

- SR PCE is only required for SR-TE if more information is needed than is available on a head-end
  - For example: multi-domain paths or disjoint paths from different head-ends
- SR PCE leverages the well-known standardized PCE HA:
  - When an SR policy is instantiated, updated or deleted, the head-end sends a PCEP Report to all its connected PCEs
    - Includes optimization objectives & constraints
    - Head-end delegates control to prime SR PCE (D:1)
  - Primary SR PCE: (i) computes path, (ii) derives SID-list, (iii) updates path on head-end
  - Head-end programs SID-list and reports it to all its connected SR PCEs
  - Upon failure of the primary SR PCE, head-end re-delegates control to another SR PCE – no impact on SR policies or traffic forwarding!



Crosswork Planning

Reference

Key Features

Predictive AI

Predict the impact of network changes, traffic growth, new services, and potential failures

Capacity Planning

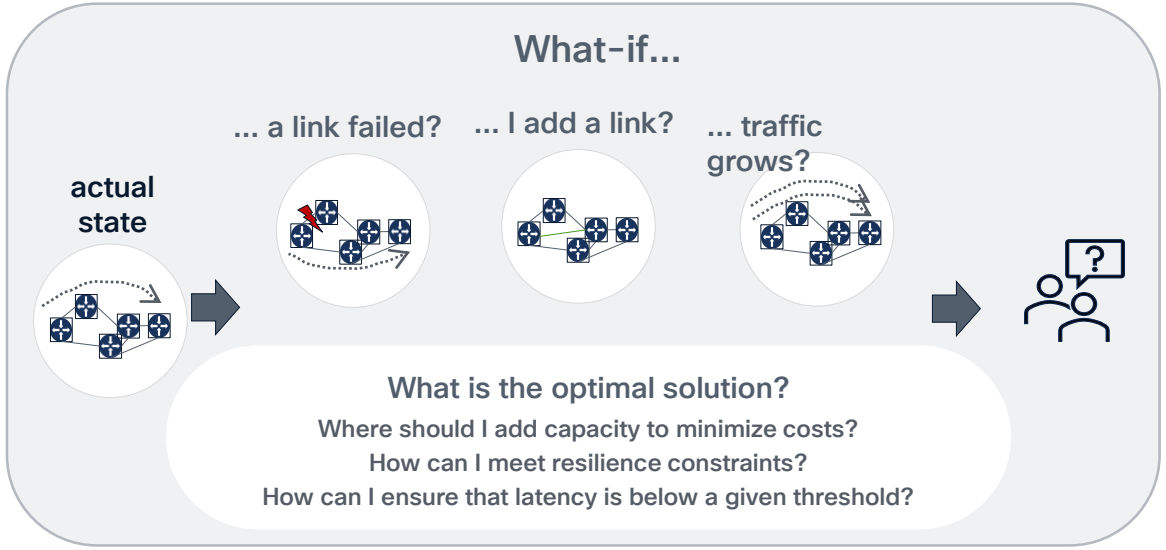
Leverage measured or simulated traffic data for accurate predictions

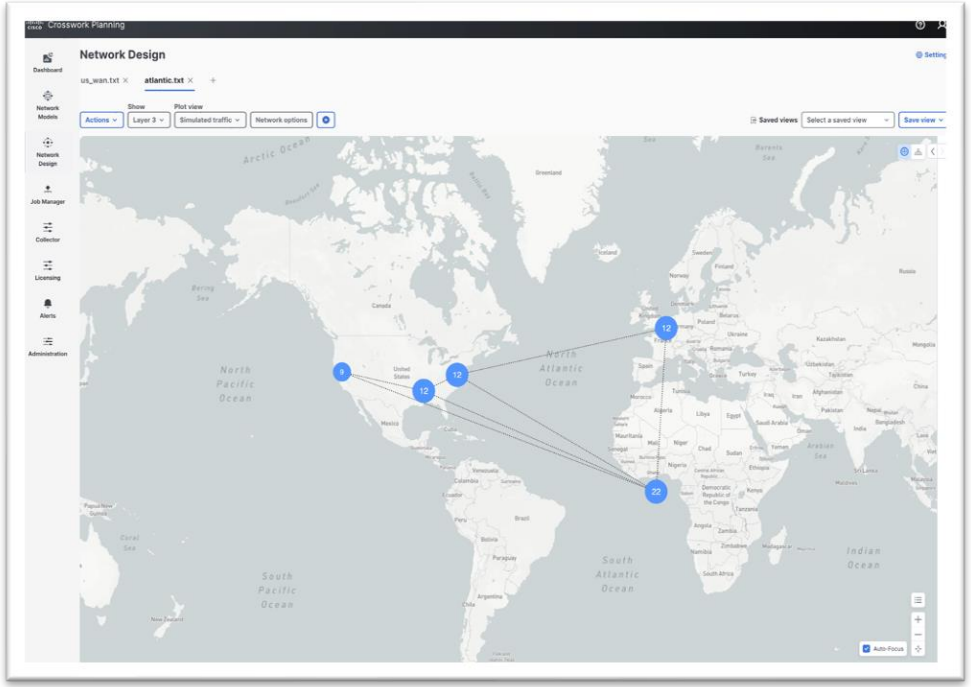
Services Optimization

Optimize network design for efficiency and reliability

Benefits

- Reduced operational costs
- Improved network performance
- Enhanced agility
- Proactive planning
- Simplified Capacity planning





Summary

Summary

- Architectural best practices can help reduce the impact of failures and minimize the risk of network outages
- Implementing all applicable best practices is recommended to ensure high availability in IP/MPLS networks

Further Reading (1)

Reference

- J. Evans and C. Filsfils. Deploying IP and MPLS QoS for Multiservice Networks. Morgan Kaufmann, 2007.
- O. Hashmi. Cisco IOS XR Deployment Best Practices for OSPF/IS-IS and BGP Routing. Cisco.com, 2022.
- M. Mishra and S. Krier. A Deep Dive into Basic and Design Best Practices for BGP and L3VPN. BRKMPL-2103, Cisco Live, 2024.
- C. Oggerino. High Availability Network Fundamentals. Cisco Press, 2001.
- G. Schudel and D. Smith. Router Security Strategies: Securing IP Network Traffic Planes. Cisco Press, 2008.
- K. Lee, F. Lim and B. Ong. Building Resilient IP Networks. Cisco Press, 2005.
- Documentation blogs and tutorials on all things IOS XR: https://xrdocs.io/
- Segment Routing: www.segment-routing.net

Further Reading (2)

Reference

- S. Brady. How Complex Systems Fail. LinkedIn.com, July 2024.
- J. Evans. No Packet Left Behind: Minimising Packet Loss Through Automated Network Operations. NANOG 88, 2023.
- N. McKeown, G. Appenzeller and I. Keslassy. Sizing Router Buffers (Redux). ACM SIGCOMM, pp. 69–74, 2019.
- G. Appenzeller, I. Keslassy and N. McKeown. Sizing Router Buffers. ACM SIGCOMM, pp. 281–292, 2004.
- C. Villamizar and C. Song. High performance TCP in ANSNET. ACM Computer Communications Review, 24(5):45–60, 1994.
- C. Mosig, et al. Revisiting Recommended BGP Route Flap Damping Configurations. Proc. of IEEE/IFIP Network Traffic Measurement and Analysis Conference, 2021.
- Understand BGP RPKI with XR7 Cisco 8000 Whitepaper. Cisco.com, October 2022.

Acknowledgements

- Phil Bedard, Luc De Ghein, Les Ginsberg, Lampros Gkavogiannis, Jakob Heitz, Serge Krier, Mankamana Mishra, Peter Psenak, Marius Stoica, Ketan Talaulikar

Complete Your Session Evaluations

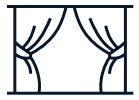
Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn exclusive prizes!

Complete your surveys in the Cisco Live mobile app.

Continue your education



Visit the Cisco Showcase for related demos

Book your one-on-one Meet the Engineer meeting

Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs

Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Contact us at: djsmith@cisco.com and lokhanna@cisco.com

Thank you

cisco Live !


CISCO