Architectural Best Practices for Ensuring High Availability

In Service Provider IP/MPLS Backbone Networks

David J. Smith - Distinguished Solutions Engineer

BRKSPG-2080

Agenda

- Network redundancy
- IP control plane best practices
- IP forwarding plane best practices
- Network simplification

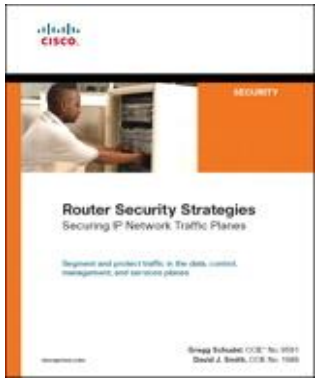
Out of Scope

- Operational guidelines
- Operation and maintenance
- Network management tools
- Incidents and recovery

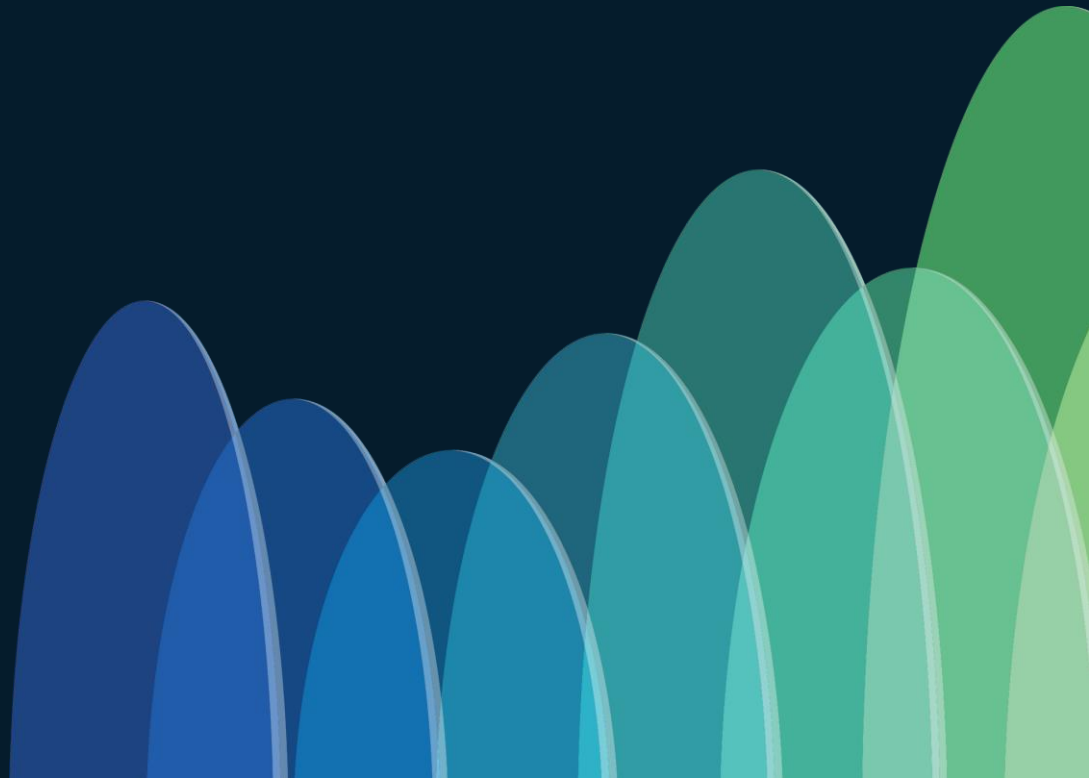
Refer to BRKSPG-2695:
“Resilient Networks: From Prevention to Recovery”

About Me

- Live in the New York City area (USA)
- Joined Cisco in August 1995
- SE supporting Service Providers in the Americas for 25+ years
- Contact: djsmith@cisco.com



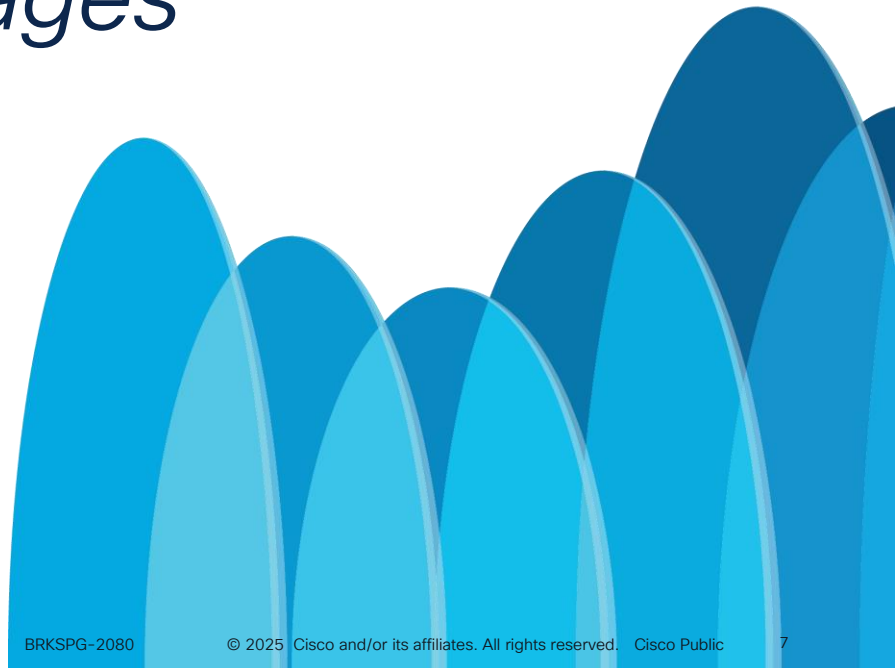
Network Redundancy



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?

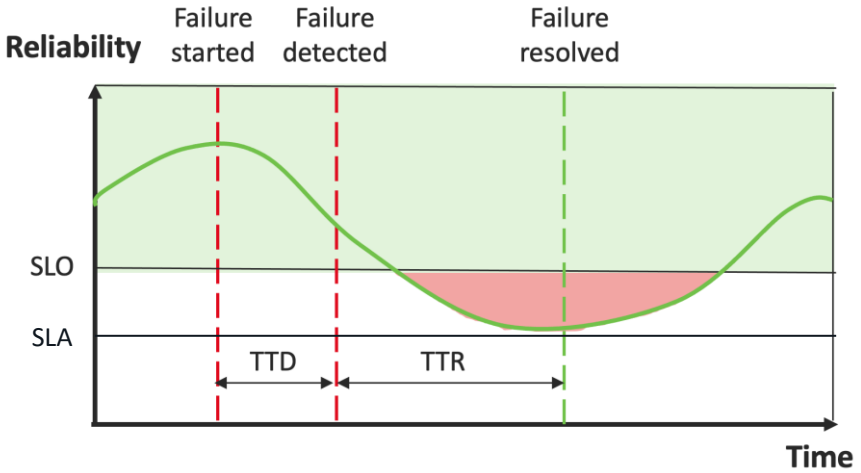


Network Redundancy Considerations

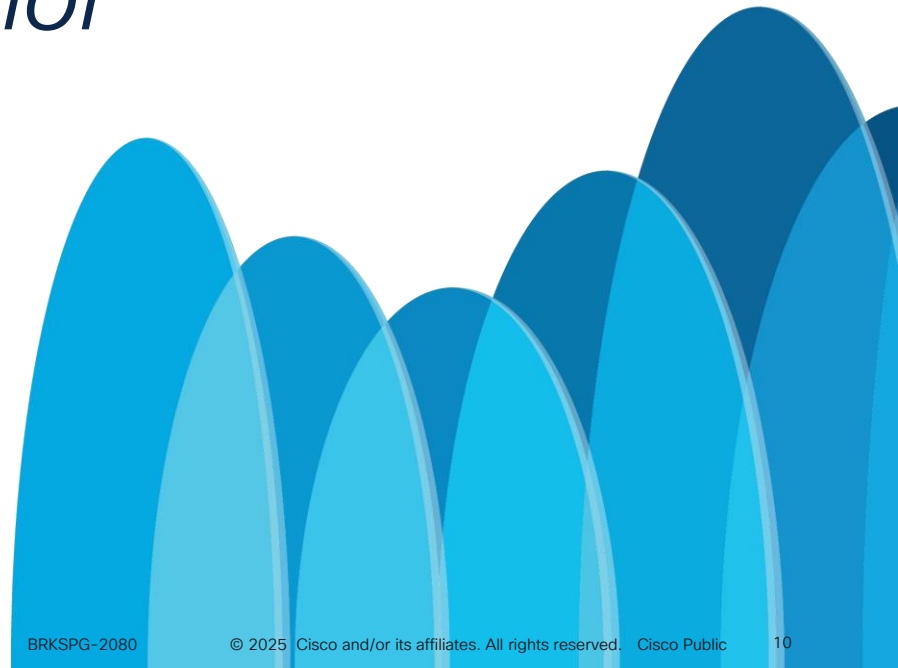
- Redundancy is highly effective when components fail hard and fast
- However, when a component malfunctions without completely failing, the consequences can be severe, even with redundancy

Goals for Network Availability

- 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

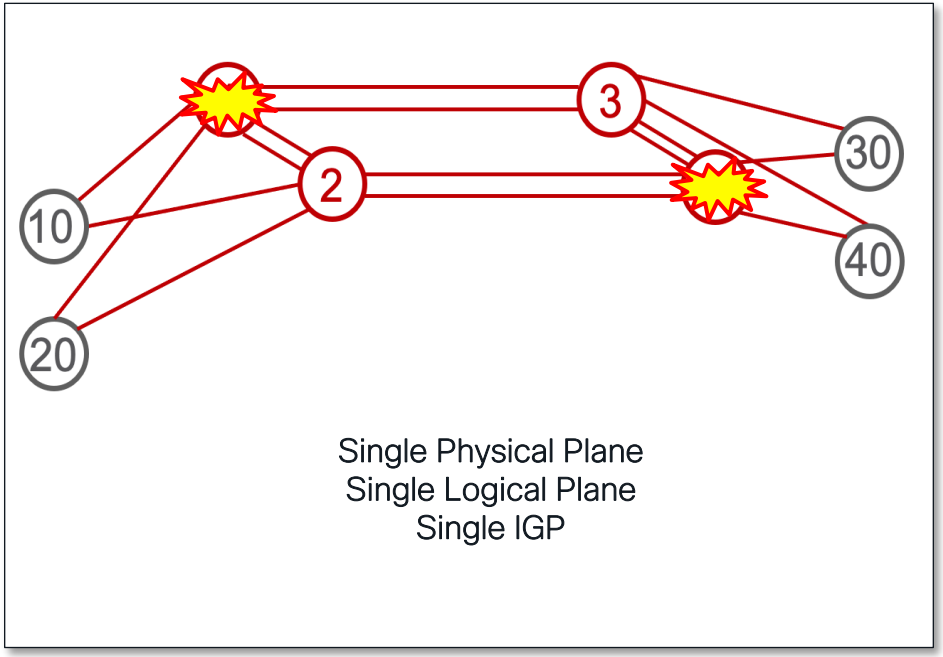
- Asymmetric relationship between cause and effect:
  - Relatively small causes can produce very large effects

Avoid the Critical State

Architectural Principles For Highly Available IP/MPLS Networks

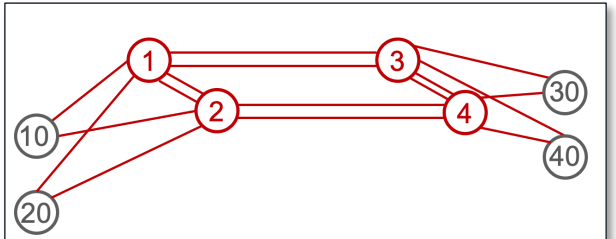
- Network redundancy
- IP control plane best practices
- IP forwarding plane best practices
- Network simplification

Network Redundancy Considerations

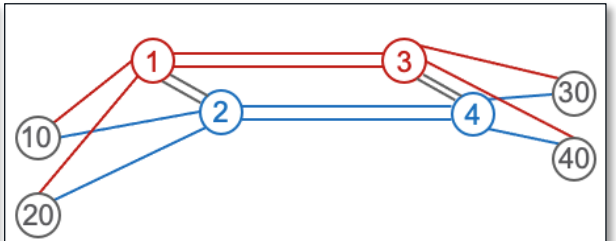


* Assumes diverse optical paths

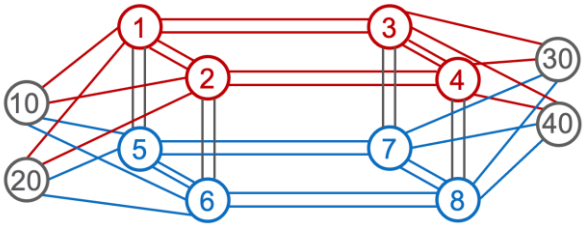
Multi-Plane Network Redundancy



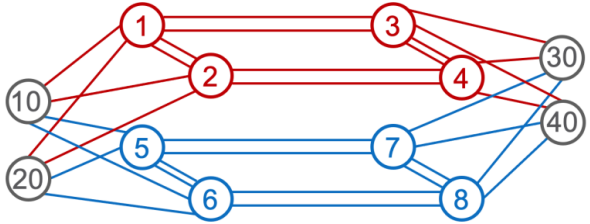
1 Single Physical Plane
Single Logical Plane
Single IGP



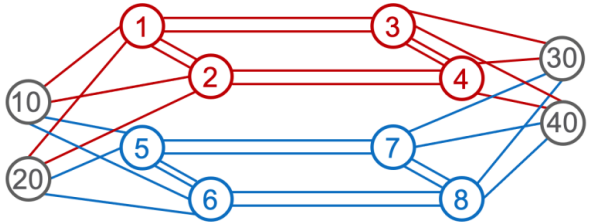
2 Single Physical Plane
Two Logical Planes
Single IGP



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



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



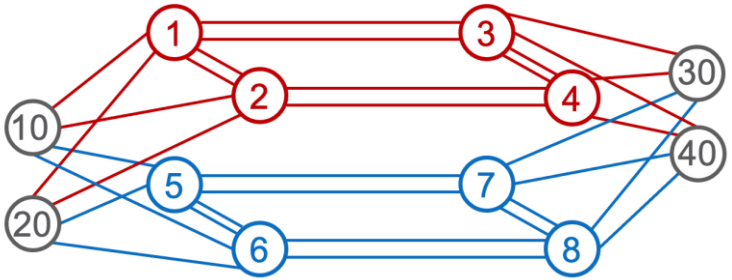
5 Two Physical Planes
Two Logical Planes
Two IGPs (one per plane)

* Assumes diverse optical paths

Multi-Plane Traffic Steering

- Use cases 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
  - MP-BGP policies
  - MPLS/RSVP-TE tunnels
  - SR-TE policies
  - SR Flexible Algorithms

} Recommended

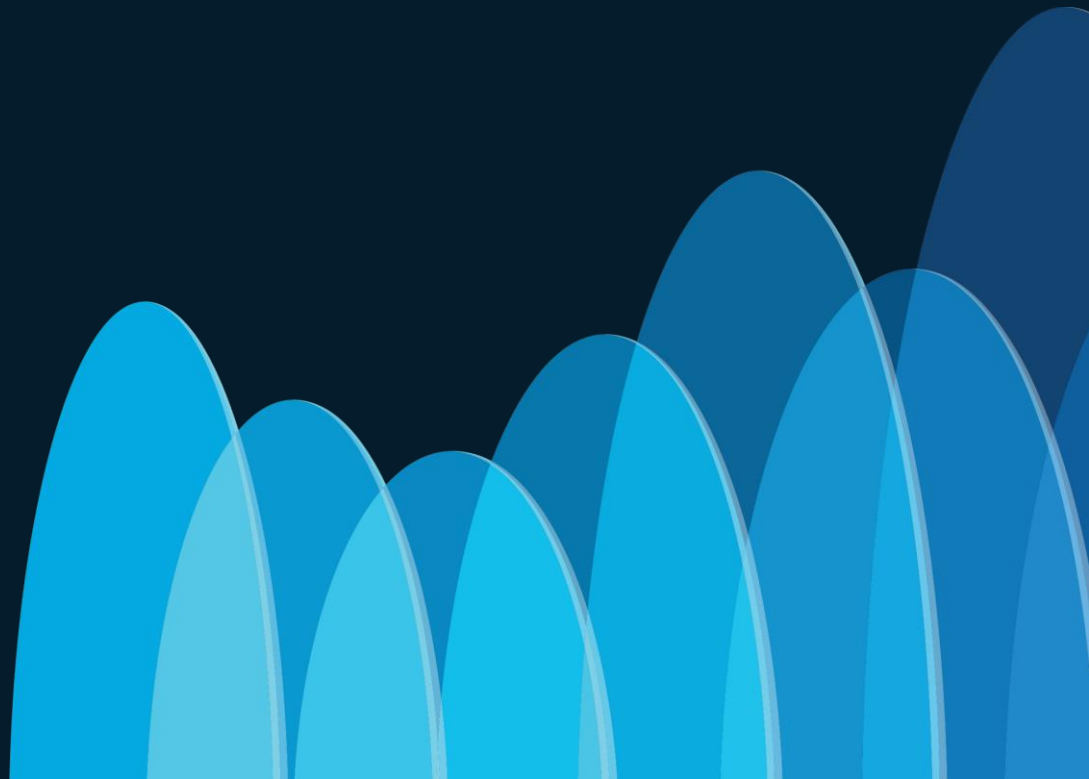


Node Redundancy Considerations

	Distributed	Centralized	Fixed
HW redundancy	Full redundancy (RP, fabric, power, fans)	Full redundancy (RP, SC, power, fans)	Partial redundancy (power/fans only, not RP)
Scaling	Scales vertically; Facilitates BW scaling	Scales horizontally; Facilitates service scaling	Scales horizontally; Facilitates service scaling
Blast radius	Large	Small	Small

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

IP Control Plane Best Practices



cisco Live!

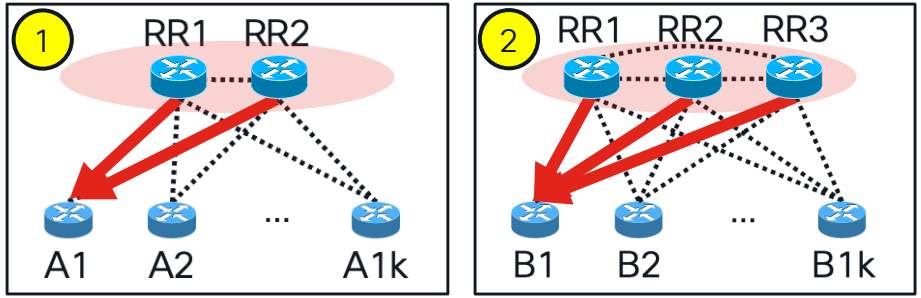
BGP Considerations

- 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

BGP Route Reflectors

- RRs simplify BGP control plane provisioning in large-scale BGP networks
- RR redundancy eliminates single points of failure, however, excessive redundancy can be detrimental
  - Increases BGP I/O packet processing
  - Increases the number of BGP paths
  - Delays BGP convergence

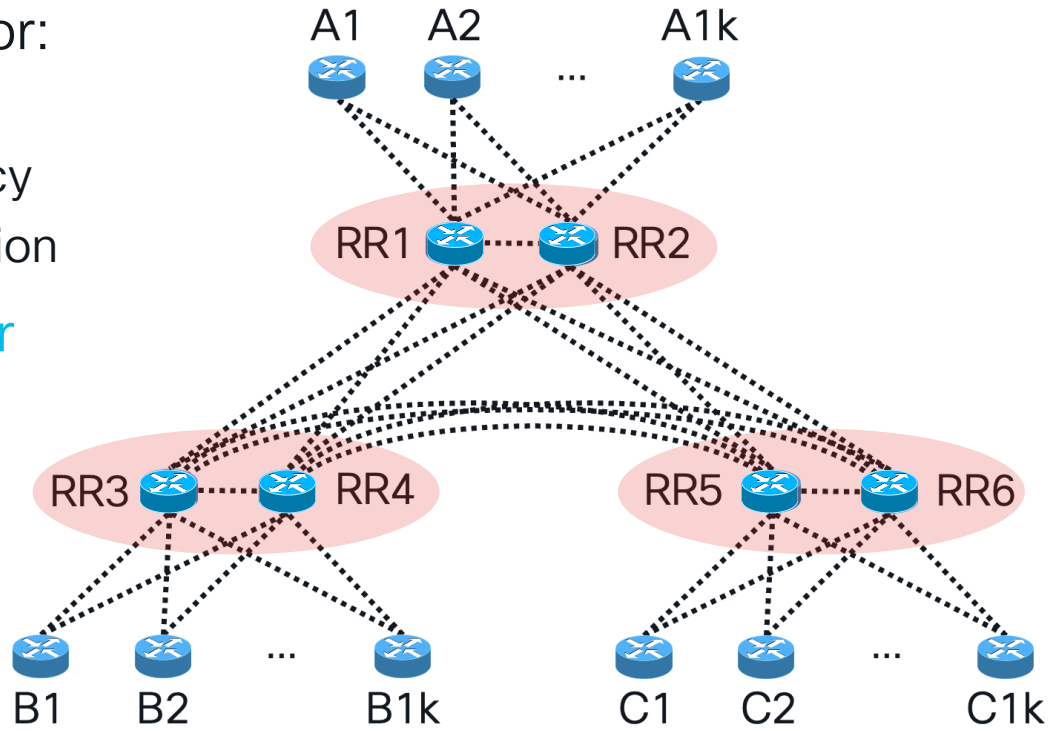
BGP RR Redundancy



- The more redundant RRs, the more:
  - BGP sessions per client
  - BGP I/O processing per client
  - BGP paths carried per client
  - Example: B clients get 50% more copies than A clients (e.g., 3M vs. 2M)
- RR clients also cannot remove an I-BGP route until a WITHDRAWN is received from each of its RRs
  - Delays or absence of a BGP WITHDRAWN for an unavailable route may slow convergence or disrupt traffic, respectively
- In large-scale BGP networks, it's recommended to avoid more than three (3) RRs in a cluster to reduce BGP overhead per above

BGP RR Scaling

- Large-scale BGP networks may require multiple RR clusters for:
  - BGP scaling
  - Fault isolation / geo-redundancy
  - Optimal routing based on location
- Maintaining smaller RR cluster sizes (e.g., <1K clients) helps reduce the blast radius of RR cluster failures



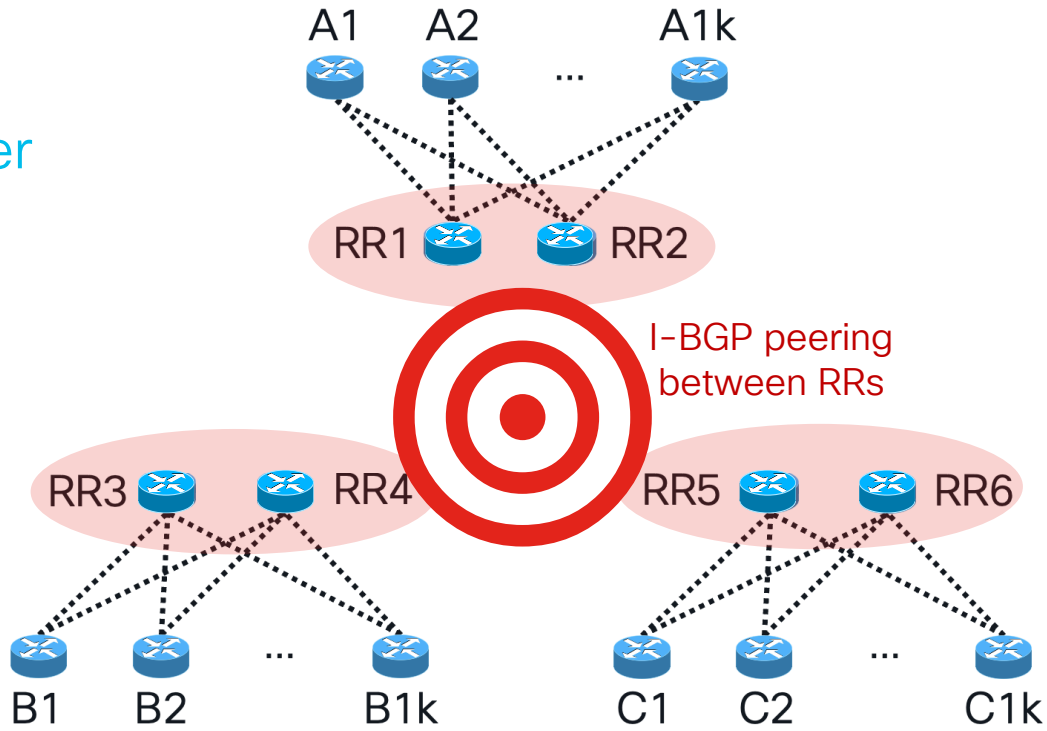
I-BGP peering

RR within a cluster

RR client

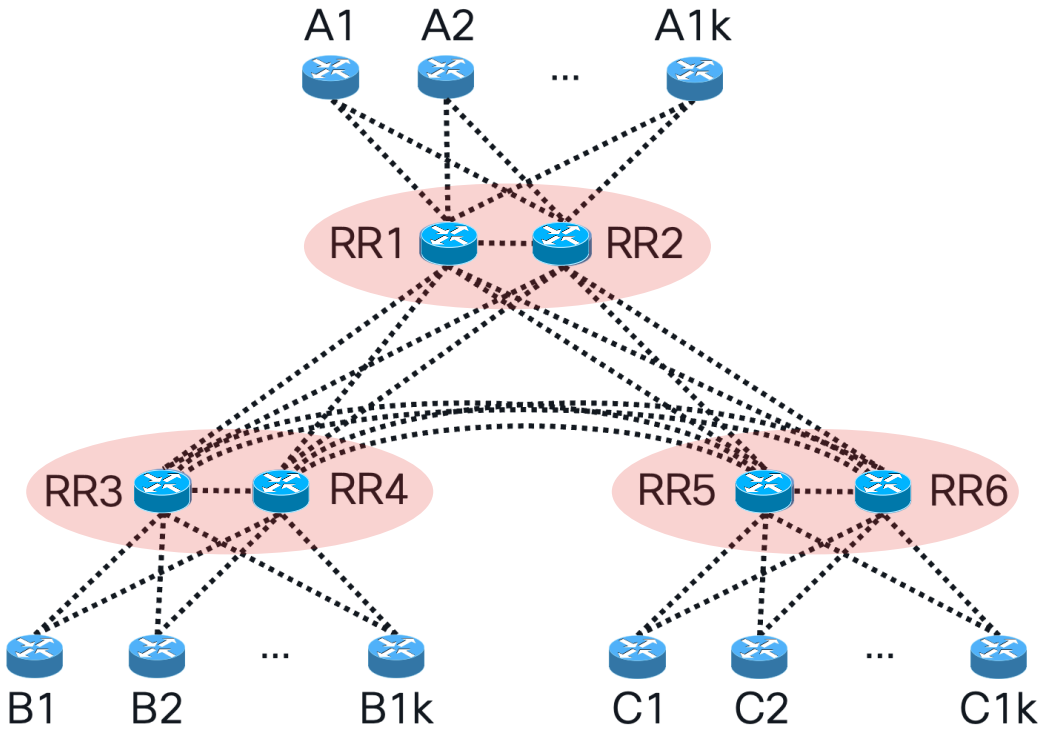
Multi-Cluster BGP RR Designs

- In general, there are three (3) main variations of multi-cluster BGP RR designs:
  1. Full mesh design
  2. Multi-plane design
  3. Multi-tier design
- Combinations of these variations can also be implemented together in a hybrid design(s)



Multi-Cluster BGP RR Design: Full Mesh

- Full peering mesh between RR clusters
  - N(N-1)/2 BGP peering's amongst the RRs



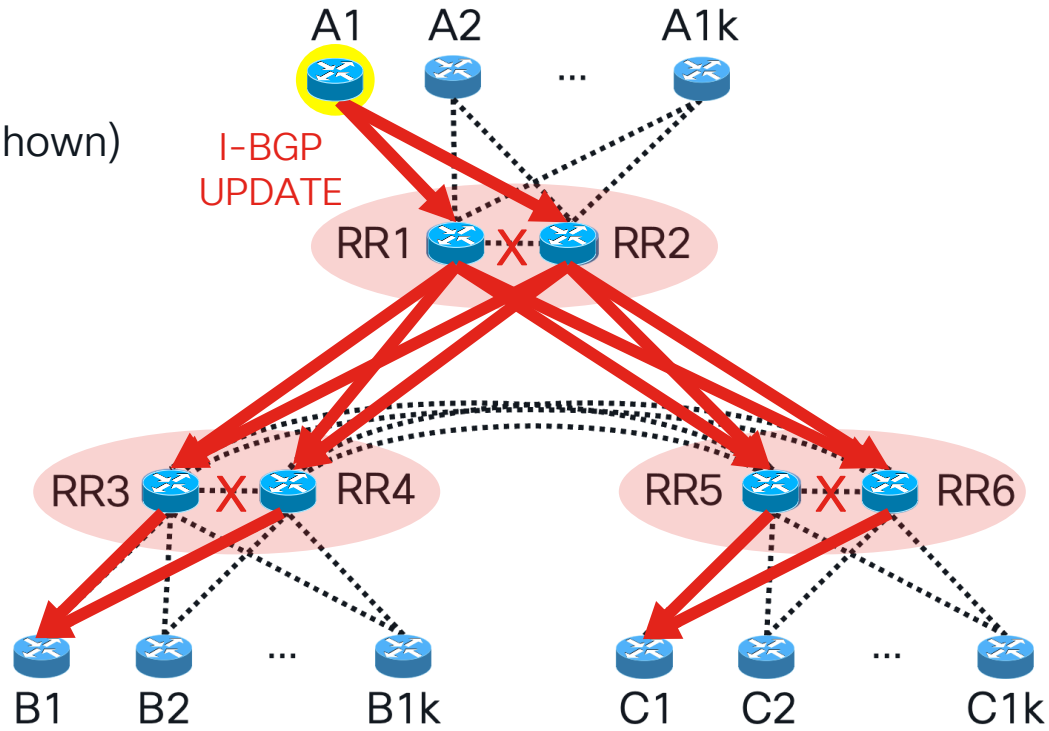
I-BGP peering

RR within a cluster

RR client

Multi-Cluster BGP RR Design: Full Mesh

- RRs advertise BGP UPDATES from clients to each of their I-BGP peers
  - One copy per RR peer
  - One copy per RR client peer (not shown)
- Adding more redundant RRs, adds:
  - More BGP sessions per RR
  - More BGP I/O processing per RR
  - More BGP paths carried per RR
  - Similar impacts to RR clients



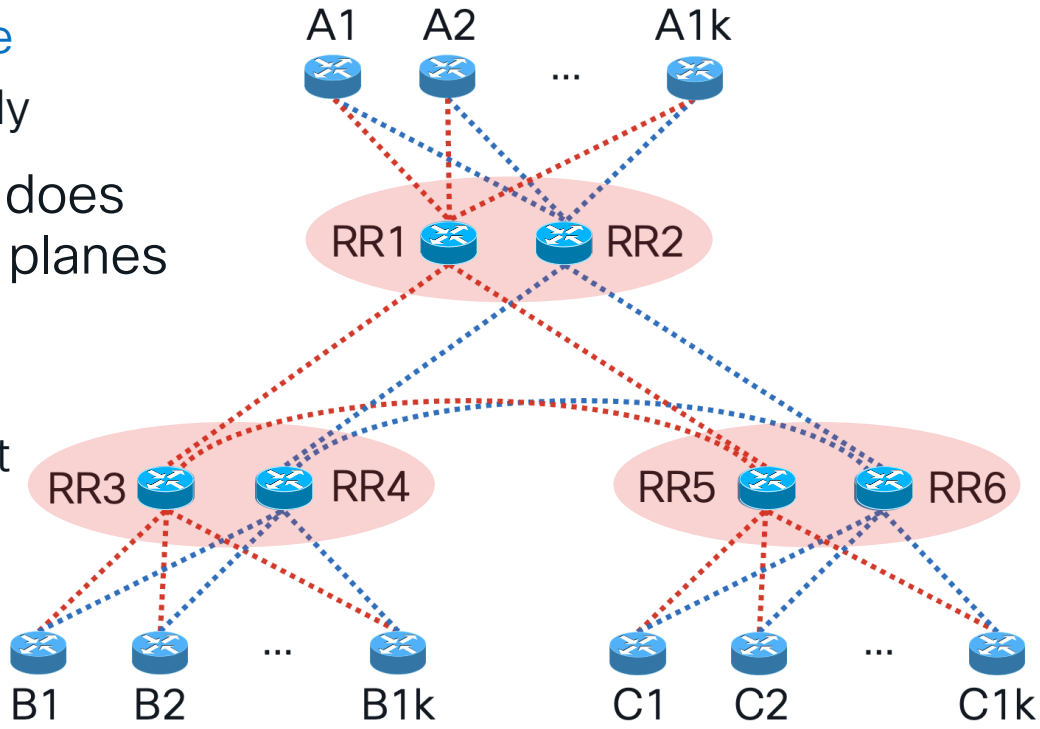
Origin of I-BGP UPDATE

I-BGP UPDATE

X Discarded I-BGP Update due to same Cluster ID

Multi-Cluster BGP RR Design: Multi-Plane

- Multiple independent RR planes
  - Example: Red plane, Blue plane
  - Full mesh within each plane only
- Adding more redundant RRs, does not adversely affect other RR planes
  - Affects RR clients only
  - However, additional RR planes may improve protection against multiple BGP failures

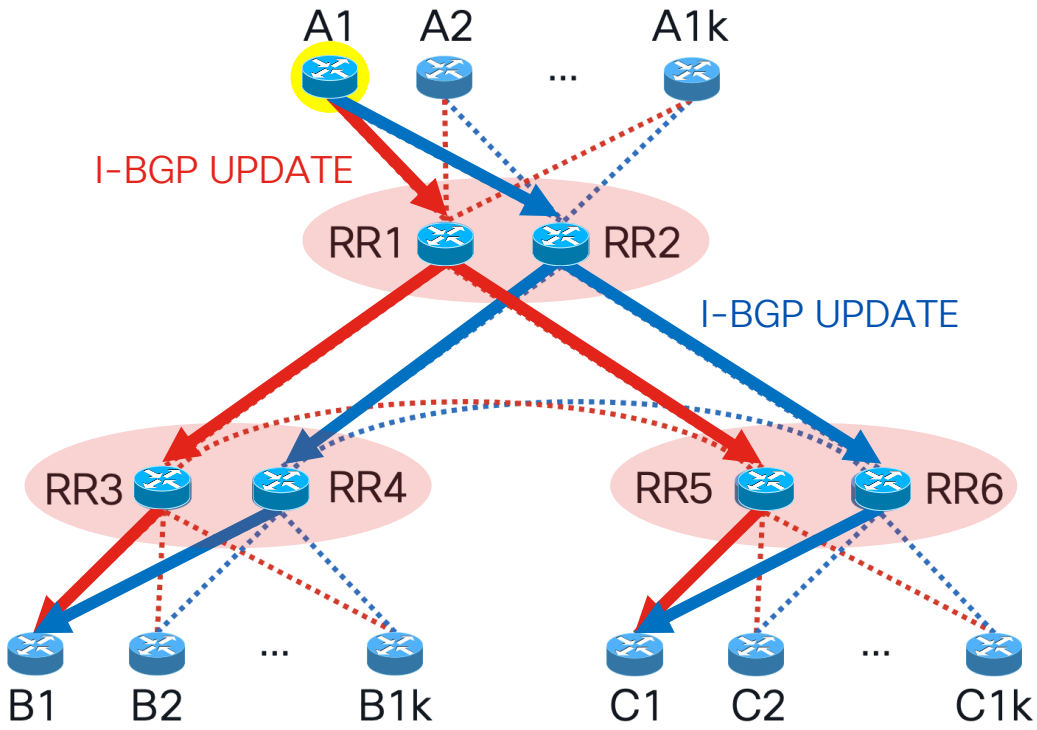


I-BGP peering (blue-plane)

I-BGP peering (red-plane)

Multi-Cluster BGP RR Design: Multi-Plane

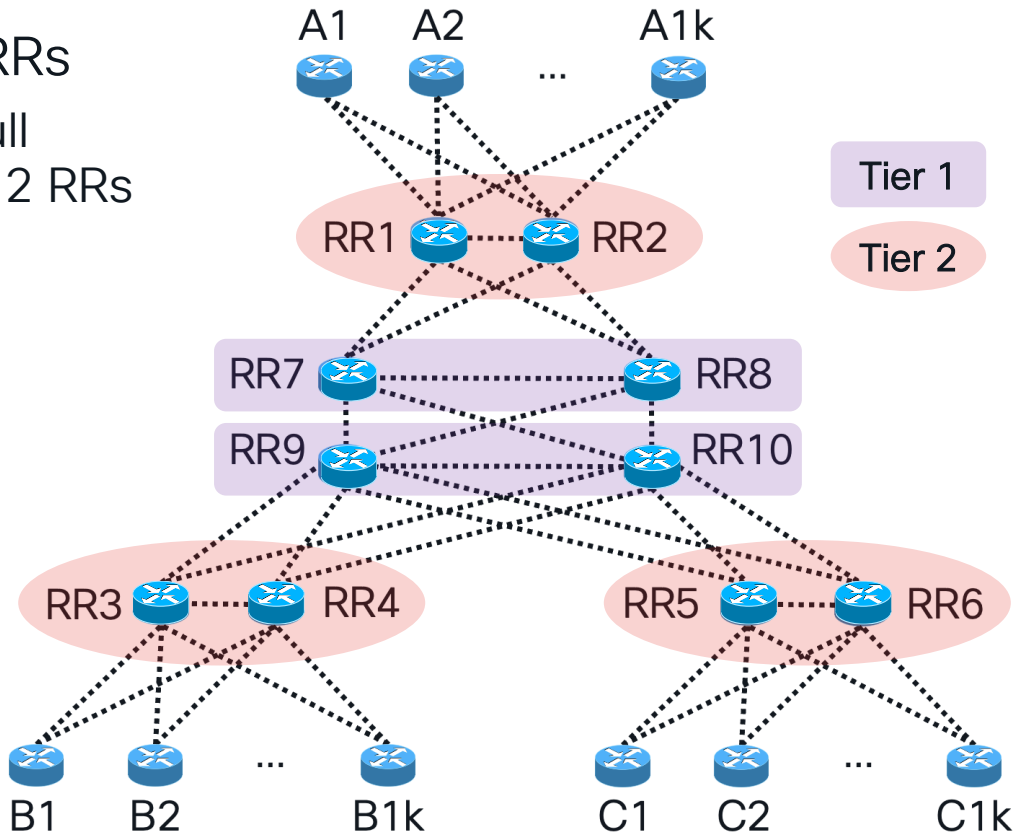
- Reduces the number of:
  - BGP sessions per RR
  - BGP I/O processing per RR
  - BGP paths carried per RR
- Eases transitions during RR software upgrades



Origin of I-BGP UPDATE
I-BGP peering (blue-plane)
I-BGP peering (red-plane)
RR within a cluster
RR client

Multi-Cluster BGP RR Design: Multi-Tier

- Tier 2 RRs act as RR clients to a cluster of redundant Tier 1 RRs
  - Avoids the need for N(N-1)/2 full peering mesh amongst the Tier 2 RRs
- Tier 1 RRs are fully meshed



Tier 1 RR within a cluster

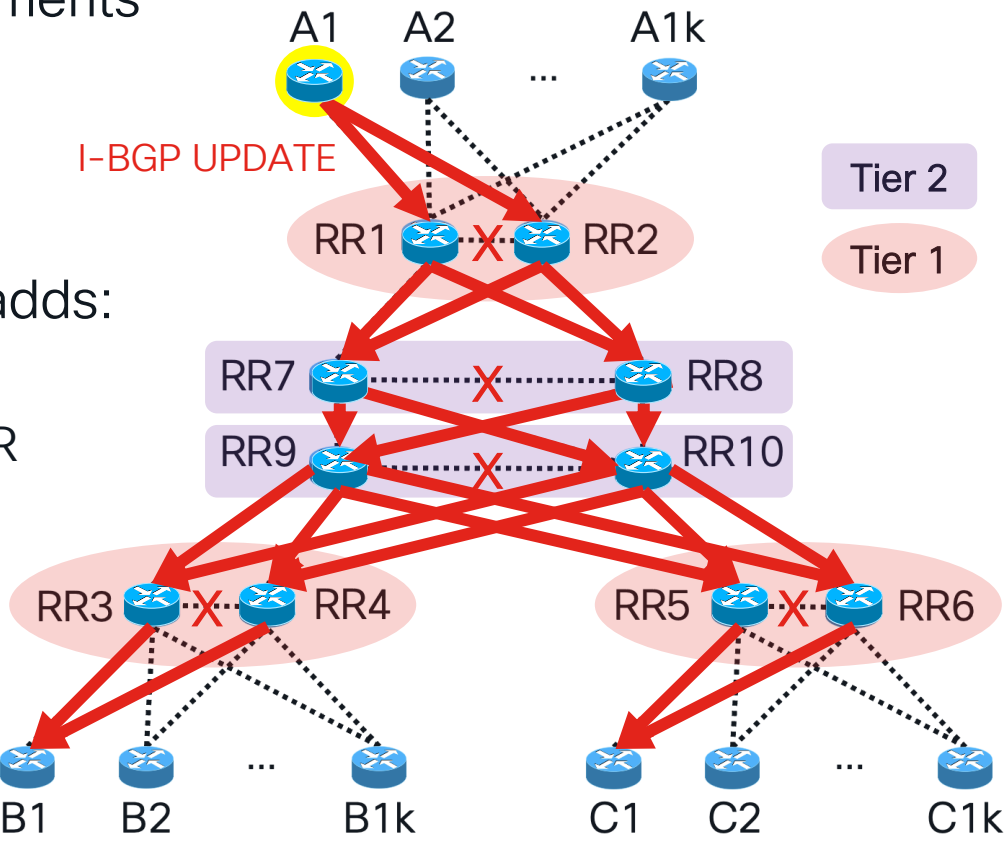
I-BGP peering

Tier 2 RR within a cluster

RR client

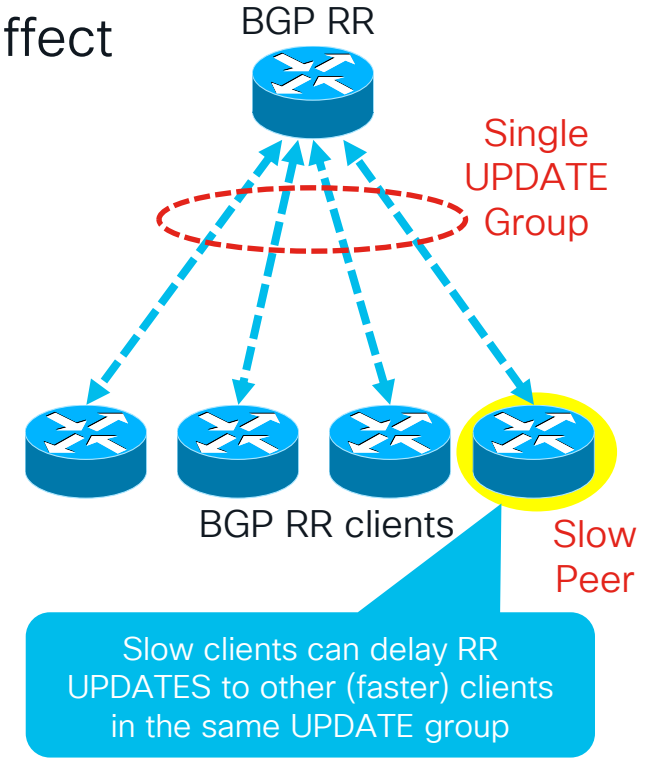
Multi-Cluster BGP RR Design: Multi-Tier

- Used for RR scaling in environments with many RR clusters
- However, the additional tier may slow BGP convergence
- Adding more redundant RRs, adds:
  - More BGP sessions per RR
  - More BGP I/O processing per RR
  - More BGP paths carried per RR
  - Similar impacts to RR clients



BGP RR Slow Peers

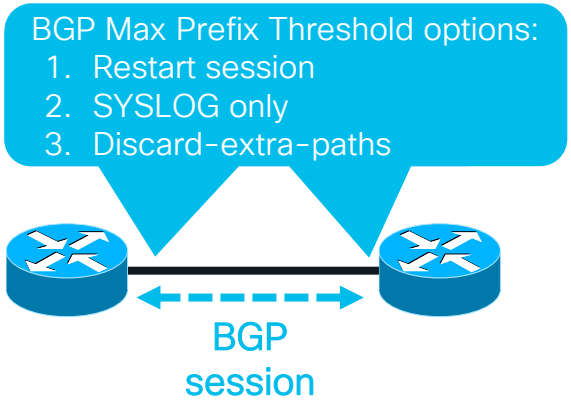
- BGP RR clients that are slow may adversely affect other clients within the same UPDATE group
  - Can delay BGP RR UPDATES to faster clients
- IOS XR 7.3.1 introduced BGP slow peer automatic isolation; Enabled by default; Not recommended
- IOS XR 7.9.1 introduced an updated version of slow peer automatic isolation; Disabled by default; Recommended
- Alternatively, configure BGP RR clients that are 'permanently slow' in their own UPDATE group separate from fast clients



Other BGP RR Considerations

- Recommend the use of Multi-instance BGP or 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 cause suboptimal routing or route oscillations per RFC 3345
  - BGP Add Paths recommended to prevent this
  - Alternatively, BGP ORR (Optimal Route Reflection)
- Recommend BGP table-policy on BGP RRs deployed outside of the forwarding plane to avoid installing BGP Internet routes in their FIBs

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 were hit, the session would be taken down
- As of IOS XR 7.3.1 the 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

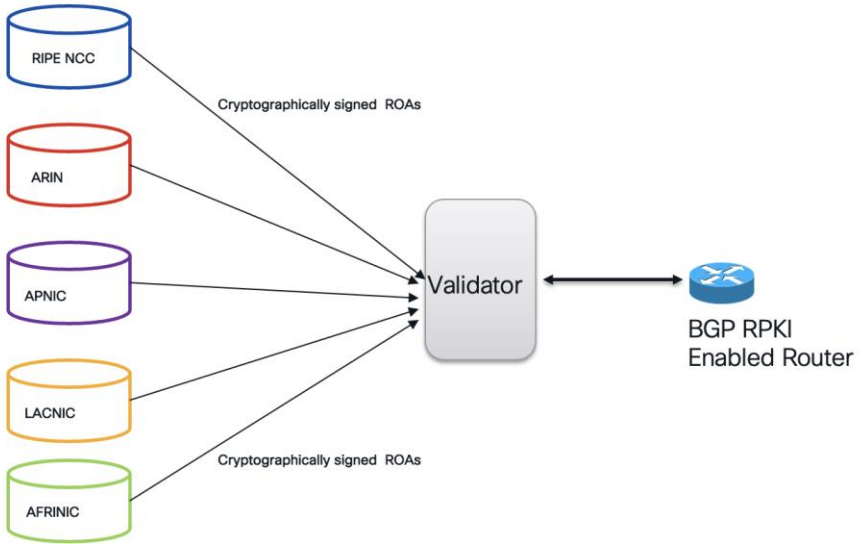
BGP Error Handling

- RFC 4271 required that a BGP speaker reset a session if it received an UPDATE with a malformed attribute over that session
- RFC 7606 revised the error handling for BGP UPDATES
  - To avoid BGP session resets whenever possible
  - Note, fatal errors still result in session resets
- In IOS XR:
  - Basic BGP error handling of less severe errors is enabled by default
  - Extended BGP error handling of rare errors is disabled by default
- Recommend: Enable extended BGP error handling

```
router bgp 65530
  update in error-handling extended ebgp
  update in error-handling extended ibgp
```

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)
- Helps reduce the risk of:
  - BGP prefix hijacking
  - BGP prefix mis-announcements



Reference

Other BGP Best Practices (1)

- 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
- 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
- Limit E-BGP peering to only explicitly configured neighbors
  - Restrict the number of dynamic BGP sessions on routers under your administration

Other BGP Best Practices (2)

Reference

- BGP session authentication using TCP Authentication Option (AO) for session integrity
- 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)
- 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)
- IETF BCP 194 (RFC 7454: BGP Operations and Security) if providing Internet service

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 Route Target Constrain (RTC) filter changes can cause a lot of churn, so use RTC only where it can dramatically reduce the number of L3VPN routes to update and store
  - Also, RRs should advertise RT-filter default to clients while RR clients send specific RT-filters to the RRs, not the other way around
- BGP update wait-install to postpone advertising UPDATES until the RIB confirms that BGP routes have been installed

IGP Hardening / 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

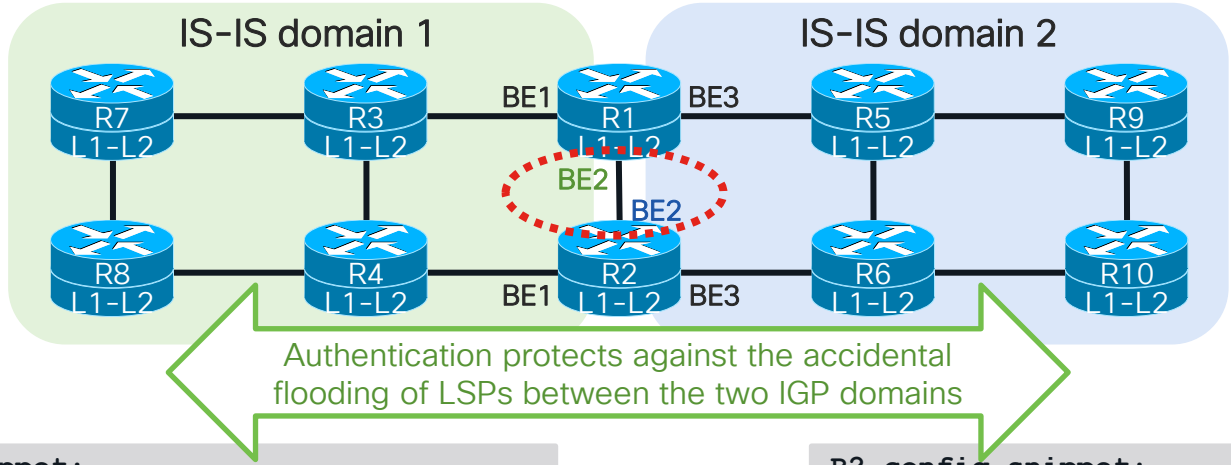
```
maximum redistributed-prefixes 10000 75 /* IOS XR default */
```

- 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 Hardening / 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



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:

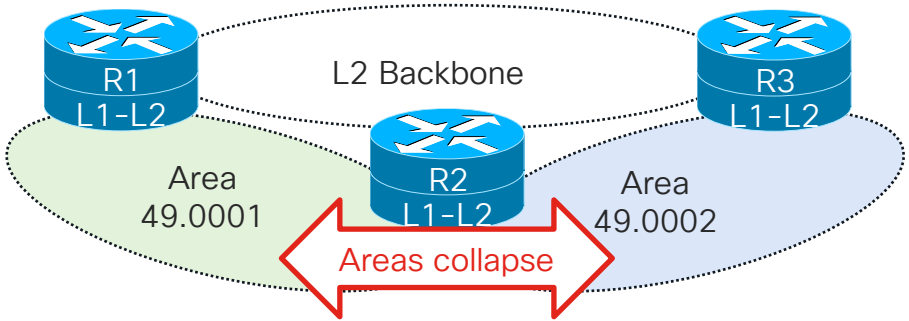
```
router is-is 2
  net 49.0002.0001.0001.0001.00
  interface Bundle-Ether 2
    address-family ipv4 unicast
    hello-password hmac-md5 encrypted [pwd2]
```

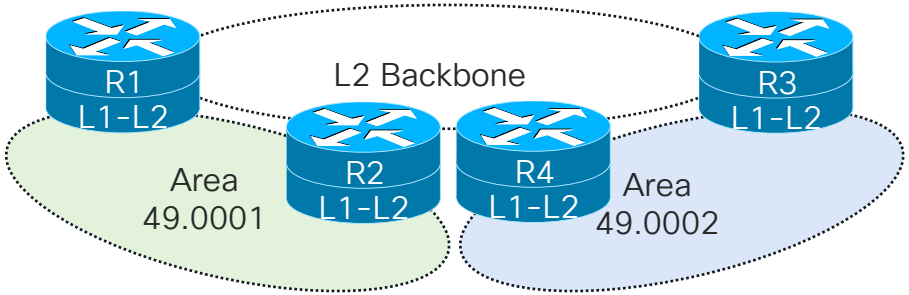
IGP Hardening / Best Practices

- 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`** 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
  - **`is-type level-1-2`** specifies that a router act as a gateway (e.g., ABR) to connect different areas (IOS XR default)
- Recommend use of L1-L2 mode on ABRs only

IGP Hardening / 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 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 routing nodes (not interfaces) to an area
- Use separate IS-IS ABRs per area

IGP Hardening / Best Practices - Scaling

- Managing IGP scale is critical to IGP and wider network stability
- Multiple well-known techniques are available to manage IGP scale:
  - Hierarchical IGP (i.e., using multiple areas)
  - Multi-domain IGP with BGP-LU (aka Unified MPLS / Seamless MPLS)
  - Multi-domain IGP with IP route summarization (SRv6 only)
  - Converged SDN Transport using SR-PCE
  - Multi-plane architecture with each plane having its own distinct IGP
- These IGP scaling techniques also offer isolation, reducing the impact (i.e., blast radius) of churn in one area/domain from affecting others

Reference

Other IGP Hardening / Best Practices (1)

- 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

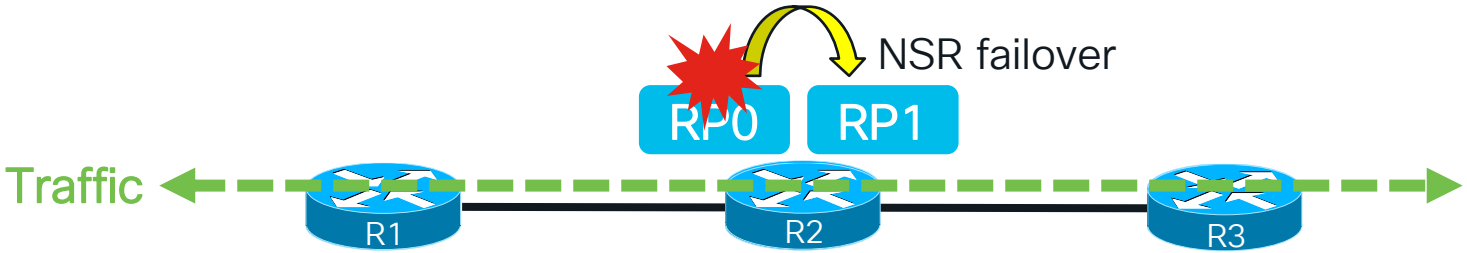
Reference

Other IGP Hardening / Best Practices (2)

- 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 for host-only and LDP label advertise to permit I-BGP next hops only (e.g., PE loopbacks) so that only transit traffic is MPLS LSP forwarded
- 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

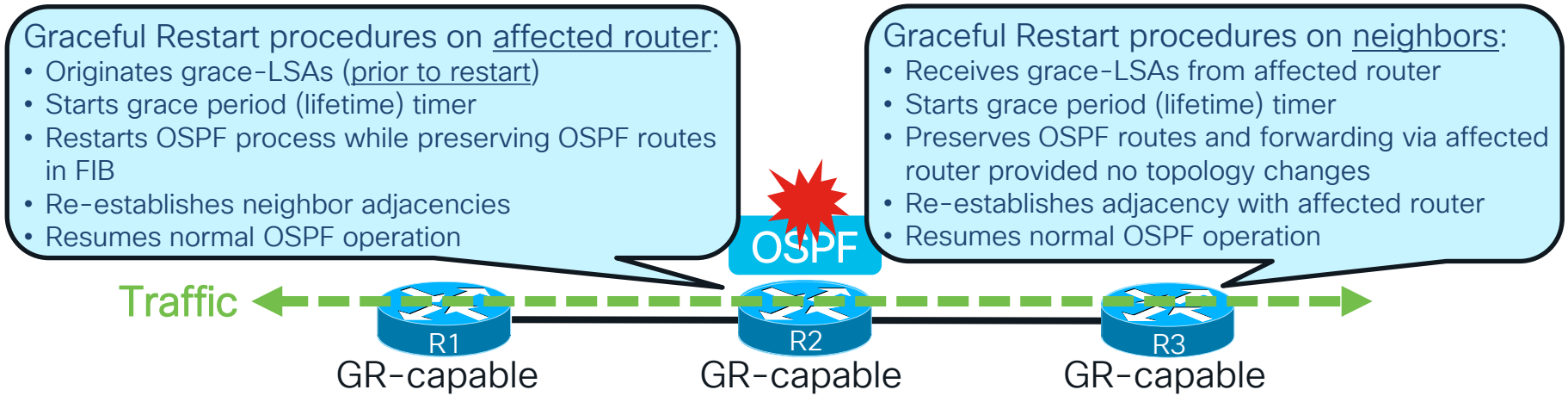
Non-Stop Routing (NSR)

- 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.



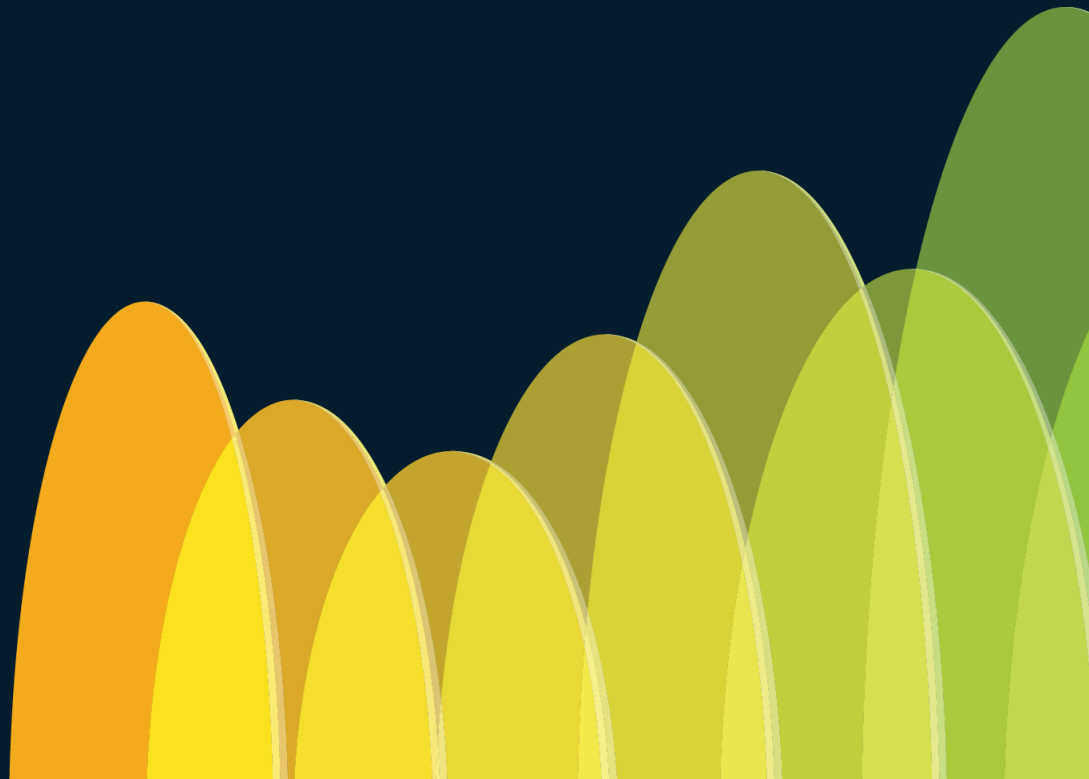
Non-Stop Forwarding (NSF) with Graceful Restart

- 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



- 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

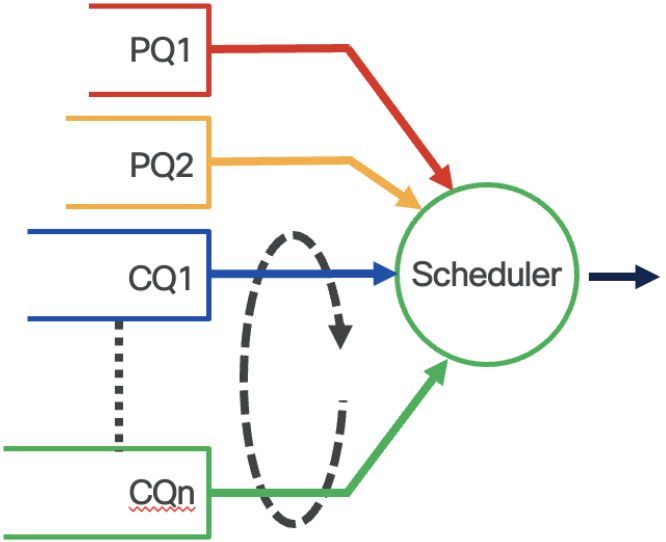
IP Forwarding Plane Best Practices



cisco Live!

QoS Considerations

- DiffServ QoS plays a vital role in ensuring high network availability
  - Isolates traffic classes and helps guarantee priority traffic
- Recommend: Designate a traffic class with ample bandwidth exclusively for control and management plane traffic to avoid drops
  - If control plane sessions / adjacencies go down, IP reachability to affected prefixes may be lost
- Recommend: On ingress to the network edge, properly mark (aka color) traffic to ensure proper packet classification and scheduling downstream



QoS Buffer Sizing Considerations

- Be aware of reduced QoS buffering on modern routers:

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

- Proper buffer configurations are essential when migrating between platforms to avoid indiscriminate 'no buffer' packet drops during periods of congestion across multiple ports
- These modern NPUs are still considered to have deep buffers compared to DC-optimized NPUs, which do not have HBM

MPLS Label Scale Considerations

Configuration	ASR 9900	NCS 5500/5700	8000
LDP label allocation for host-routes only	Recommended	Recommended; May be Required	Recommended; May be Required
MP-BGP label allocation on PEs for L3VPNs and 6PE	Per-Prefix* Per-VRF Per-CE	Per-VRF (Required)	Per-VRF (Required)
I-BGP next hop for IPv4 and IPv6 address families	next-hop-unchanged* next-hop-self	next-hop-self (Recommended)	next-hop-self (Recommended)

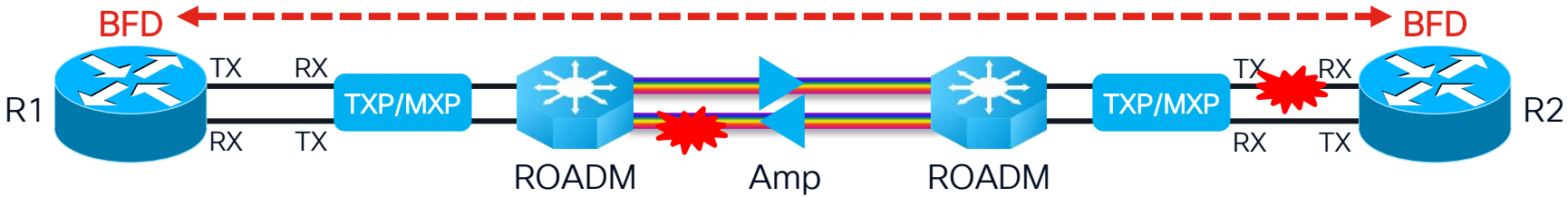
- For NCS 5500/5700 and 8000 series routers, configure settings above to help prevent Out-of-Resource forwarding issues as well as improve routing convergence
- Note, Inter-AS L3VPN Option B requires ‘Per-Prefix’ or ‘Per-NextHop-Received-Label’ BGP label allocation on ASBRs including NCS 5500/5700 and 8000 series

* Default IOS XR configuration

BFD (Bidirectional Forwarding Detection)

- 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 ms	• Relies on optical network to propagate remote faults
Routing protocol timers	>30 secs	• Lower timers affect routing scale
Ethernet CFM w/ EFD	12 ms or more	• HW offload enables fast timers
BFD	12 ms or more	• HW offload enables fast timers • Supports advanced capabilities: BFD strict mode, BFD dampening, BFD over Bundles, BFD multi-hop



BFD (Bidirectional Forwarding Detection)

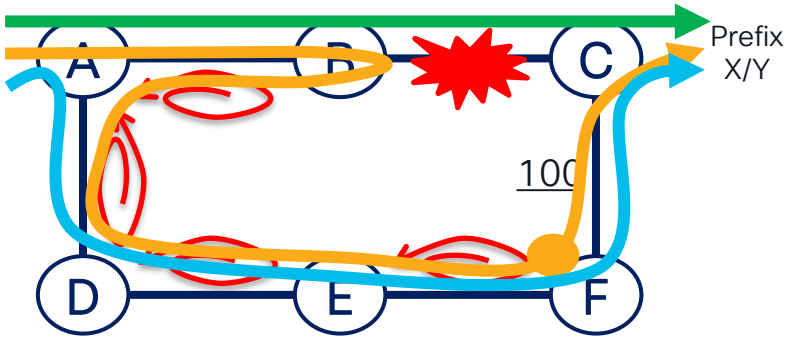
- BFD for Ethernet Bundles includes two (2) deployment modes:
  1. BFD over Bundles (aka "BoB") - defined in RFC 7130. "BoB" is where a (micro) BFD session is run on each member link within the bundle. Cisco supports two "BoB" modes: Cisco mode (pre-standard) and IETF mode. IETF mode is recommended and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

Fast Re-Route (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 (Recommended) – Topology independent, no stateful tunnels, no targeted LDP sessions (enabled by SR)

* LFA (Loop Free Alternate)

TI-LFA FRR Protection



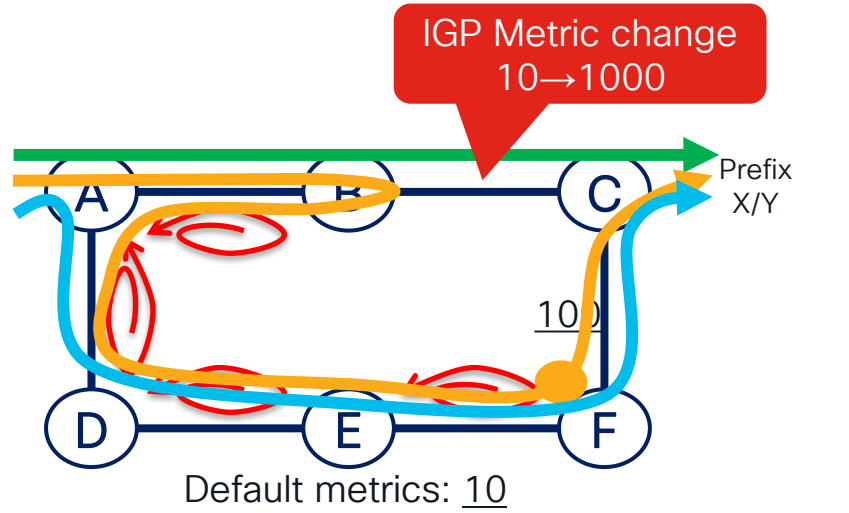
Default metrics: 10

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

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

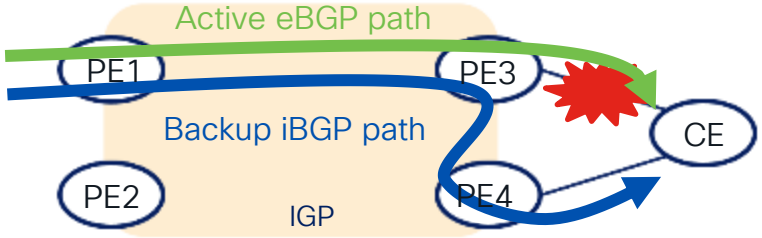
SR Microloop Avoidance

- 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]



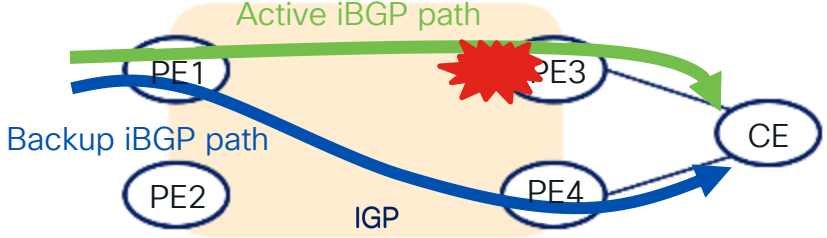
BGP Prefix Independent Convergence (PIC)

BGP PIC Edge Link Protection



- Egress PE-CE link failure (PE3-CE1) triggers BGP PIC Edge Link Protection on egress PE3
- PE3 has backup path via PE4 programmed in its FIB in advance of the failure
- PE3's convergence onto its PE4 backup path is prefix independent and pre-programmed, making it fast (~sub-second)
- BGP PIC Edge Link Protection requires Per-Prefix or 'Resilient' Per-CE label allocation

BGP PIC Edge Node Protection



- Egress PE node failure (PE3) triggers BGP PIC Edge Node Protection on ingress PE1
- Triggered by removal of PE3 from PE1's IGP database
- PE1 has backup path via PE4 pre-programmed in its FIB in advance of PE3 node failure
- PE1's convergence onto its PE4 backup path is prefix independent and pre-programmed, making it fast; however, it depends on IGP convergence and the removal of PE3

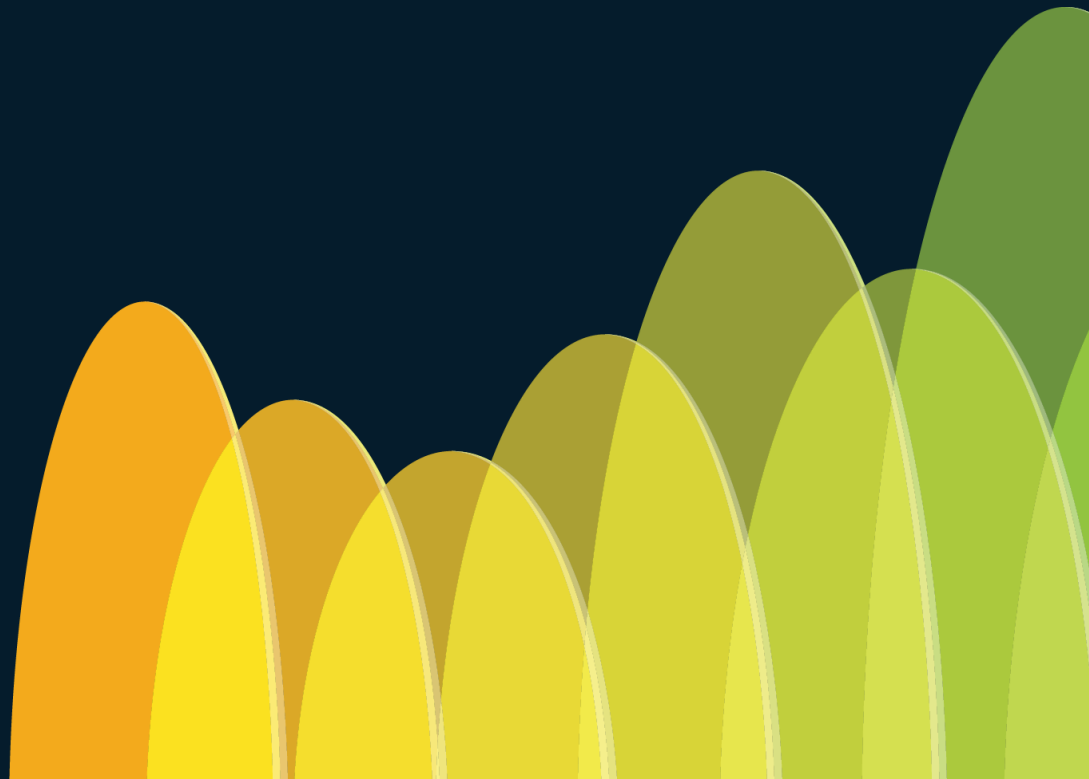
See also: IETF draft-ietf-rtgwg-bgp-pic-20

Other IP Forwarding Plane Best Practices

Reference

- Carrier Delay to delay the processing of link-up notifications (IOS XR default)
- Event Dampening to avoid churn in the control plane caused by frequent interface state changes
- Interface ACLs / packet filtering at edge – automation is key in maintaining accuracy
- If providing Internet services, Unicast RPF or ACL 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

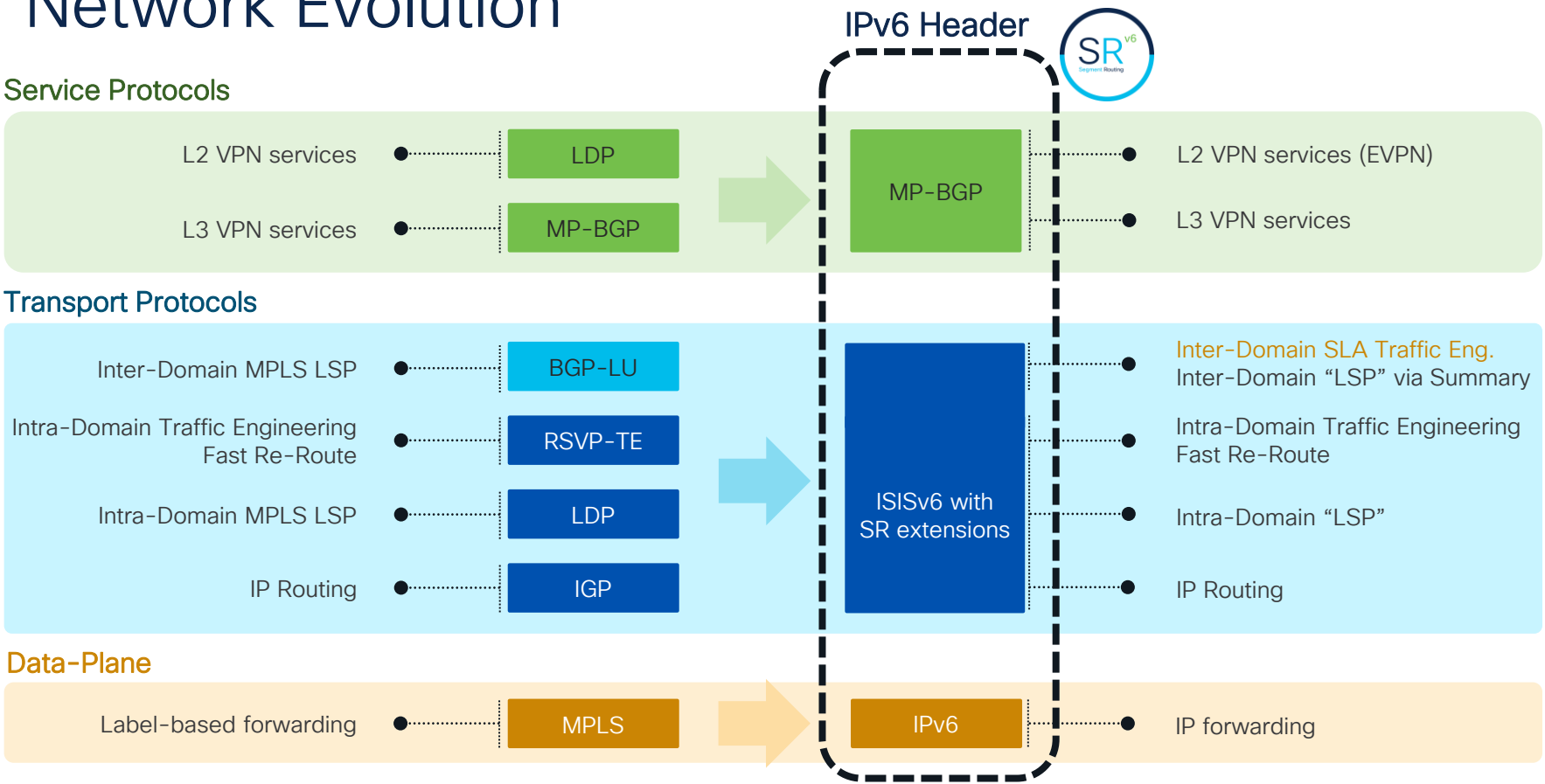


cisco Live!

Network Simplification

- A simplified network with fewer technology layers and protocols, helps to reduce complexity and potential failure scenarios as well as makes automation easier and more sustainable long-term
- Key tenets of a simplified IP/MPLS network:
  - MP-BGP as the unified service overlay control plane
  - Segment Routing (SR) as the unified forwarding plane
  - SR-TE (Traffic Engineering) for advanced control of traffic
  - Centralized (SDN) Controller for network-wide orchestration of SR-TE policies
  - BGP-LS for export of topology link-state and TE information to Controller
  - BFD for fast failure detection
  - DiffServ QoS to isolate traffic classes and guarantee priority traffic
  - YANG model-driven programmability & telemetry for management and automation
  - Routed Optical Networking for greater network efficiencies and economics

Network Evolution



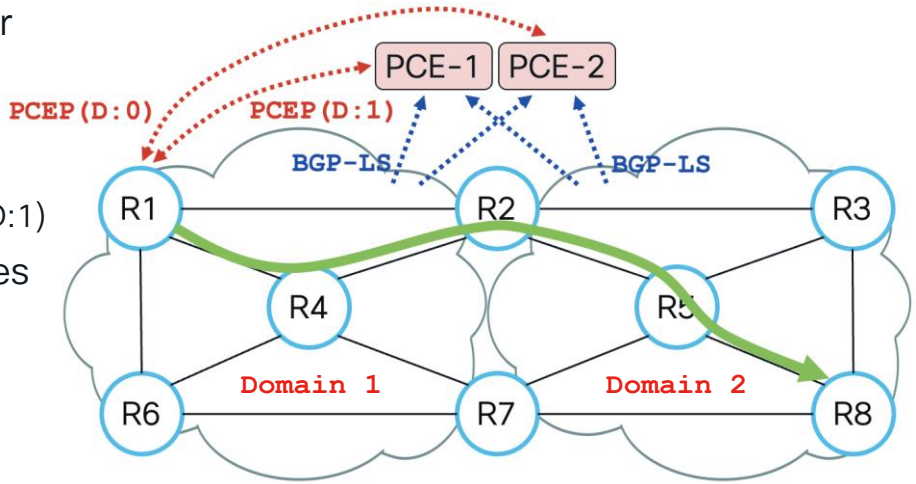
Segment Routing (SR)

- 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

- An SR PCE is only required if more information is needed than is available on a head-end; e.g., multi-domain paths or disjoint paths from different head-ends
- SR PCE leverages the well-known standardized PCE HA:
  1. 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 primary SR PCE (D:1)
  2. Primary SR PCE: (i) computes path, (ii) derives SID-list, (iii) updates path on head-end
  3. Head-end programs SID-list and reports it to its primary SR PCE (D:1) and redundant SR PCEs (D:0)
  4. Upon failure of the primary SR PCE, head-end re-delegates control to another SR PCE – No impact on SR policies or forwarding!



Reference

Crosswork Planning

Key Features

Predictive AI

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

Capacity Planning

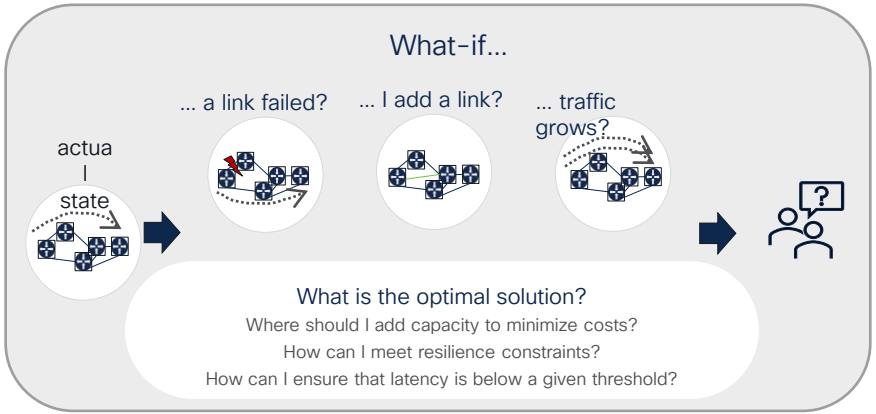
Leverage measured or simulated traffic data for accurate predictions

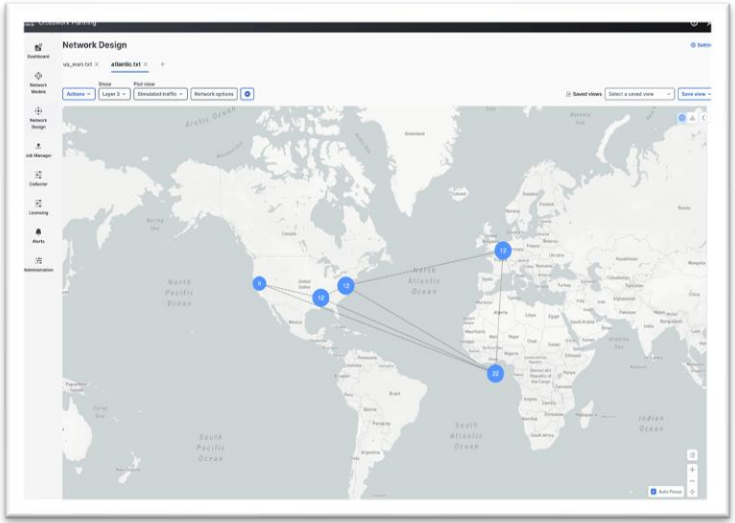
Services Optimization

Optimize network design for efficiency and reliability

Benefits

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





Summary

Summary

- Failure events may adversely affect network availability
- Architectural best practices can help mitigate the risks
- Operators need to balance risks, complexity and costs

References (1)

- 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

References (2)

- 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.

Related Breakout Sessions

Code	Title
BRKMPL-1123	Multicast with EVPN, Segment Routing, and Traffic Engineering
BRKOPT-2016	Building Transport-Grade Packet-Based Networks with Routed Optical Networking
BRKSPG-1180	Impact of AI Traffic in Transport Networks
BRKSPG-1583	5G Non-Terrestrial Networking using Cisco Converged SDN Transport
BRKSPG-2029	Designing Routed Optical Networks: IP/MPLS Considerations
BRKSPG-2063	Modernizing Private WAN Architecture for Critical Networks Infrastructure
BRKSPG-2203	Introduction to SRv6 uSID Technology
BRKSPG-2227	Design, Deployment, and Management of Next-Generation Network Fabrics
BRKSPG-2643	Differentiating B2B Services and Transport with QoE and Proactive Service Assurance
BRKSPG-2695	Resilient Networks: From Prevention to Recovery
BRKSPG-2944	Cisco 8000 Technical Update: Powered by Silicon One and IOS XR
BRKSP-2133	SP Architectural & Service Evolution with the Cisco SP NaaS framework
BRKSP-2551	Introduction to Segment Routing
IBOSPG-2000	Let's Talk Security: A Service Provider's Perspective

Acknowledgements

- Phil Bedard, Les Ginsberg, Jakob Heitz, Lokesh Khanna, Serge Krier, Peter Psenak, Marius Stoica, Ketan Talaulikar

Webex App

Questions?

Use the Webex app to chat with the speaker after the session

How

1. Find this session in the Cisco Events 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 February 28, 2025.

Fill Out Your Session Surveys

Participants who fill out a minimum of 4 session surveys and the overall event survey will get a unique Cisco Live t-shirt.

(from 11:30 on Thursday, while supplies last)

All surveys can be taken in the Cisco Events mobile app or by logging in to the Session Catalog and clicking the ‘Participant Dashboard’

Content Catalog

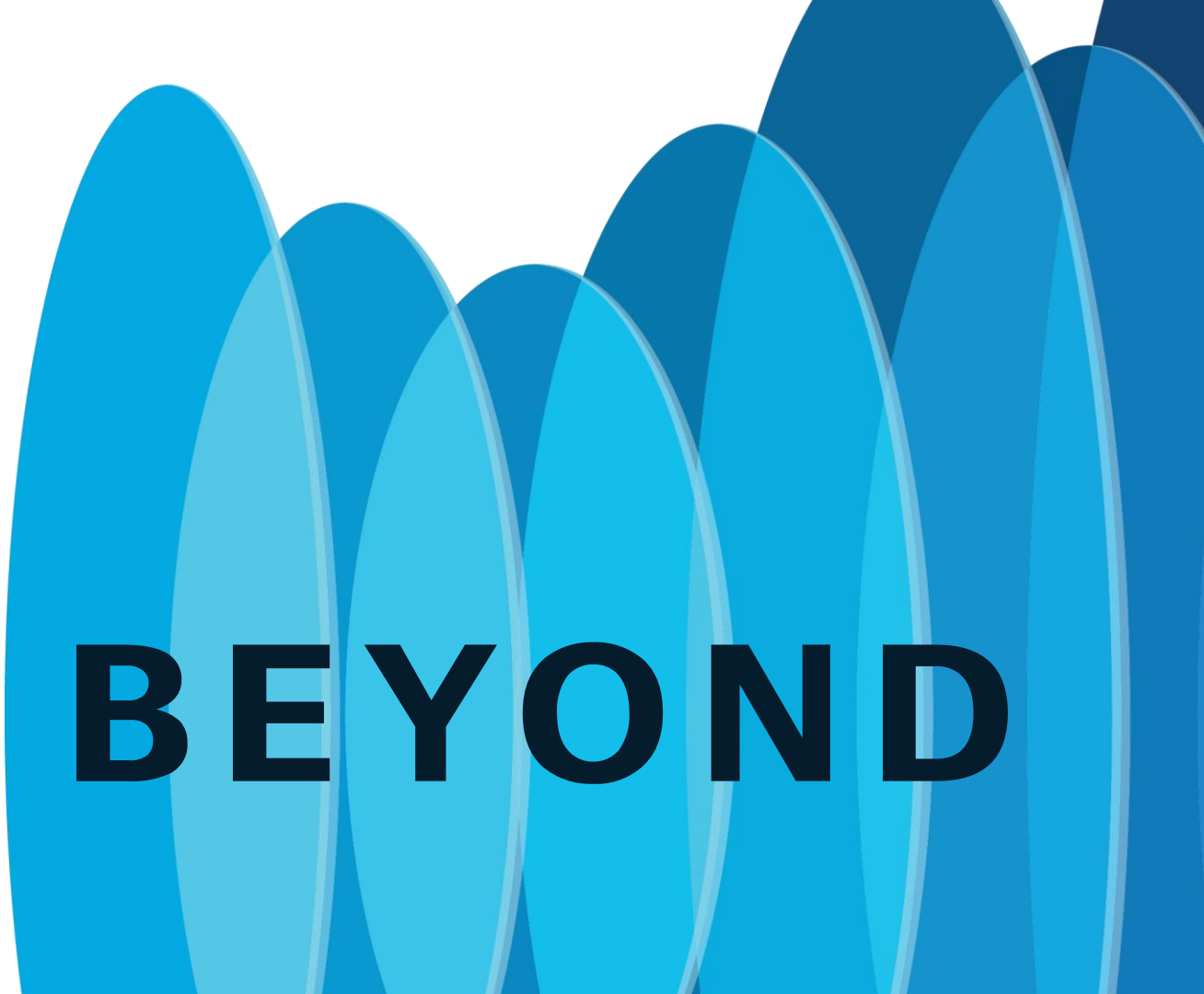
Continue your education

- 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 ciscolive.com/on-demand. Sessions from this event will be available from March 3.

Contact me at: djsmith@cisco.com


cisco
Thank you
cisco Live!

cisco Live!



GO BEYOND