



The bridge to possible

A Deep Dive into Basic and Design Best Practices for BGP and L3VPN

Mankamana Mishra, Serge Krier
mankamis@cisco.com, sekrier@cisco.com

BRKMPL-2103

CISCO *Live!*

#CiscoLive

Cisco Webex App

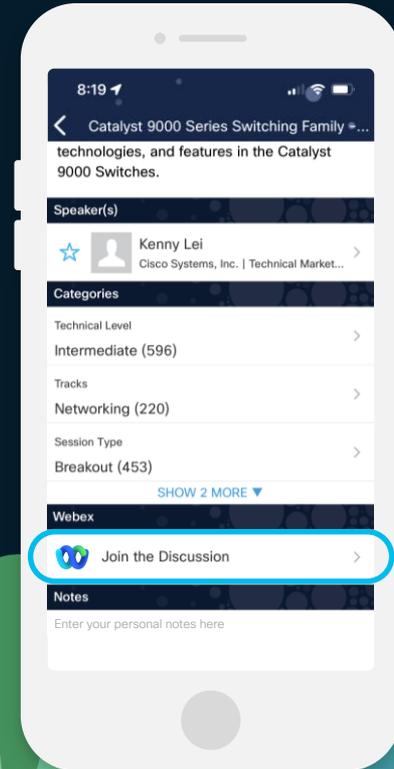
Questions?

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

How

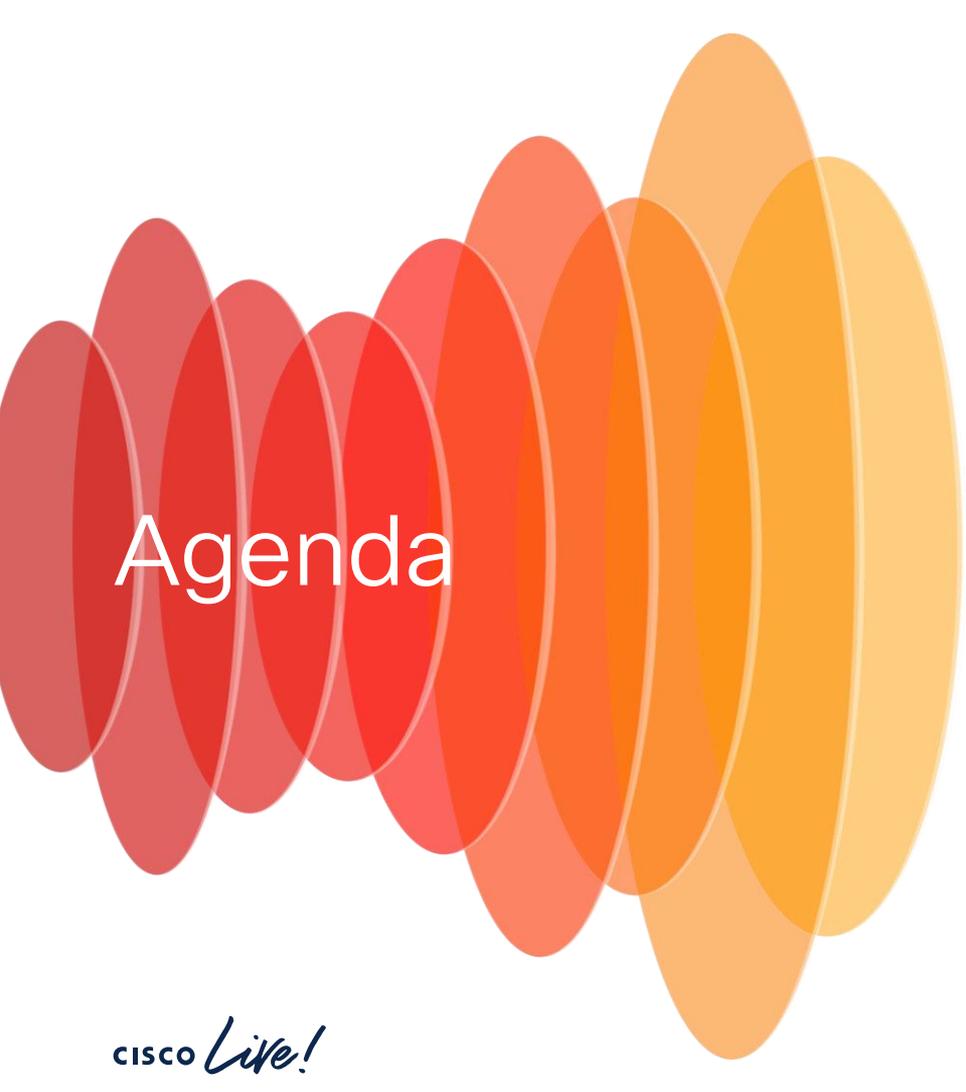
- 1 Find this session in the Cisco Live Mobile App
- 2 Click “Join the Discussion”
- 3 Install the Webex App or go directly to the Webex space
- 4 Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 7, 2024.



Before we Start

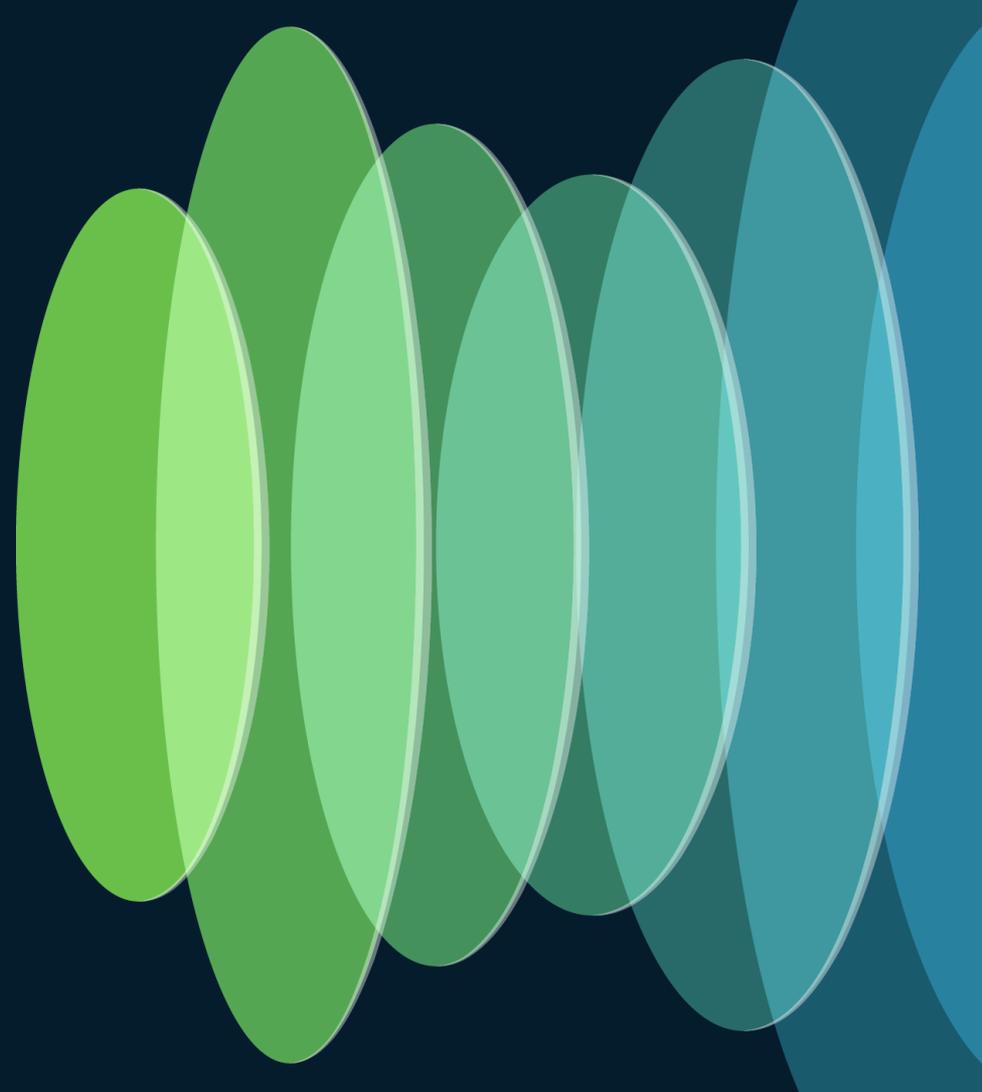
- The session is **introductory**, so basic BGP and L3VPN will be covered.
- If you are already a BGP expert, you may find it as just a revision.
- Just an hour session, so it gives overall introduction of basics of BGP.
- In interest of time, preference for Q&A the end.
- Open to discuss any BGP related topic afterwards. Please setup Meet the Engineer or reach out on Webex.



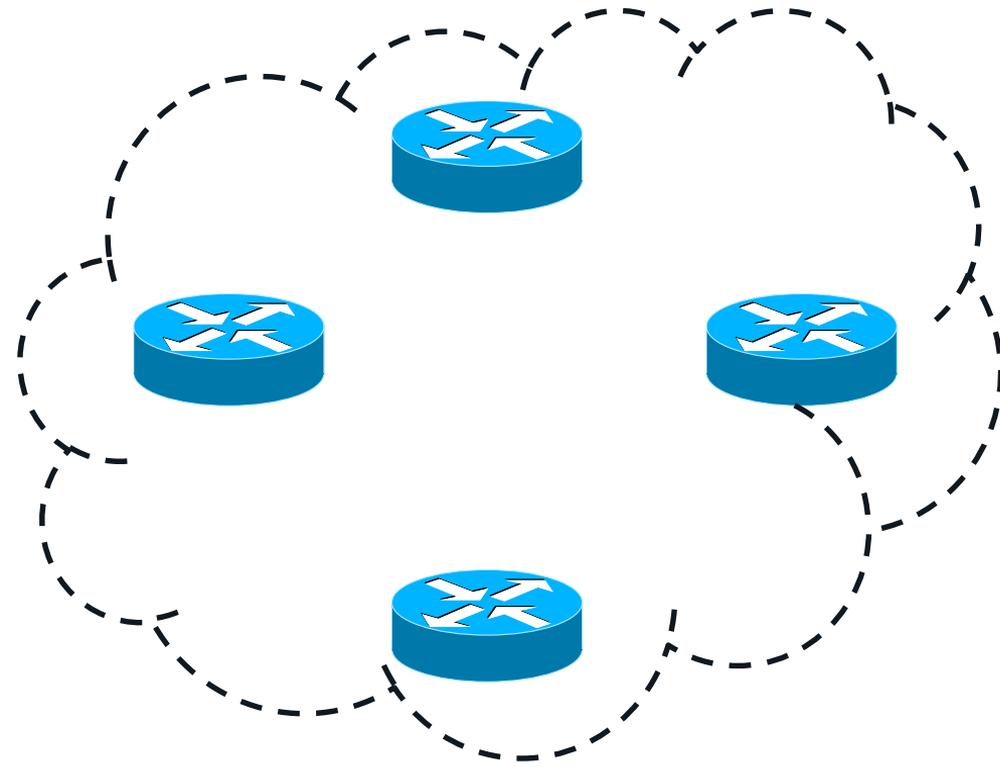
Agenda

- BGP Basics
- L3VPN Basic
- Scaling BGP networks
- BGP Security
- Best Practices

Fundamental terminology



Autonomous system (AS)

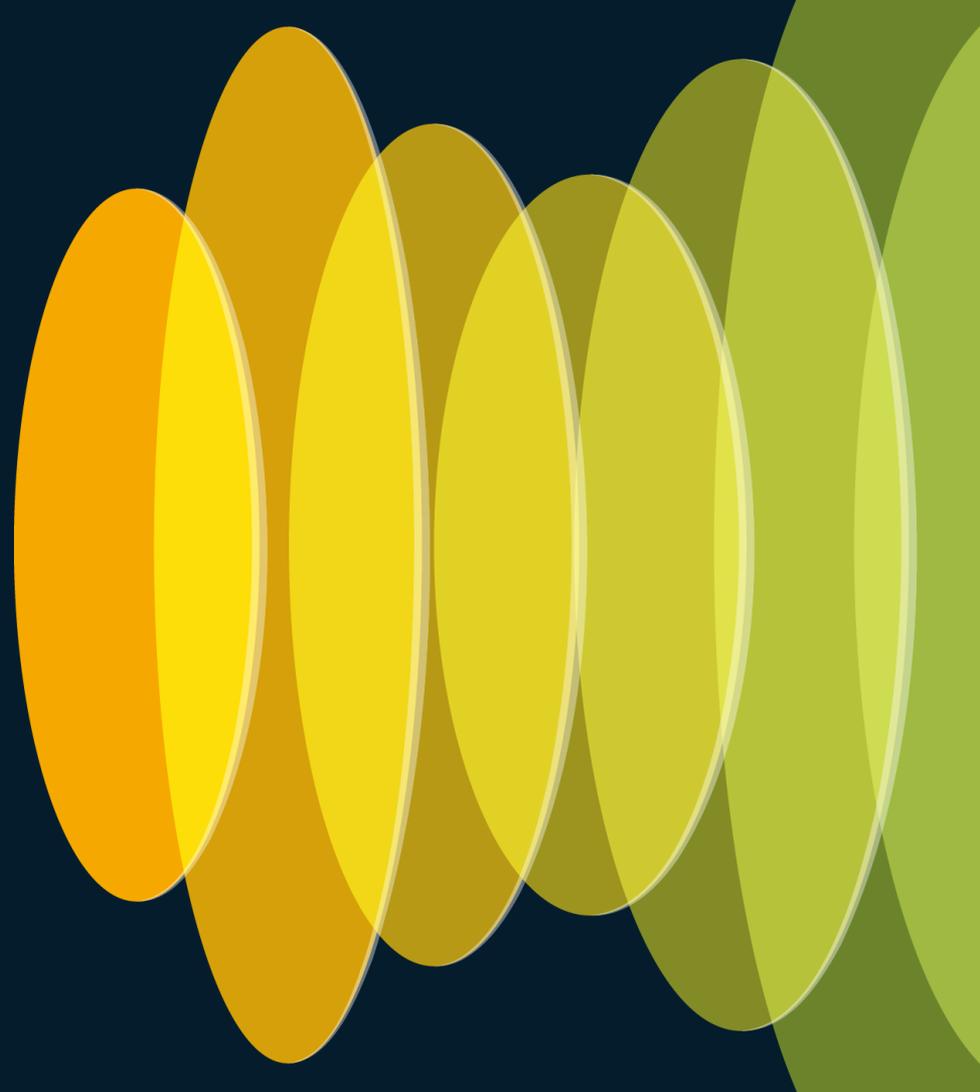


- An Autonomous System (AS) is a set of Internet routable IP prefixes belonging to a network or a collection of networks that are all managed, controlled, and supervised by a single entity or organization.
- An AS utilizes a common routing policy controlled by the entity

ASN (Autonomous System Number)

- The AS is assigned a globally unique 32bits number (0-4294967295) identification number by the Internet Assigned Numbers Authority (IANA).
- The 64512 to 65535 range and above 4.2B is reserved for private use
- Remaining ASes are available by IANA for global use. Registration needed.
- Autonomous Systems were introduced to regulate networking organizations such as Internet Service Providers (ISP), educational institutions and government agencies.
- AS ownership information is publicly available.

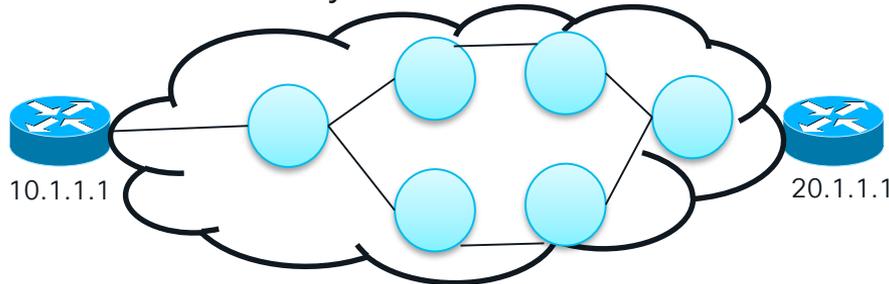
BGP Basics



Border Gateway Protocol (BGP)

- The Border Gateway Protocol (BGP) is an Inter Autonomous System routing protocol (RFC 4271).
- The primary function of a BGP speaking system is to exchange network reachability information with other BGP systems in scalable and robust way.
- BGP is decoupled from the IGP (It runs on top of the IGP). Peering always between two routers via explicit config.
- A BGP session runs over a TCP session port 179. The peering address must be reachable through the IGP or directly connected.
- Peers do not need to be directly connected. For example in below network both BGP peer are not connected directly.

```
router bgp 100
  bgp router-id 10.1.1.1
  address-family ipv4 unicast
  neighbor 20.1.1.1
  remote-as 100
  update-source Loopback0
  address-family ipv4 unicast
```

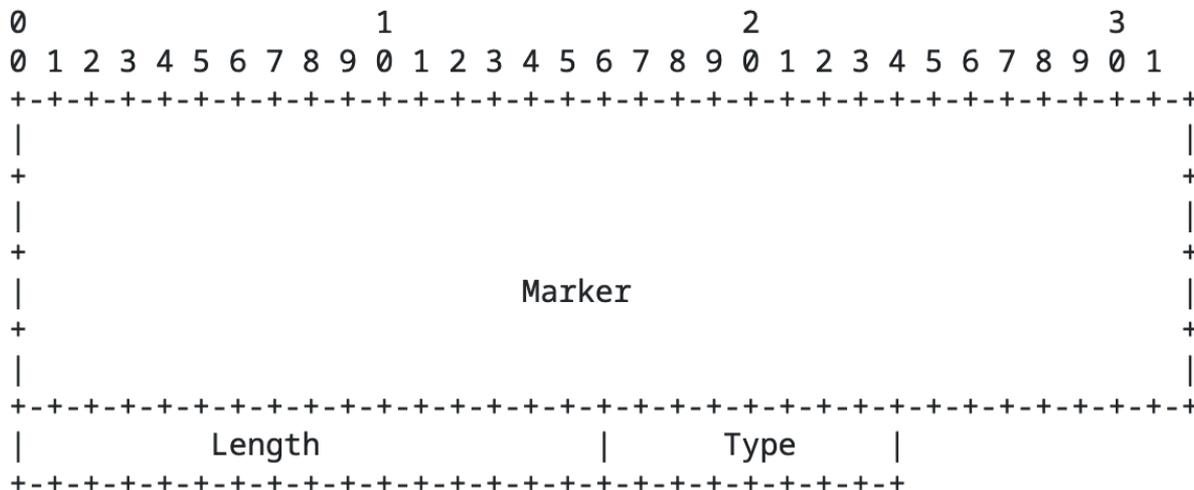


```
router bgp 100
  bgp router-id 20.1.1.1
  address-family ipv4 unicast
  neighbor 10.1.1.1
  remote-as 100
  update-source Loopback0
  address-family ipv4 unicast
```

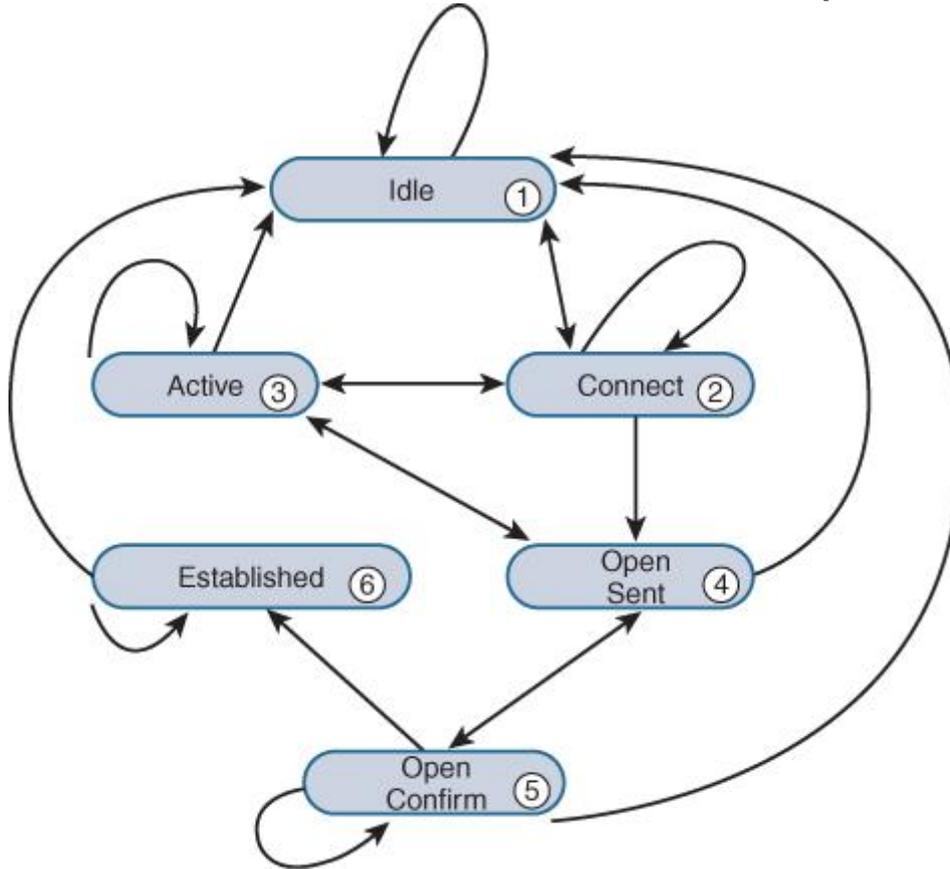
BGP Message Header

Marker: 16 times 0xFF

Type: 1 = OPEN, 2 = UPDATE, 3 = NOTIFICATION,
4 = KEEPALIVE, 5 = ROUTE-FRESH



BGP connection state (Finite State Machine)



1. **Idle:** BGP detects the start event. Tries to initiate TCP connection with peers and listens for other TCP connections.
2. **Connect:** BGP is waiting for the TCP connection to be completed
3. **Active:** BGP is trying to acquire a peer by listening for, and accepting, a TCP connection.
4. **Open Sent:** Open sent message has been sent and is waiting for peers open message
5. **Open confirm:** Waiting for peers keep alive or notification.
6. **Established:** Session ready to use and ready to exchange routes

BGP NLRI (Network Layer Reachability Information)

- NLRI is key in the local database for network/mask -> this created the parent network
 - unique NLRI received from a neighbor creates a path
 - One network linked to multiple (path/nbr)
 - attributes are properties of a (path/nbr)
- AFI/SAFI (address-family/sub-address-family)
 - Originally, BGP was intended for routing of IPv4 prefixes between organizations
 - RFC4760 added Multi-Protocol BGP (MP-BGP) capability by adding extensions called address-family identifier (AFI)
 - An address-family correlates to a specific network protocol, such as IPv4, IPv6, VPNv4, VPNv6 and additional granularity through a subsequent address-family identifier (SAFI), such as unicast , labeled and multicast.

BGP Update : NLRI update/withdraw

An UPDATE message is used to advertise feasible routes that share common path attributes to a peer, or to withdraw multiple unfeasible routes from service (see 3.1). An UPDATE message MAY simultaneously advertise a feasible route and withdraw multiple unfeasible routes from service. The UPDATE message always includes the fixed-size BGP header, and also includes the other fields, as shown below (note, some of the shown fields may not be present in every UPDATE message):

```
+-----+
|   Withdrawn Routes Length (2 octets)   |
+-----+
|   Withdrawn Routes (variable)         |
+-----+
|   Total Path Attribute Length (2 octets) |
+-----+
|   Path Attributes (variable)           |
+-----+
|   Network Layer Reachability Information (variable) |
+-----+
```

BGP Attributes (most used)

- AS Path : sequence of ASes the route has traversed
- MP_REACH (14)/MP_UNREACH(15)
- Next_Hop attribute (by itself for ipv4, otherwise in MP_REACH)
 - Do not change in iBGP updates
 - It is the forwarding gateway. We may need a recursive lookup
- Multi_EXIT_DISC : metric set on EBGP (better = lower)
- Local Preference : degree of preference (better = higher) announced to IBGP
- Community: arbitrary number for prefixes sharing common property, for policy control
- Attributes are :
 - Optional: transitive or non-transitive
 - Well-known : mandatory or discretionary

External BGP vs Internal BGP

- IBGP session is formed between neighbors within the same AS
- EBGP is formed between neighbors in different AS.
- A route learnt from an IBGP peer will not be advertised back to another IBGP by default.
 - Exception: special iBGP Route-Reflector role
- Routes received from an EBGP peer can be advertised to EBGP and IBGP peers
- Local AS is not prepended to the AS path when advertised to an IBGP peer.
- Local preference attribute is sent to the IBGP peers but not to an EBGP peer.

BGP Best Path Selection (BP)

For every NLRI, BP decides the best among all paths to install in the routing table for traffic forwarding and advertise to other BGP neighbors

1 - Prefer the path with the Highest Weight

2 - Prefer the path with the Highest LOCAL PREF

3 - Prefer the path that was locally originated

4 - Prefer the path with the shortest AS_PATH

5 - Prefer the path with the lowest origin type

6 - Prefer the path with the lowest MED

7 - Prefer eBGP over iBGP paths

8 - Prefer the path with the lowest IGP metric to the BGP nexthop

9 - Determine if multiple paths require installation in the routing table for BGP multipath

10 - When both paths are external, prefer the path that was received first

11 - Prefer the route that comes from the BGP router with the lowest router ID

12 - prefer the path with the minimum cluster list length

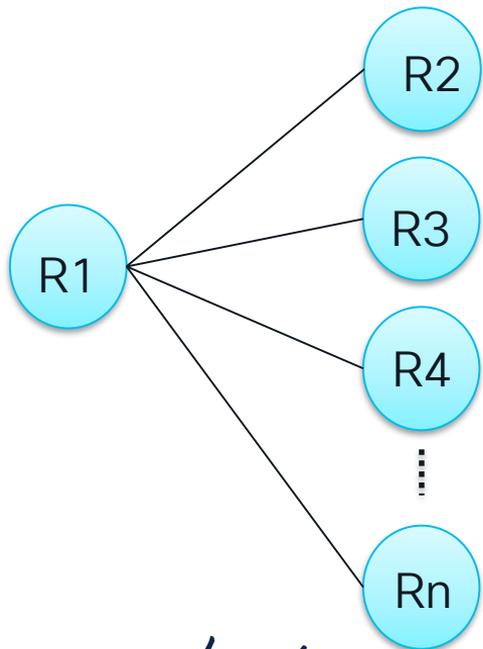
13 - Prefer the path that comes from the lowest neighbor address.

BGP update processing

- Incoming update processing
 - Receiving update from neighbor in Input Q
 - Parsing and validating update format (RFC7606)
 - Apply inbound route-policy to accept/modify/drop paths
 - Runs best path selection
- Label allocation (if needed)
- Routing Information Base (RIB) install
- Outgoing update-generation
 - Propagate best-path to other neighbors via neighbors Output Q
 - Efficiently (attribute-packing) and scalable way (update-group)
 - Incremental : only new best path or existing best path with attribute changes

Update group

- Neighbors sharing the same outbound route-policy
 - Member of the same update-group
 - Format one update, replicate it , send it to all the members



BGP formats update per neighbor

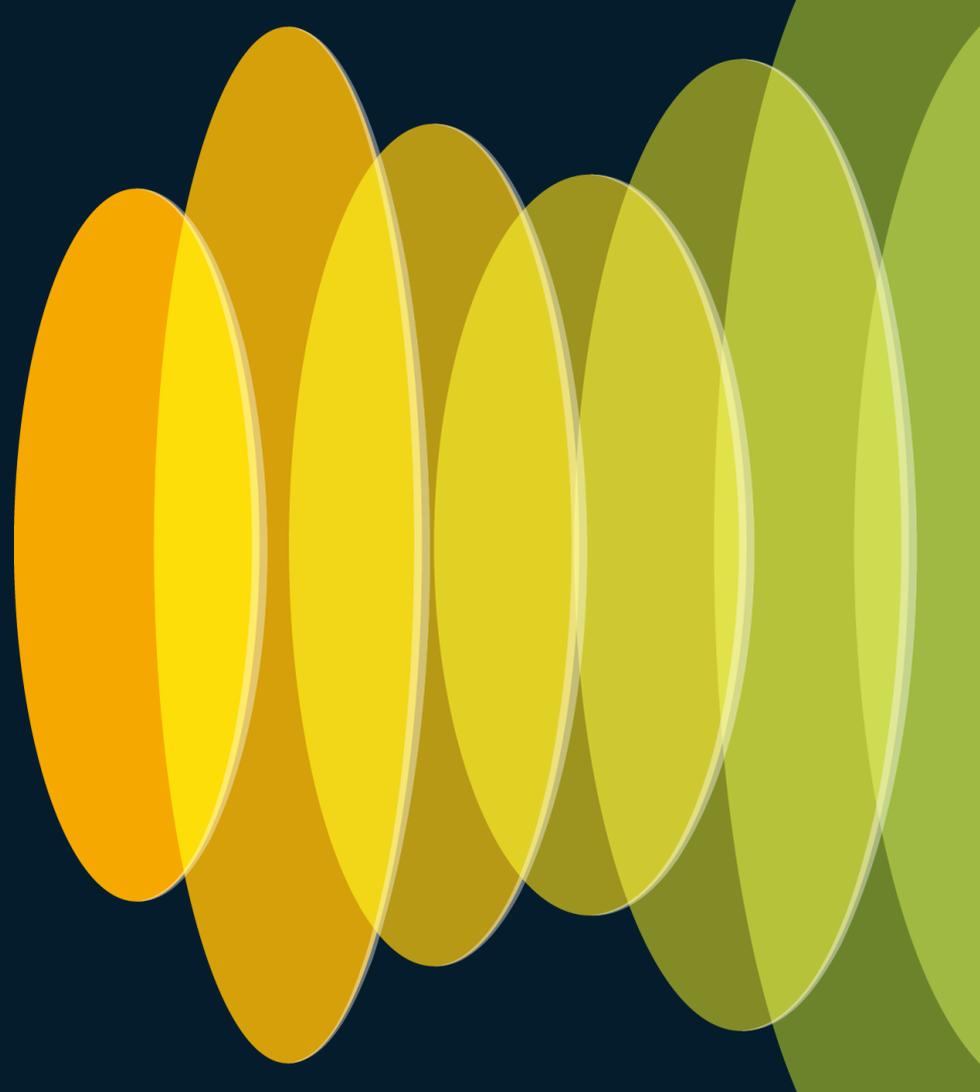


BGP formats one update per neighbor-group and replicates

BGP Update messages and processing

- Between ASes: eBGP peering
 - Loop protection: Rejects routes that have traversed our AS
- Inside the ASes: iBGP peering
 - Loop protection: Never propagate iBGP learned routes to iBGP peer

Layer 3 Virtual Private Network L3VPN

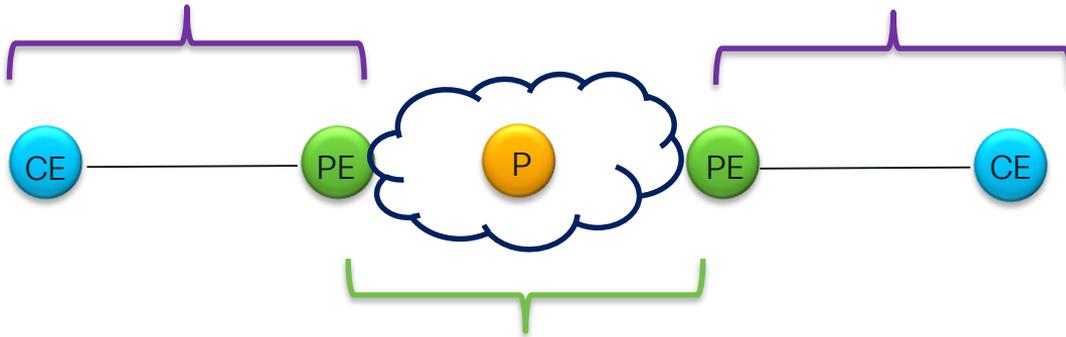


L3VPN in a Nutshell

- Customers exchange routes with the provider
 - A routing protocol runs between the customer (CE) and ISP (PE)
- The provider isolates the customer A routes from other customer B routes using Route-Distinguisher, VPN label and MPLS transport across their core.
- The routes are then provided to each customer site via an CE-PE routing protocols and the customer makes their own routing decisions
- “VPN” is isolation, not security. No inherent encryption/confidentiality in the provider's cloud

L3VPN moving parts

- The **Core**:
 - MP-BGP between PEs
 - IGP : mainly ISIS or OSPF routing between core PEs
 - Label transport options : LDP, RSVP/TE, Segment-Routing IGP, SR-TE policy
- The **Edge**:
 - **Any** routing protocol between the PE and CE



VRF Overview

- VRF = VPN Routing Forwarding instance
- Isolated routing table, kind of like a container
- Easiest to think of each VRF like a virtual routing table in a container
- Interfaces are assigned to a VRF
- Everything not in a VRF is in “the **global**” default VRF (global table)
- In MPLS-VPN **each customer** has a VRF
- VRFs for customers, default global vrf for the Provider



MP-BGP (Multi Protocol BGP)

- MP-BGP extends BGP to carry more than just IPv4 prefixes
- Introduced “address family” style configuration
 - Allows for IPv6, MPLS and other info in same BGP session between neighbors
 - When session is established the capabilities are negotiated
- No new rules, still requires full mesh or Route-Reflectors
 - RRs need to support additional capabilities
- L3VPN Relies on Extended Communities attributes
 - Extended Communities are arbitrary TLVs attached to BGP prefixes
 - Most common for vrf import/export filtering is route-target extended community.

MP-BGP: Address-Families

- Address-family “vpn4”, “ipv4 unicast vrf” introduced
- vpn4 AFI for PE to PE (label information)
- ipv4 unicast vrf for PE to CE
- Neighbor must be “activated” for each AFI supported
- Equally applies to vpn6 and ipv6

RTs and RDs: Creating the VRF

- VRFs have 3 parts:
 1. **VRF name** : case sensitive, local scope only
 2. **Route Distinguisher (RD)** : global scope
 3. **Route Target(s) (RT)**
- RD and RT are for VPNv4; RD must **always** be defined
- RD must be unique to the VRFs on the local PE
- If there is no MPLS, called “VRF-lite”

Understanding the RD

- **Route Distinguisher**

Every CE route from all VRFs are placed in a single VPNv4 table

How are routes from one VRF distinguished from another VRF?

=> By prepending the RD to the route to create a unique global **VPNv4 route**

Only used to make routes unique VPNv4 prefixes, thus VRF CE routes can overlap.

VRF IPv4 Route: 192.168.1.0/24

RD: 100:100

VPNv4 Route NLRI:

- NLRI Length: 112 bits (= 14 bytes)
- Label MPLS (3 bytes)
- RD: **100:100 (8 bytes)**
- IPv4 network: **192.168.10/24 (3 bytes)**

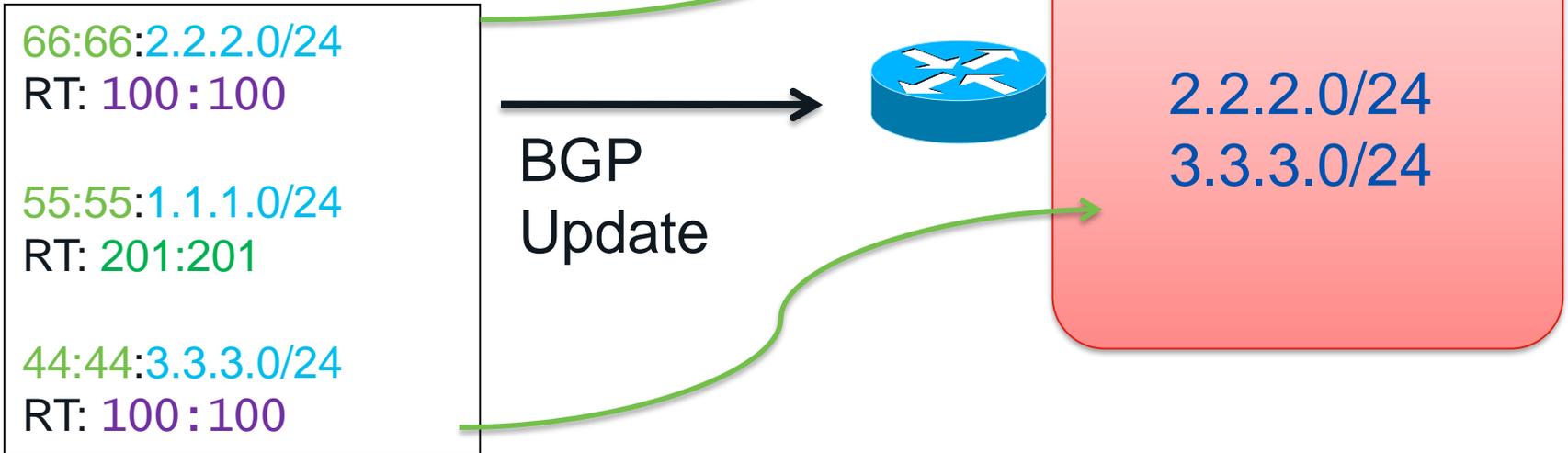
Understanding the Route-Target

Route Target

- RT is a BGP extended community (attribute attached the NLRI)
- “route-target **export**” adds the community to the **outbound** update
 - from local VRF to VPNv4
- “route-target **import**” defines which routes to bring **into the VRF**
 - from VPNv4 to local VRF
- Multiple imports and exports RT allowed
- Allows to leak routes from one/many VRF to one/many other VRFs

Route-Target in Action

```
vrf red
rd 14:23
route-target import 100:100
route-target export 201:201
```



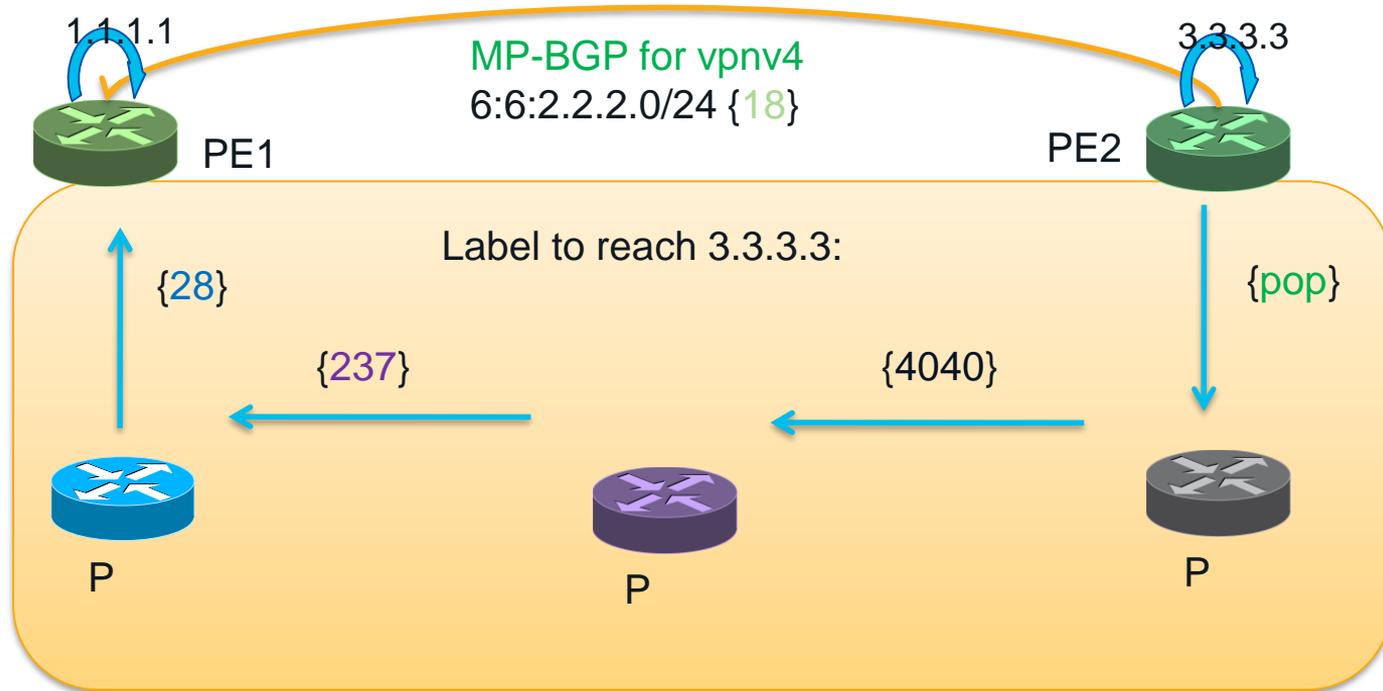
MP-BGP: Advertising CE Routes

- BGP maintains a table for each AFI (vpngv4, ipv4, vrf...)
- CE routes are placed into the vpngv4 BGP table
 - BGP routes in a vrf AFI are automatically turned into vpngv4 routes
 - If BGP is not PE-CE protocol routes must be redistributed into `ipv4 vrf AFI`
- All vpngv4 routes get an assigned service vpn label for mpls forwarding
 - When inserted into vpngv4 from vrf
 - When propagated to other BGP neighbors if nexthop is changed
- vpngv4 routes are exchanged between vpngv4 peers (PEs)

VPN Labeling

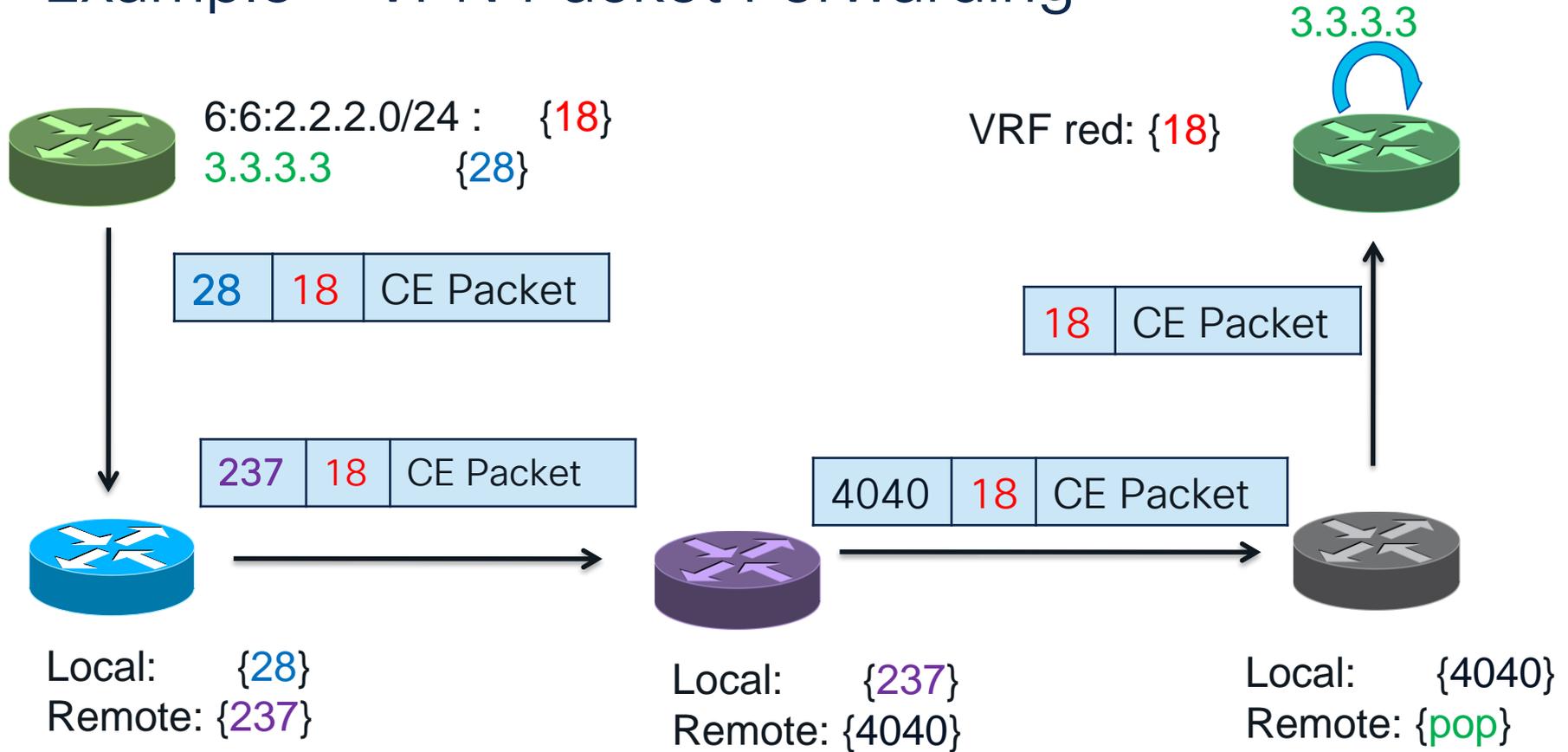
- VPN Labeling uses two labels: **Transport Label** and **VPN Label**
- **Transport label** gets packet from local PE to remote PE
- **VPN label** identifies the VRF to the remote PE
- P Routers are unaware of **VPN label**
- Within the MPLS cloud forwarding is just PE to PE
- Forwarding happens in two parts:
 1. Get the packet through the MPLS cloud to the remote PE (loopback)
 2. Put the packet in the correct VRF for a routing lookup via the **VPN label**

Example – Label Exchange

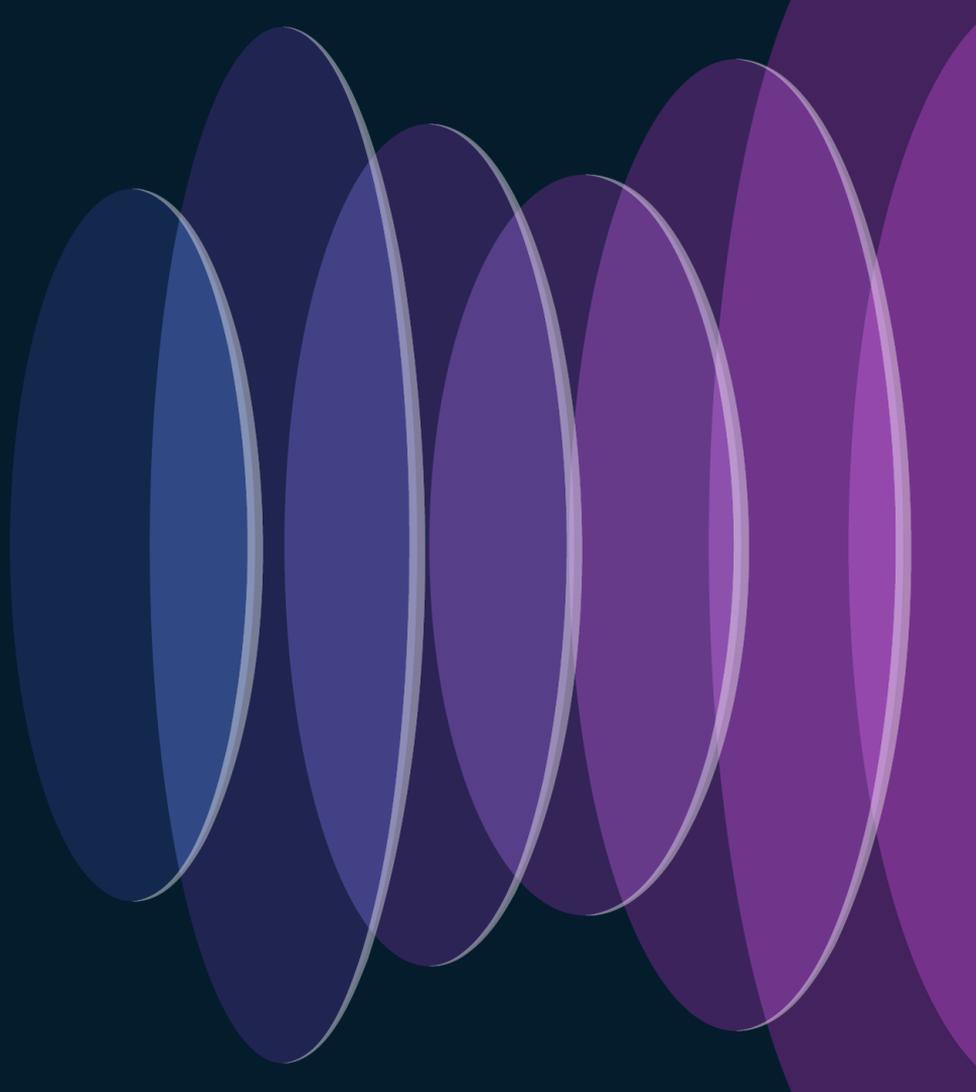


- PE2 allocates VPN label {18} for local vrf prefix
- MP-BGP exchanges VPN label {18} from PE2 to PE1
- IGP exchanges labels to reach BGP peer IP (3.3.3.3)

Example – VPN Packet Forwarding

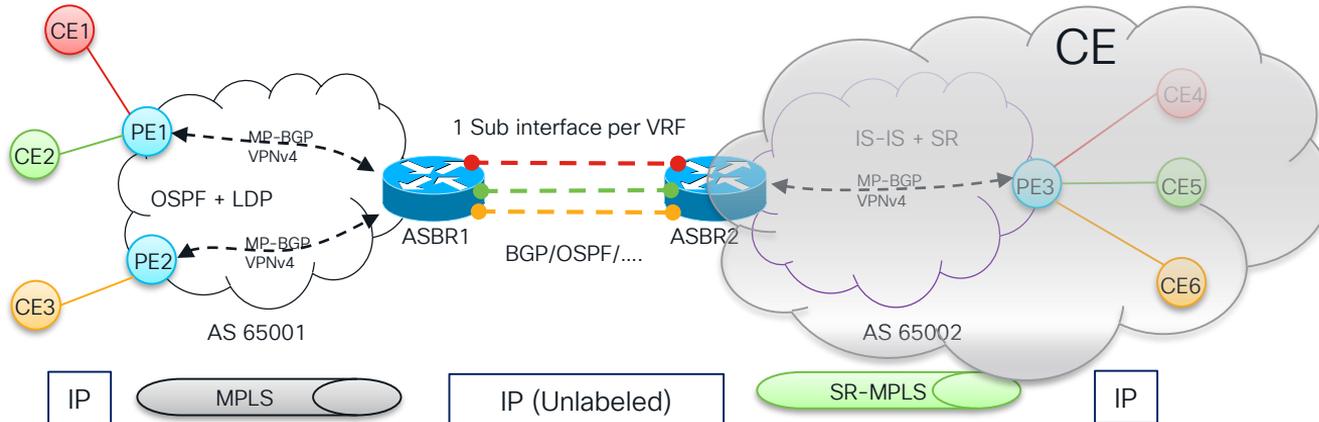


BGP Deployment scenarios L3VPN



Inter-AS Option A

- Option A is the simplest of the interconnection options.
- The AS Border Router (ASBR) of each AS defines an interface or sub-interface per VRF. Once defined, the ASBR will instantiate the VRF assigning the sub-interface to the VPN. This needs to be done per VPN requiring Inter-AS service.
- The sub-interfaces facing the other AS doesn't transport labeled traffic, only regular IP traffic. In order to exchange routing information with the remote ASBR, any routing protocol can be used.
- From the ASBR1 point of view, the remote AS is seen like any other regular CE device.



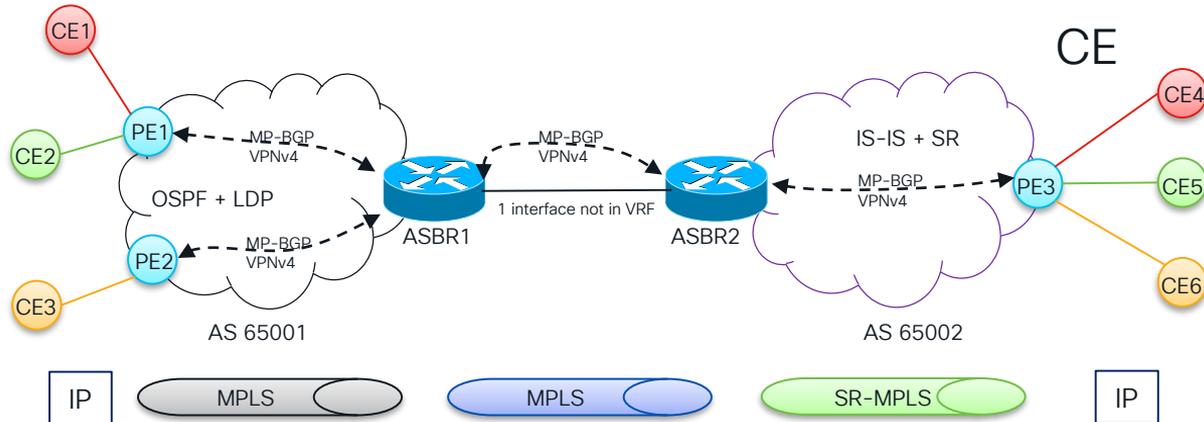
Inter-AS option A

- Simplicity
- Flexibility
- Defines clear demarcation points between MPLS L3 VPN service providers
- Easy of deployment
- IP access-list can be used for traffic filtering

- Poor scalability

Inter-AS Option B

- The Option B is the second option covered in RFC 4364 for interconnecting sites of VPN customers connected to different autonomous systems
- Inter-AS Option B tries to avoid the operational complexity needed to set up a new VPN customer with inter-as connectivity *by moving complexity*. The new procedure partially solve scalability problems but introduces some new ones we didn't have with Option A.
- There is no need to configure one VRF per-VPN customer demanding interconnection. The ASBRs should be directly connected and perform the route exchange using a single interface (physical or logical) not assigned to a VRF.



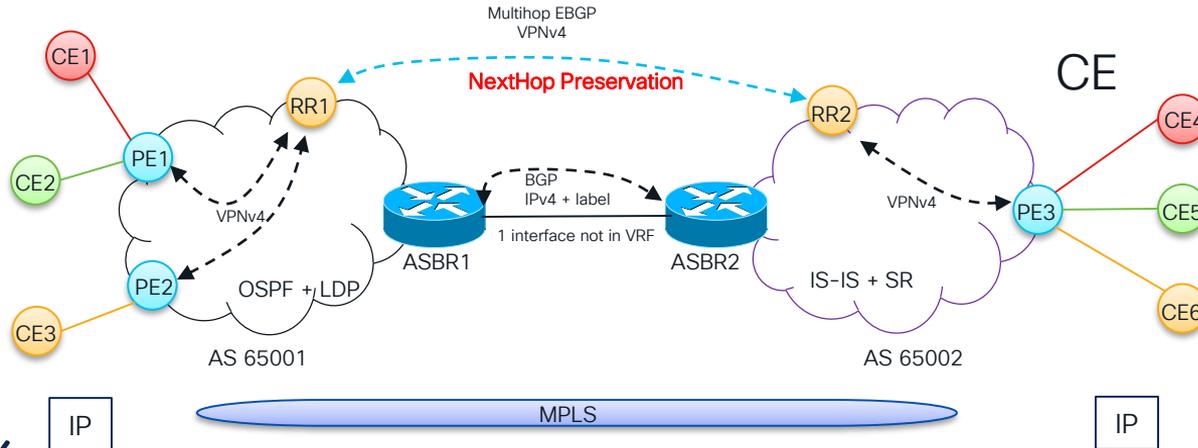
Inter-AS option B

- Enhances scalability
- Simplifies deployment

- Diffuse demarcation points due to aggregation (interface)
- It's hard to enforce IP filtering policies
- Stronger trust relationships between providers
- Additional security measures.

Inter-AS Option C

- Inter-AS Option C is the third option for interconnecting multi-AS backbones covered in RFC 4364. It's the most scalable option of the three so far and it has its own applicability scenarios that we must be aware of to apply this design properly.
- the ASBRs don't carry any of the VPN routes. ASBRs only take care of distributing labeled IPv4 routes of the PEs within their own AS.
- To improve scalability, one MP-EBGP VPNv4 session transports all VPN routes (external routes) between PEs or RR. In the case of using RR to exchange the external routes, the next hop of the VPNv4 routes must be preserved.
- The ASBR use EBGP to exchange the internal PE routing information between AS (internal routes). These internal routes correspond to the BGP next-hops of the external routes advertised through the multi-hop MP-EBGP session between PEs or RRs. The internal routes advertised by the ASBRs can be used to establish the MP-EBGP sessions between PEs and allows for LSP setup from the ingress to the egress PE.



Inter-AS option C

Scalability

- The ASBRs do not store external routing information
- Resource conservation as the external information is not duplicated on the ASBRs.
- The RRs already store the routes.

Planes isolation

- Multi-hop EBGp VPNv4 for VPN routes
- EBGp labeled IPv4 for internal routes

Security

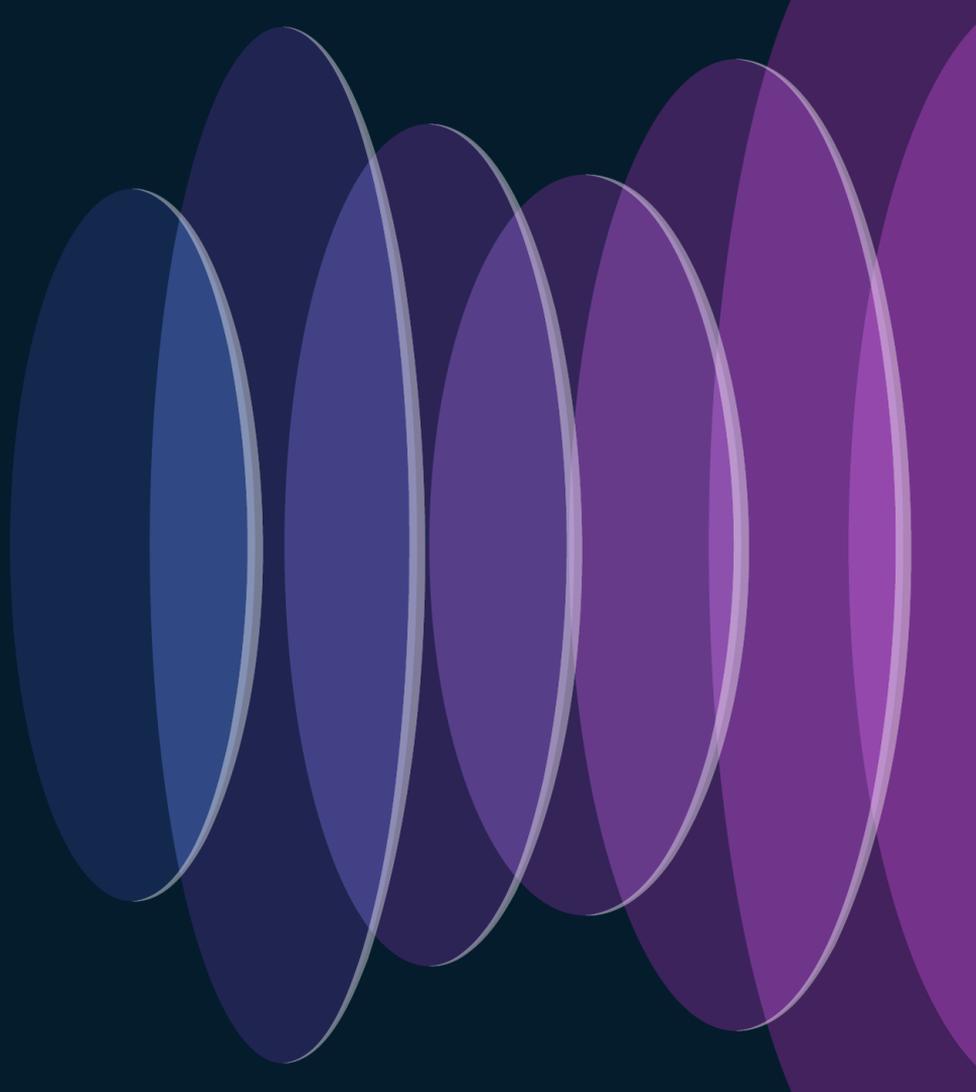
- Advertising of PE addresses to another
- not always a good option

QoS enforcement per VPN isn't possible at ASBR

- VPN context doesn't exist at ASBRs
- Not possible to perform policing, filtering or accounting with per VPN granularity at ASBR

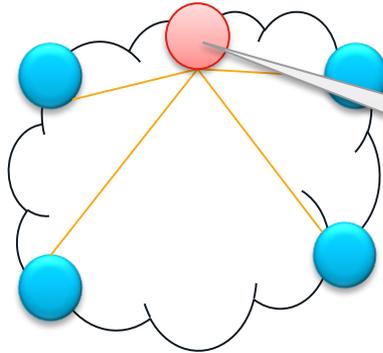
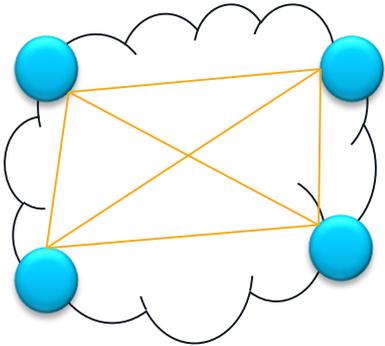
Which make this solution not a very good option when Autonomous Systems don't have a **strong trust relationship** between them

Scaling BGP networks



Route Reflector

- By default, full mesh of BGP peering is needed across all iBGP peers
 - if a new BGP speaker is in the network, config change is needed all over network
 - As the number of peers grows, the total number of sessions also grows like square
- A route reflector is a BGP speaker that will reflect routes learned from other iBGP peers.
- If RR is not in the forwarding plane , BGP table policy can be used to prevent route getting downloaded to RIB
- All routers from a peering relationship only with the Route Reflector.
 - New addition in network does not require config change at any other BGP peer

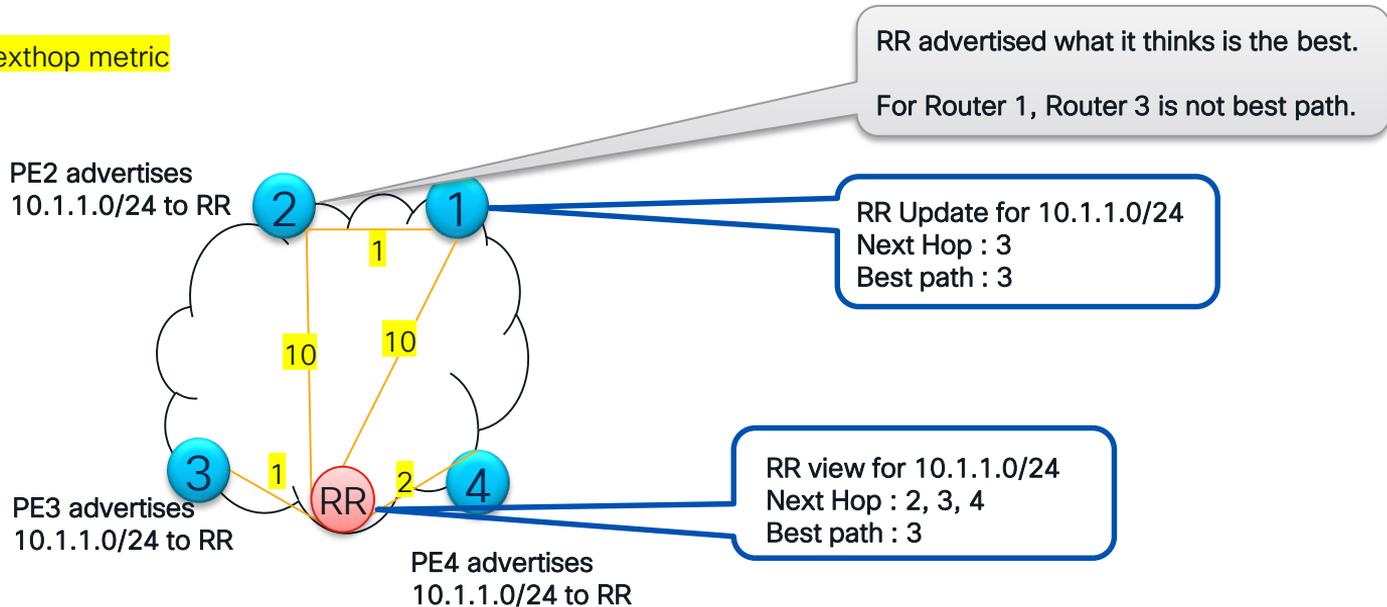


Any challenge with just having RR ?

Route Reflector – Drawbacks?

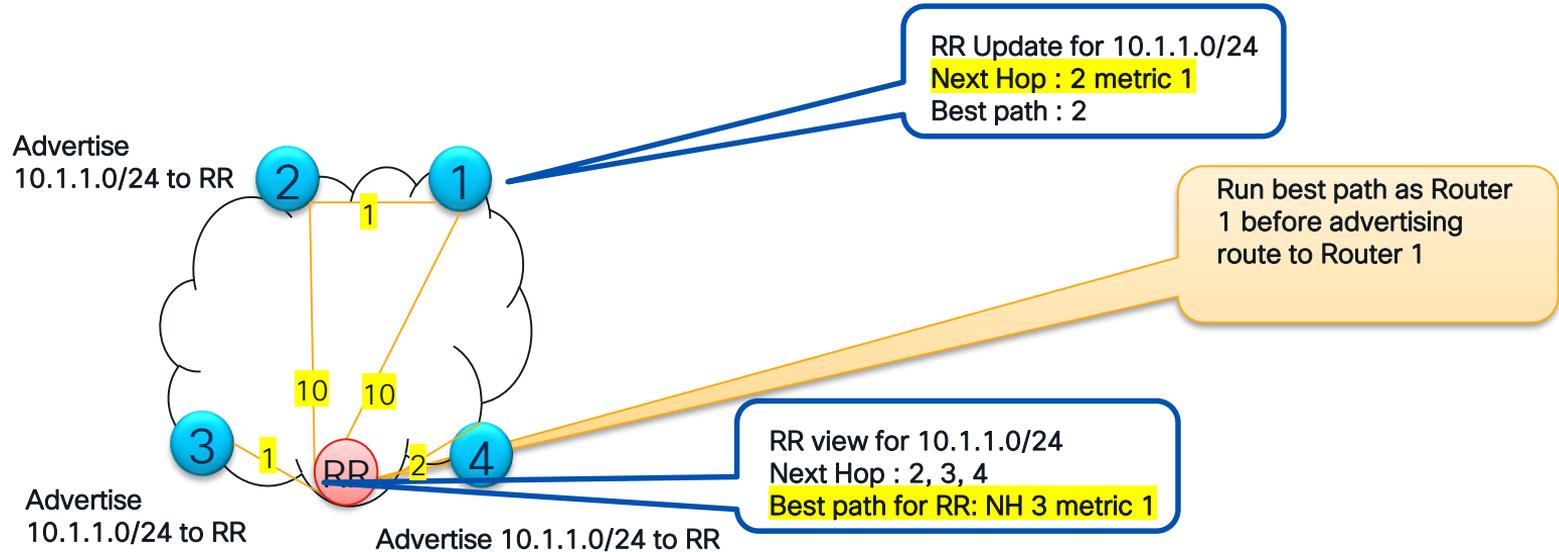
RR compute only 1 best path based on its own view of Nexthop metric, likely not the best for all clients point of view

IGP nexthop metric



Optimal Route Reflector

RR would compute best-path based on Nexthop IGP metric from client point of view.



Requirement: IGP link-state protocol (ospf/isis) for RR to compute NH metric on RR client's behalf.

Label allocation mode

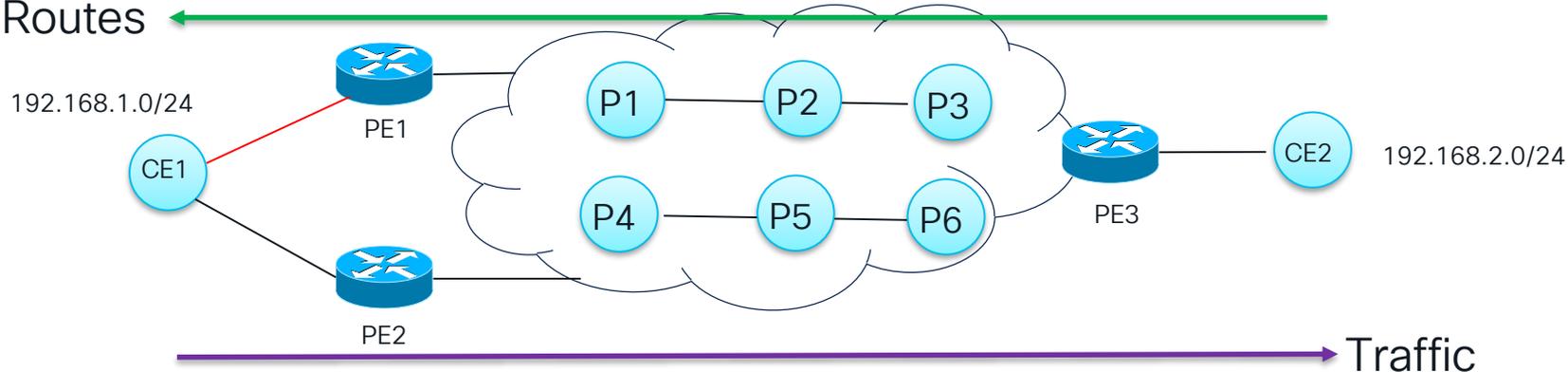
- Per vrf
 - Good for scaling, one label for all prefixes, less load-balancing, no backup-path
 - IP lookup always
- Per CE
 - Better scaling
 - MPLS forwarding
 - can be used by backup path for faster convergence
- Per prefix
 - Limited scale (1M label available in total), good for load-balancing
 - MPLS forwarding
 - can be used with backup path for faster convergence

BGP Additional path (ADD Path)

- BGP routers and route reflectors (RRs) propagate only their best path over their sessions.
- The BGP Additional Paths feature is implemented by adding a 32 bits path identifier to each path in the NLRI.
- The path identifier ID can be considered something similar to a route distinguisher (RD) in VPNs, except that a path ID can apply to any address family.
- Path IDs are unique to a peering session and are generated for each network.
- Since path ID is part of the NLRI key, it is used to prevent a route announcement from implicitly replace/withdraw the previous one.
- Increase path diversity for the RR clients , thus able to perform better/faster path selection
- Higher path scale

BGP prefix independent convergence (PIC)

Routes ←



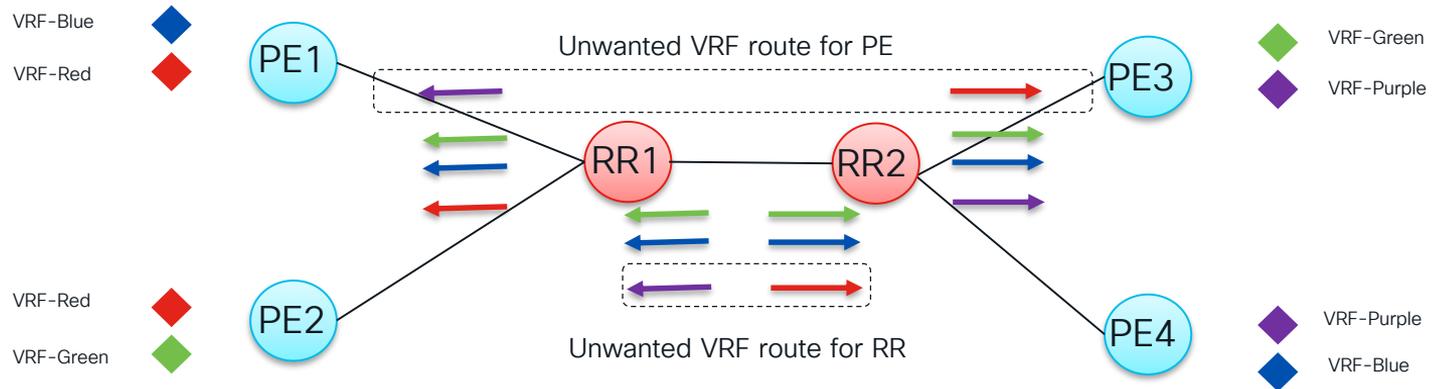
- eBGP sessions exist between the PE and CE routers.
- Traffic from CE1 uses PE1 to reach network 192.168.2.0/24 through router CE2.
- CE1 has two paths:
 - PE1 as the primary path.
 - PE2 as the backup/alternate path.
- CE1 is configured with the BGP PIC feature. BGP computes PE1 as the best path and PE2 as the backup/alternate path and installs both routes into the RIB.
- In case of CE1-PE1 link failure, traffic immediately takes alternate path in CEF (until BGP reconverge)

BGP Multipath load balancing

- A BGP routing process will install a single path as the best path in the routing information base (RIB) by default.
- The BGP multipath allows you to configure BGP to install multiple paths in the RIB for multipath load sharing.
- Load balancing over the multipaths is performed by CEF.
- CEF load balancing is done on per flow basis.
- The BGP Multipath Load Sharing for both eBGP and iBGP in an MPLS VPN allows multihomed AS and PE routers to be configured to distribute traffic across both eBGP and iBGP paths.
- By default, Equal Cost Multi Path (ECMP)
- Unequal Cost Multi Path (UCMP) weighed on nexthop igp metric also possible

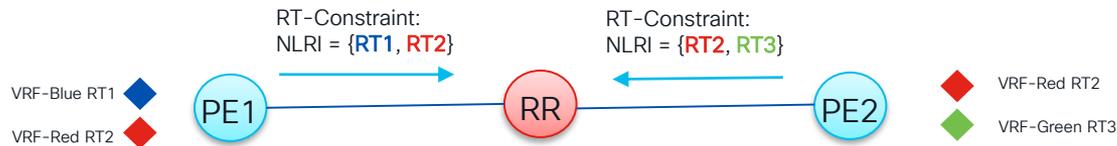
RT Constrained , why we need it ?

- Some deployments have a large number of routing updates sent from RR to PE which can require extensive use of resource
- A PE does not need routing updates for VRFs that are not locally configured; therefore, the PE determines that many routing updates it receives are "unwanted." The PE filters out the unwanted updates.

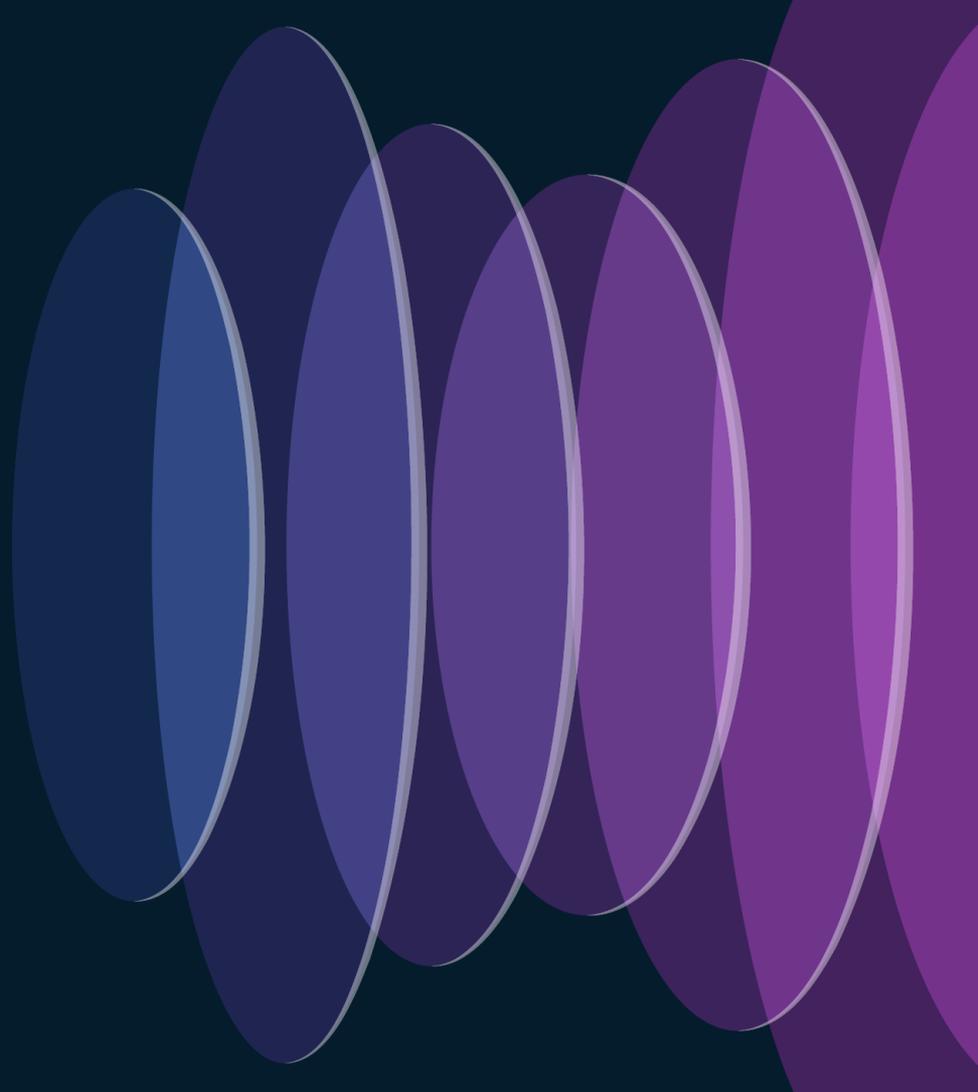


RT Constrained Route Distribution operation

- To filter out the unwanted routes described in the "Problem that BGP RT Constrained Route Distribution" the PEs and RRs must be configured with the BGP: RT Constrained Route Distribution feature (ipv4 rt-filter AF)
- The feature allows the PE to propagate RT membership and use the RT membership to limit the VPN routing information maintained at the PE and RR. The PE uses an MP-BGP UPDATE message to propagate the membership information. The RR restricts the advertisement of VPN routes based on the RT membership information it received
- This feature causes two exchanges to happen:
 - The PE sends RT Constraint (RTC) Network Layer Reachability Information (NLRI) to the RR.
 - The RR installs an outbound route filter



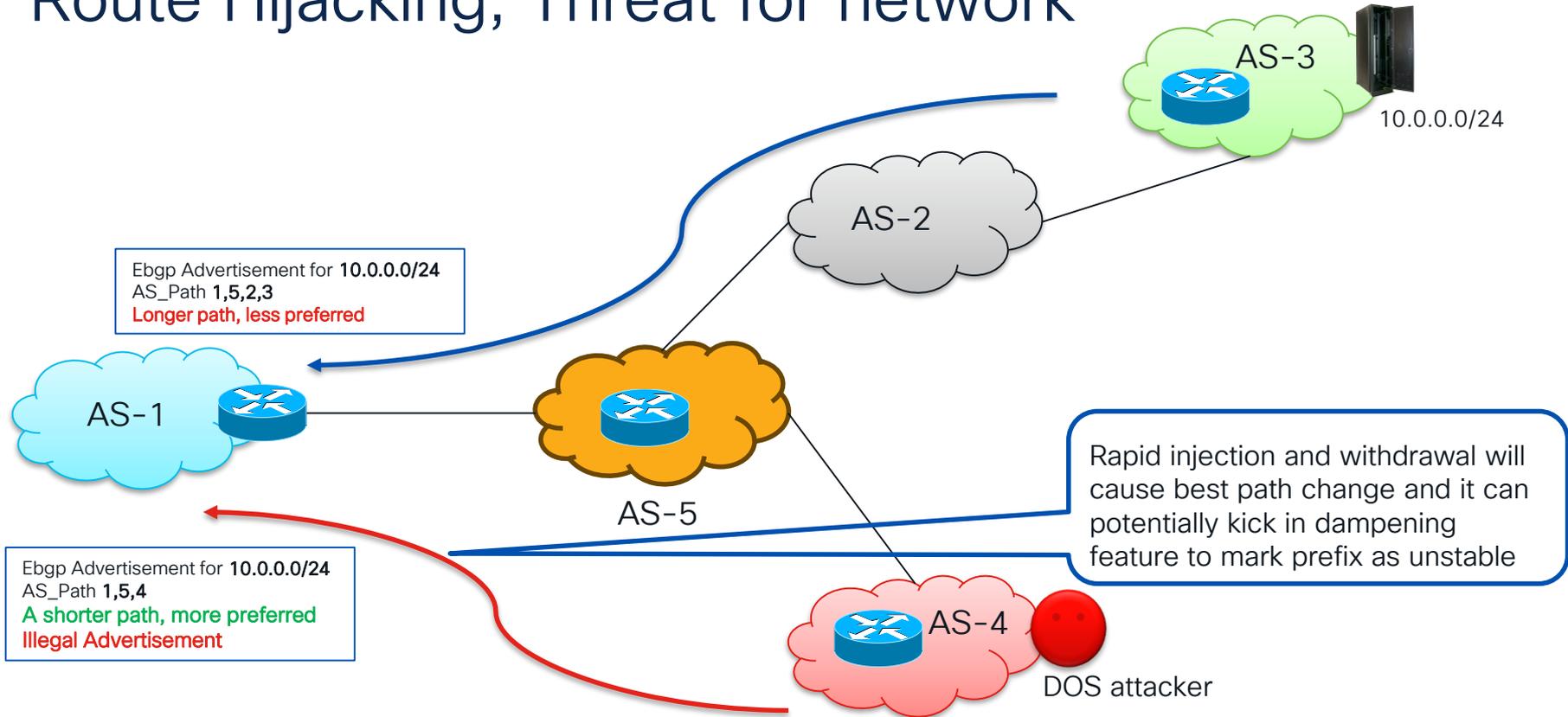
BGP security



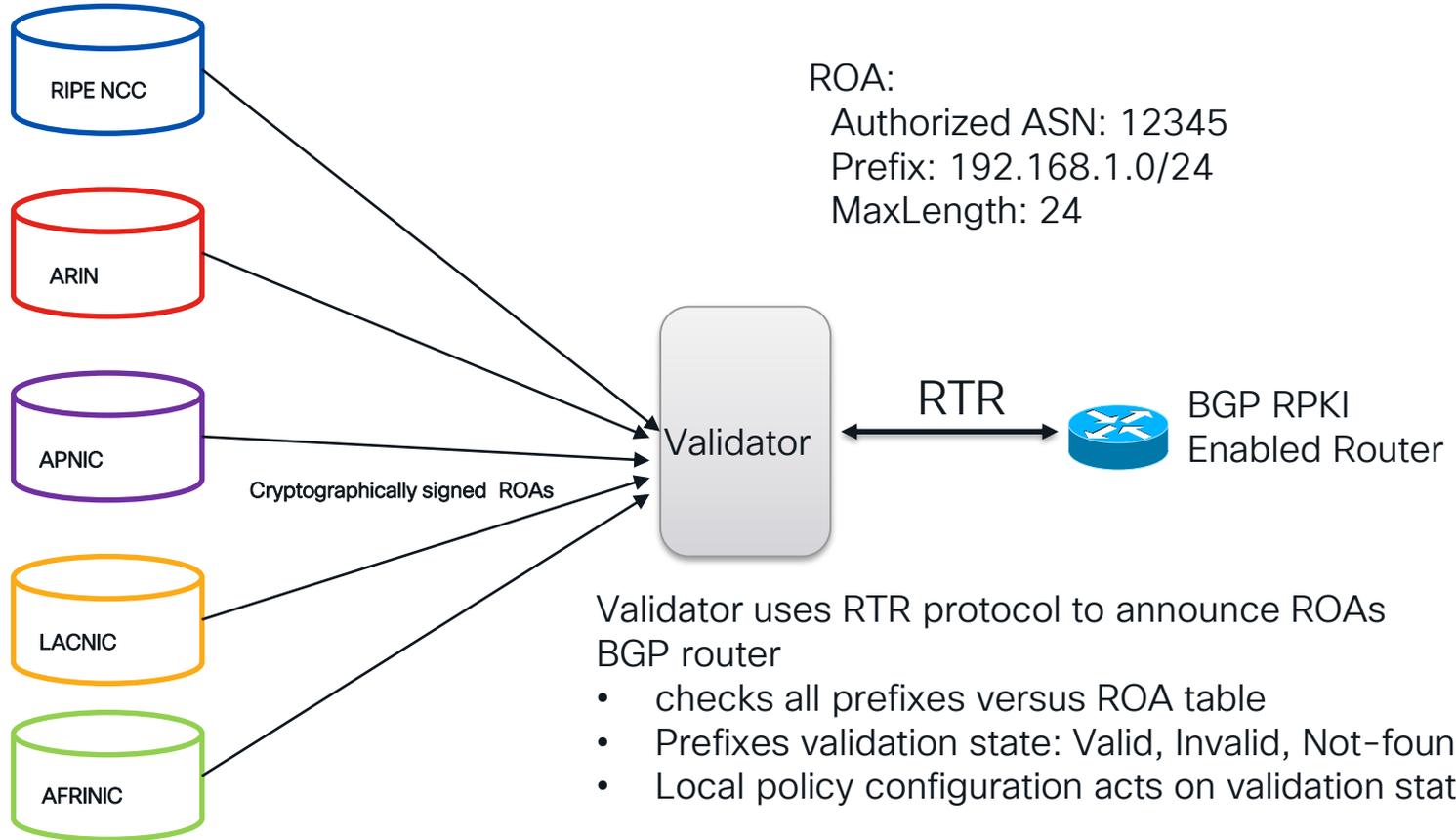
BGP: is Security important?

- BGP serves as the backbone of internet traffic.
- Even though it is the most important component of the internet core, it needs the capability to verify if the ingress BGP announcement originated from an authorized autonomous system or not.
- BGP makes it an easy candidate for various kinds of attacks. One common attack is called 'route hijack'. This attack can be exploited to:
 - Steal IPs to send spam results in IP getting rejected and hence denial of service.
 - Spy on traffic to obtain sensitive information like passwords.

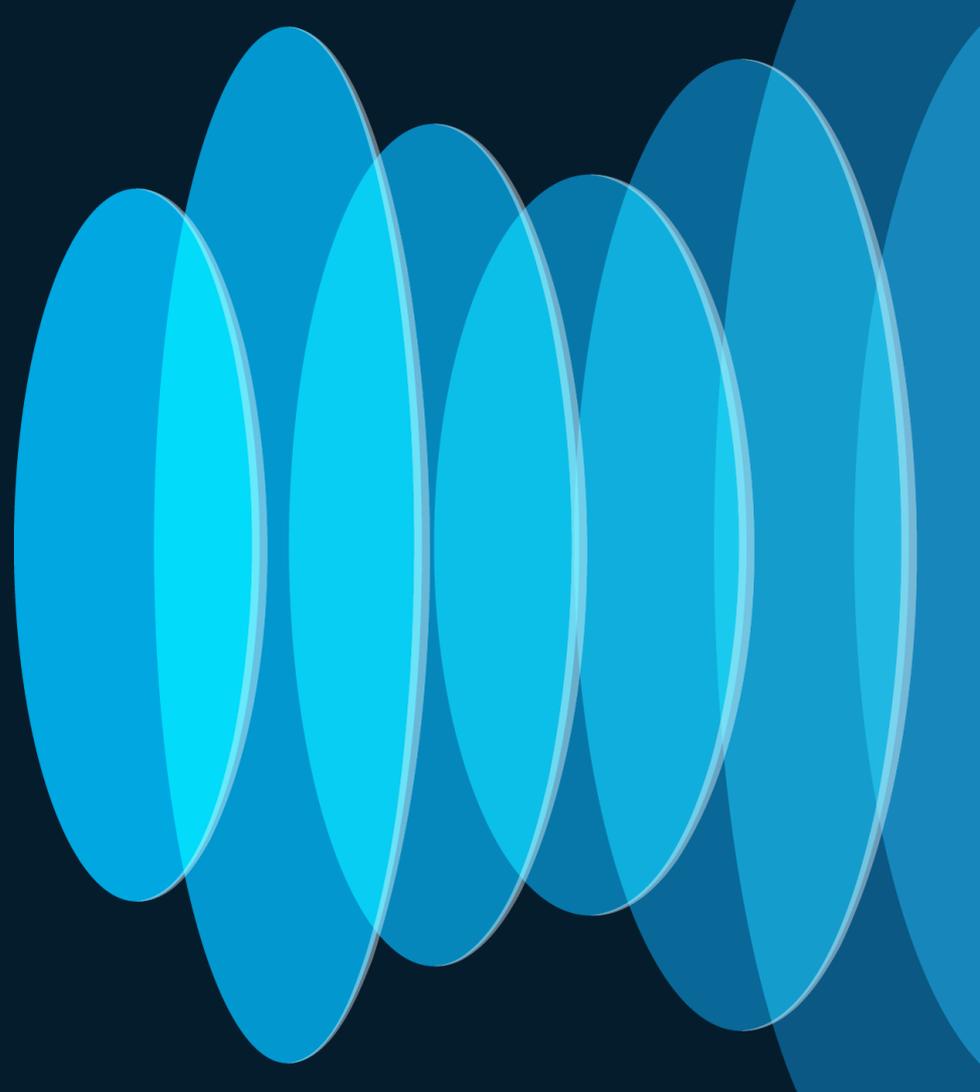
Route Hijacking, Threat for network



Solution: Resource Public Key Infrastructure (RPKI)



Best Practices



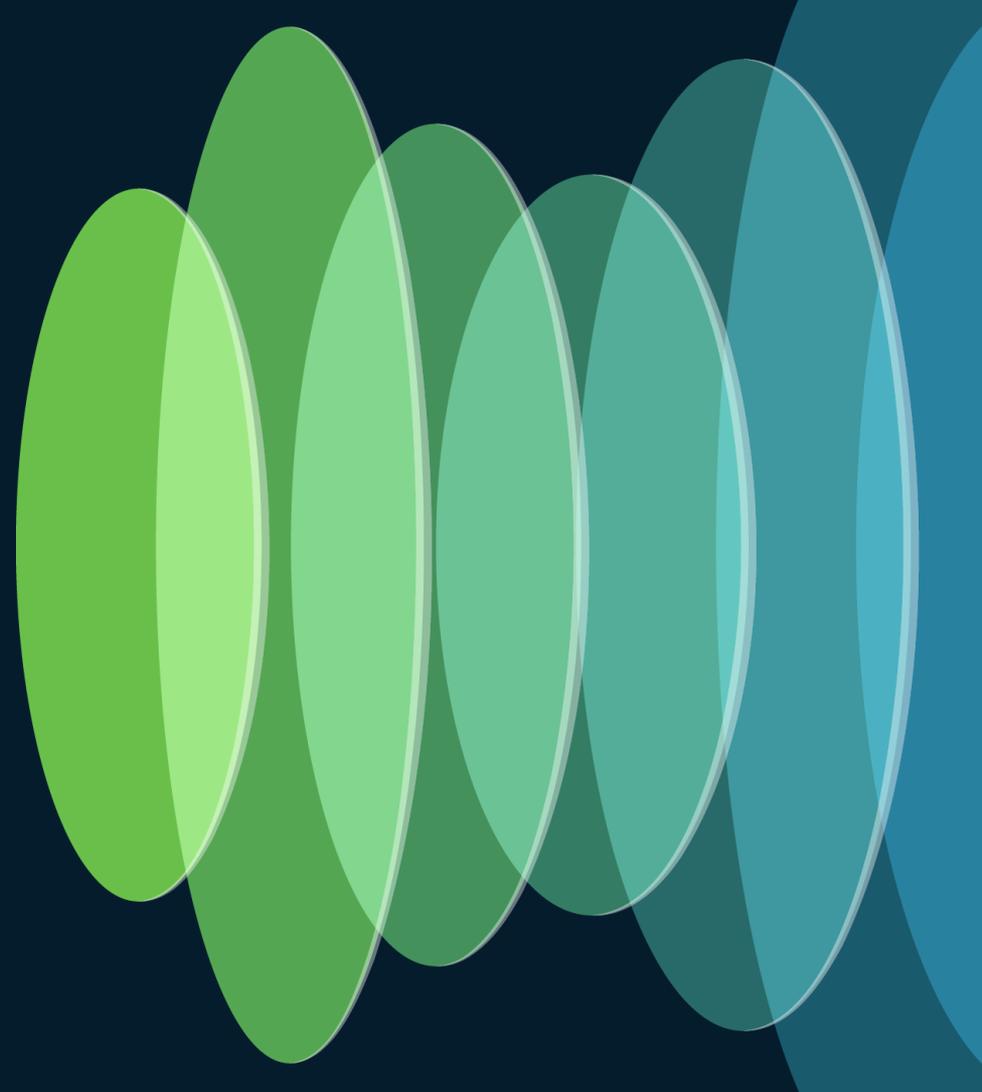
Best Practices

- **Use RR to scale BGP**; deploy RRs in pair for the redundancy
Keep RRs out of the forwarding paths and do not install BGP route in RIB
- **Choose AS format for RT and RD** i.e., ASN: X
Reserve first few 100s of X for the internal purposes such as filtering
- Consider **unique RD per VRF per PE** (vs same RD per VRF across all PEs)
Helpful for many scenarios such as multi-homing, hub & spoke.
Helpful to avoid add-path since unique RD brings uniqueness
=> all propagated by RR.
Unique RD allow VRF import policy to modify attributes since path copied from remote RD to local RD.
By opposition, same RD does not allow path attributes modification.
- **Utilize SP's public address space for PE-CE IP addressing**
Helps to avoid overlapping; Use **/31 subnetting** on PE-CE interfaces

Best Practices

- **TCP authentication (MD5/TCP-AO)** per bgp session
- **Limit number of prefixes** per-VRF and/or per-neighbor on PE
 - Max-prefix within VRF configuration; Suppress the inactive routes
 - Max-prefix for EBGP neighbor with discard/reset, threshold warning only for IBGP
- **Leverage BGP Prefix Independent Convergence (PIC)** for fast convergence <100ms
 - PIC Core : backup path with NO change in BGP nexthop.
 - PIC Edge : backup path with change in BGP nexthop.
 - Best-external advertisement
 - Next-hop tracking (ON by default)
- Consider RT-constraint for PE & RR scalability (millions of routes)
- Consider 'BGP slow peer' for PE or RR – faster BGP convergence
- Use a dedicated VPN for CE Management
- Do not configure nexthop trigger delay 0, use PIC

Q&A



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 2025.



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 me at: sekrier@cisco.com
mankamis@cisco.com



The bridge to possible

Thank you

CISCO *Live!*

#CiscoLive