
TOMORROW
starts here.
CISCO
Cisco live!

Advances in BGP

BRKRST-3371

Tom Mulvey

Network Consulting Engineer (NCE)

#clmel

Cisco live!

Agenda

- Introduction
- Motivation to Enhance BGP
- What's happened in the BGP Landscape?
- Some new cool features that may interest you

What is BGP?

- It is the plumbing technology of the Internet
- It is a protocol used to connect different autonomous systems (AS) together
- Without BGP the Internet would not exist in its current form as a stable routing platform in an unstable environment.

What is BGP? – What it truly is?

The

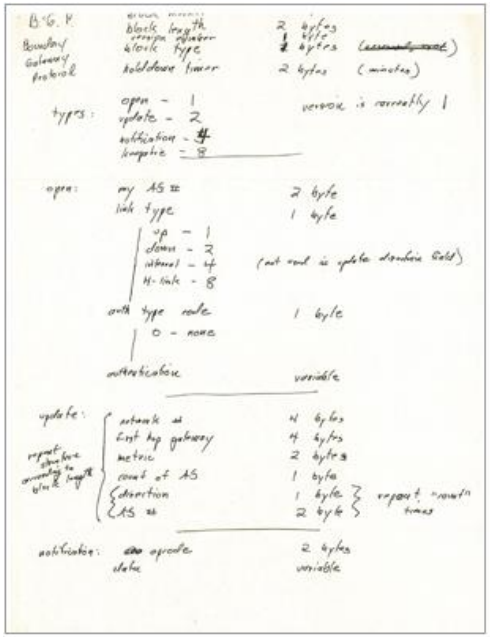
Bloody Good Protocol

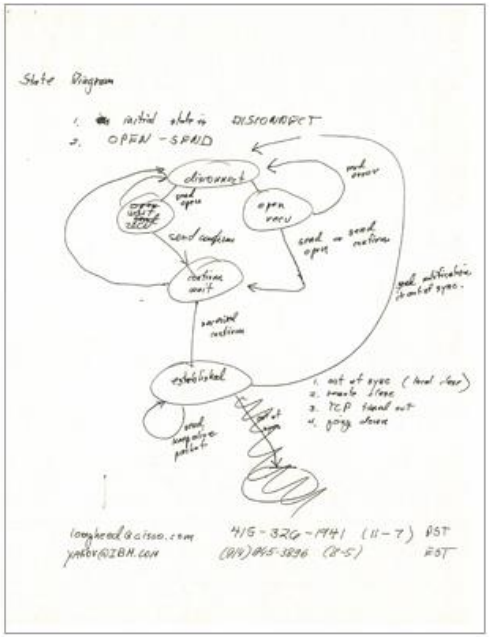
Agenda

- Introduction
- Motivation to Enhance BGP
- What's happened in the BGP Landscape?
- Some new cool features that may interest you

BGP Started in 1989…

- Motivation and Development of BGP: When the Internet grew and moved to an autonomous system (AS) mesh architecture, it required a stable, non-chatty and low CPU consuming protocol to connect all of these ASs together.
- In June 1989, the first version of this new routing protocol was formalised, with the publishing of RFC 1105, A Border Gateway Protocol (BGP).





Cisco live!

Service Provider Routing and Services Progress

- Multimedia, Mobile Internet and Cloud Services will generate massive bandwidth explosion
- Prefix growth is almost a linear curve
- Evolution of offered BGP services go from basic technologies to very advanced infrastructures

Control-plane Evolution

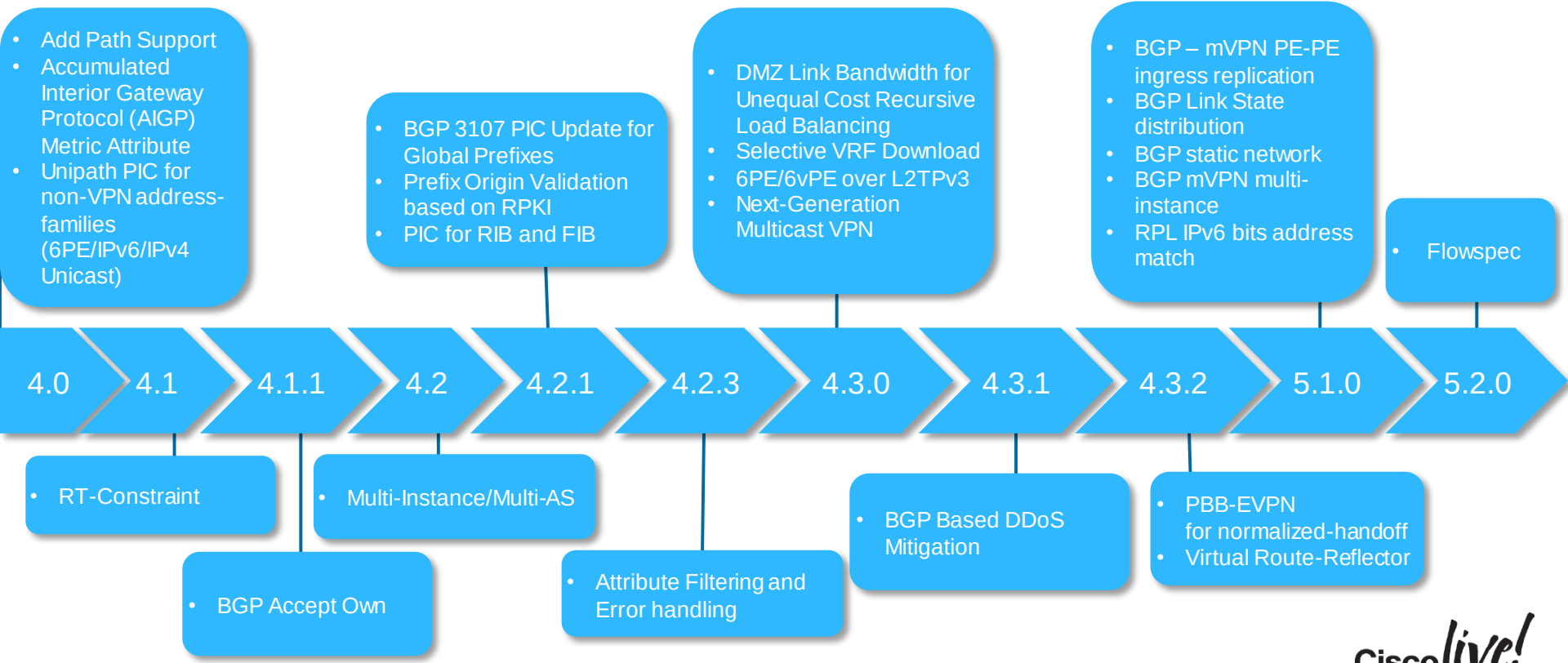
- Most services are moving towards BGP

Service/transport	200x and before	2014 and future
IDR (Peering)	BGP	BGP (IPv6)
SP L3VPN	BGP	BGP + FRR + Scalability
SP Multicast VPN	PIM	BGP Multicast VPN
DDOS mitigation	CLI	BGP flowspec
Network Monitoring	SNMP	BGP monitoring protocol
Security	Filters	BGP Sec (RPKI), DDoS Mitigation
Proximity		BGP connected app API
SP-L3VPN-DC		BGP Inter-AS, VPN4DC
Business & CE L2VPN	LDP	BGP PW Sign (VPLS)
DC Interconnect L2VPN		BGP MAC Sign (EVPN)
MPLS transport	LDP	BGP+Label (Unified MPLS)
Data Centre	OSPF/ISIS	BGP + Multipath
Massive Scale DMVPN	NHRP / EIGRP	BGP + Path Diversity
Campus/Ent L3VPN	BGP	BGP

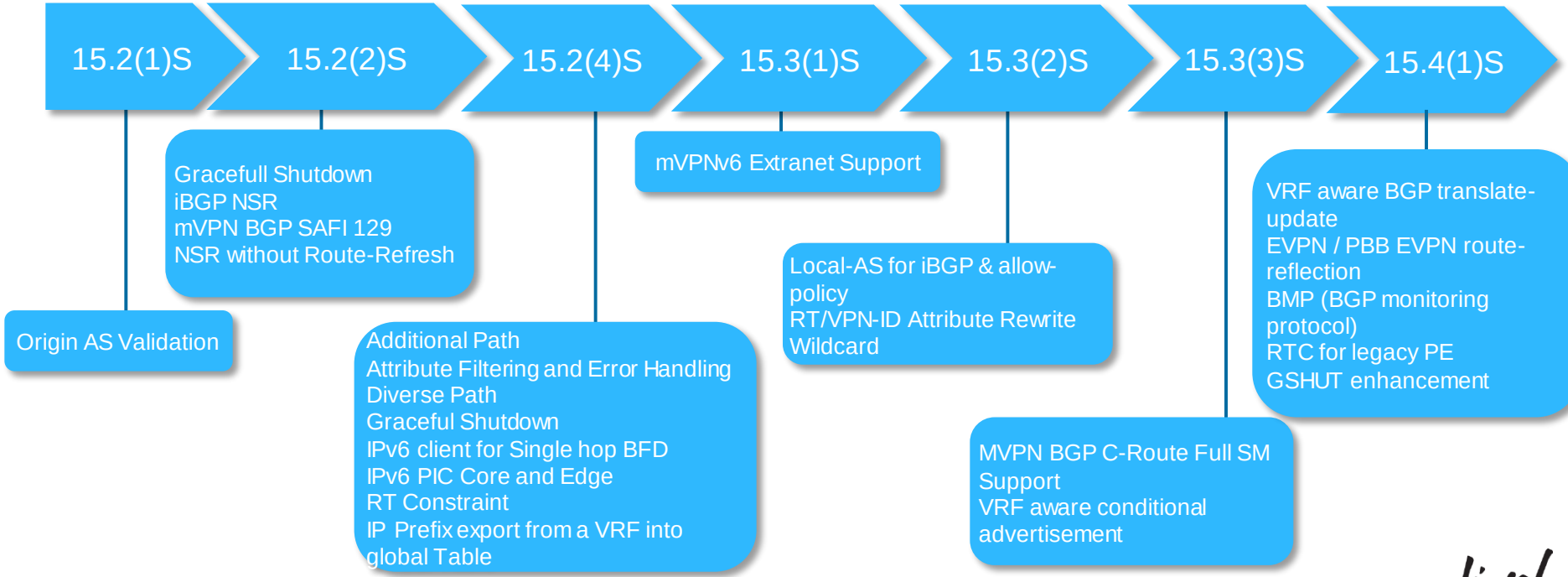
Agenda

- Introduction
- Motivation to Enhance BGP
- What's happened in the BGP Landscape?
- Some new cool features that may interest you

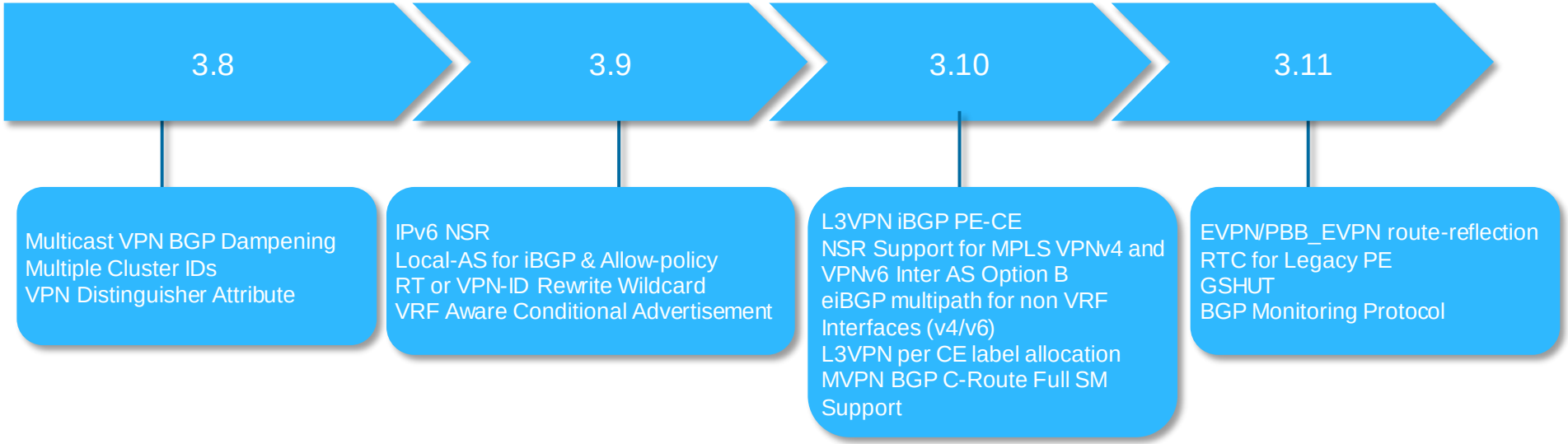
What's Happened in the XR Landscape?



What's Happened in the IOS Landscape?



What's Happened in the XE Landscape?

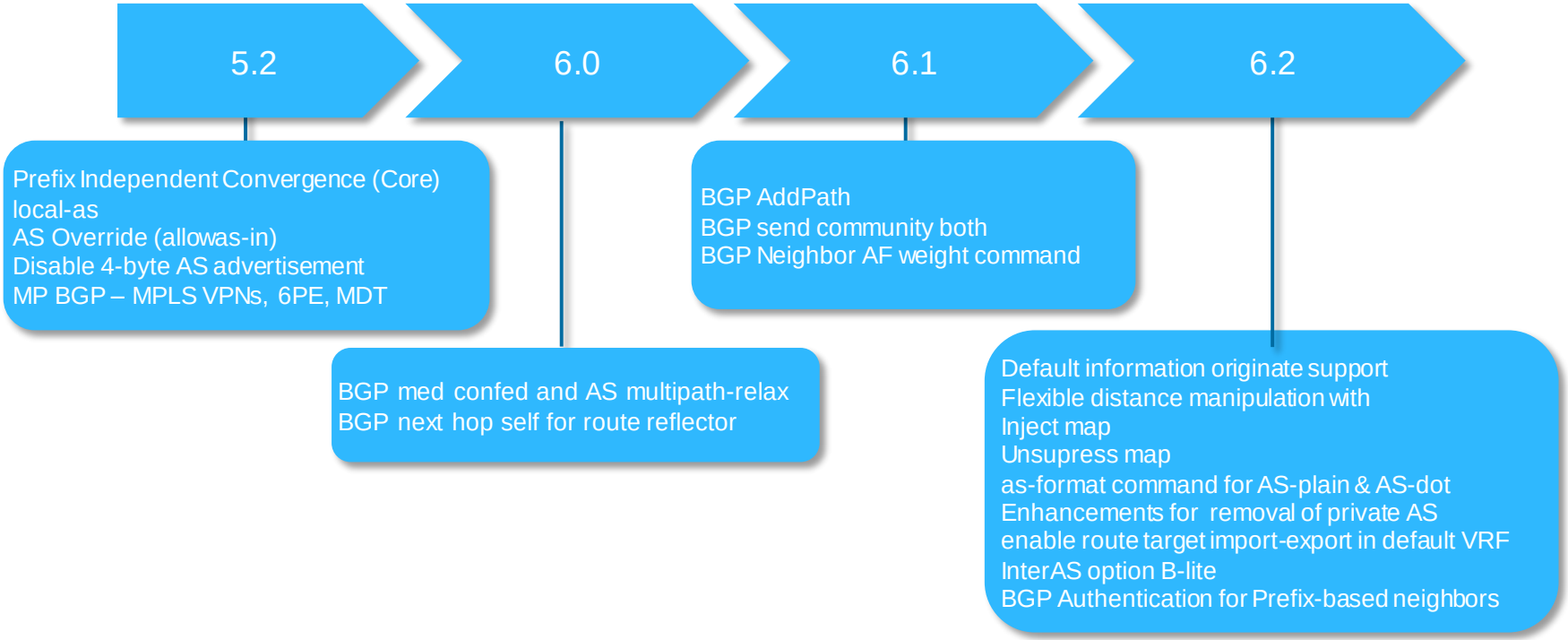


http://www.cisco.com/en/US/docs/routers/asr1000/release/notes/asr1k_rn_rel_notes.pdf

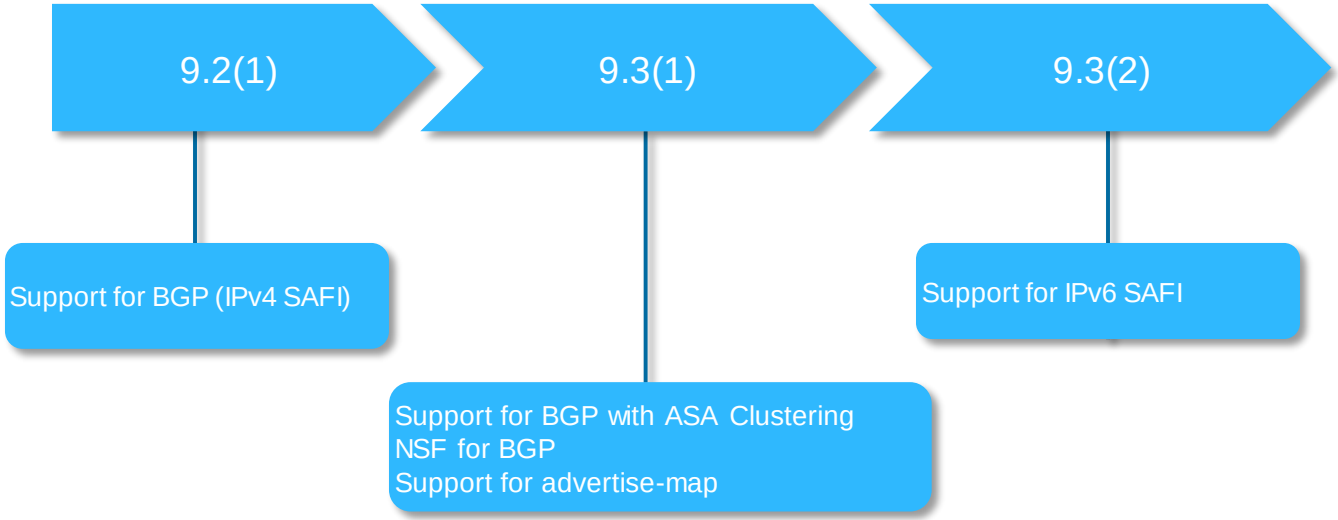
http://www.cisco.com/en/US/docs/routers/asr1000/release/notes/asr1k_rn_rel_notes.html

Cisco live!

What's Happened in the NXOS Landscape?



What's Happened in the ASA Landscape?



Agenda

- Introduction
- Motivation to Enhance BGP
- What's happened in the BGP Landscape?
- Some new cool features that may interest you

Topics We Will Cover

Convergence

- PIC Edge
- Diverse-Path (Shadow Session)
- Add-Path

Functionality/Advanced Topologies

- Accumulated IGP (AIGP)
- Multi-Instance/AS BGP
- VRF-aware Conditional Advertisement
- RT / VPN Distinguisher Attribute Rewrite
- eiBGP multipath for non-VRF interfaces
- L3VPN iBGP PE-CE
- Local-AS for iBGP and Allow-Policy
- EVPN/PBB-EVPN Route-Reflection

Scalability

- Automated Route-Target Filtering
- Router-Target Constraint (RTC) for Legacy PE
- Centralised Route Leaking/Extranetting
- Accept-Own
- Per VRF/CE Label

Security/Operations

- Attribute Filtering and Error Handling
- Origin Validation
- Graceful Shutdown
- BGP Monitoring Protocol (BMP)
- Sinkholing
- Policy-Based Routing
- Flowspec

Convergence

Cisco live!

Problem: BGP Is Slow To Converge

Problem: BGP is Slow to Converge

- Methods exist to provide fast convergence in the core
  - IGP with BFD
  - MPLS-TE
- BGP is a dinosaur – big and powerful, but slow moving
- The more prefixes we have, the slower BGP is to converge
- Can we make BGP converge like an IGP?

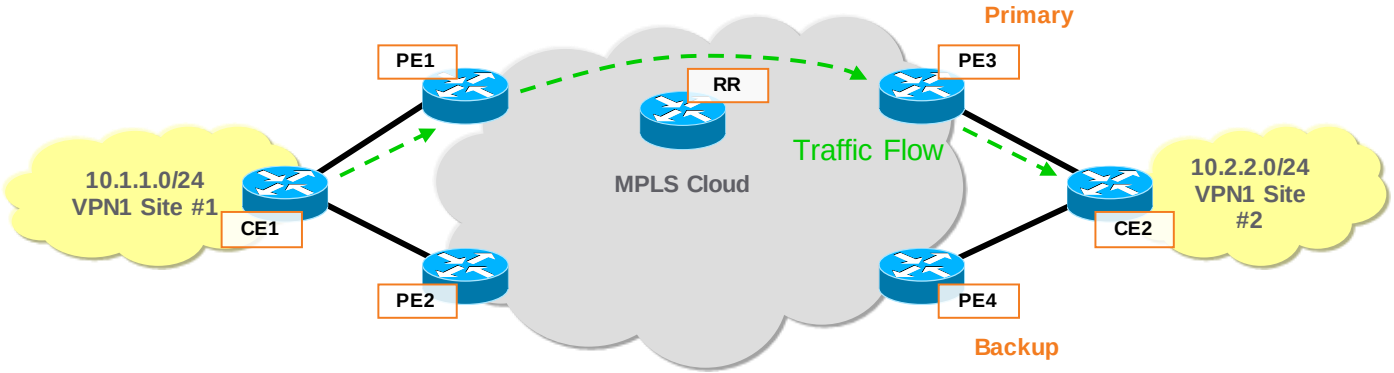
Solution: Prefix Independent Convergence (PIC) Edge

- Prefix Independent Convergence (same convergence time for 100 or 1 million prefixes)
- Reduced traffic loss
- Creates backup (shadow) path in RIB, FIB and CEF table for fast failover
- Updates data-plane while waiting for control-plane to converge after failure.

IOS: 15.2T
IOS-XE: 3.2
IOS-XR: 4.2.1
NX-OS: 6.2

PIC Edge: PE-CE Link Protection

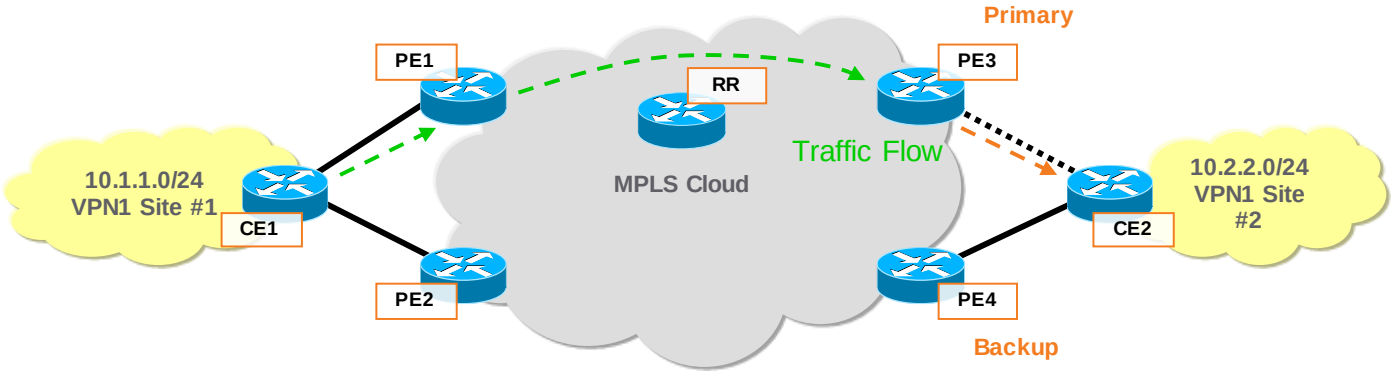
BGP Resiliency/HA Enhancement



- PE3 configured as primary, PE4 as backup
  - PE3 preferred over PE4 by local preference
  - CE2 has different RDs in VRFs on PE3 and PE4
  - PE4: advertise-best-external, to advertise route received via PE4-CE2 link
  - PE3: additional-paths install, to install primary and backup path.

PIC Edge: Link Protection

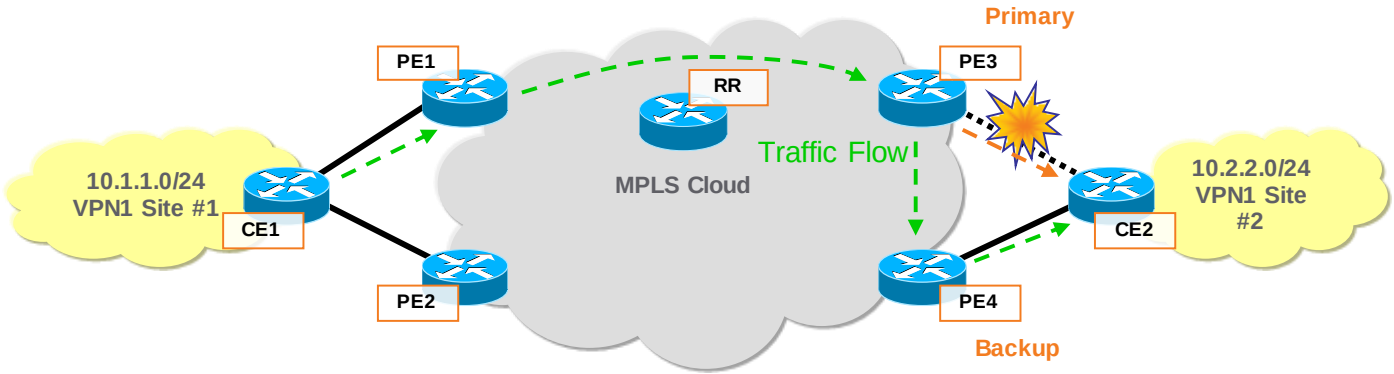
BGP Resiliency/HA Enhancement



- PE3 has primary and backup path
  - Primary via directly connected PE3-CE2 link
  - Backup via PE4 best external route
- What happens when PE3-CE2 link fails?

PIC Edge: Link Protection

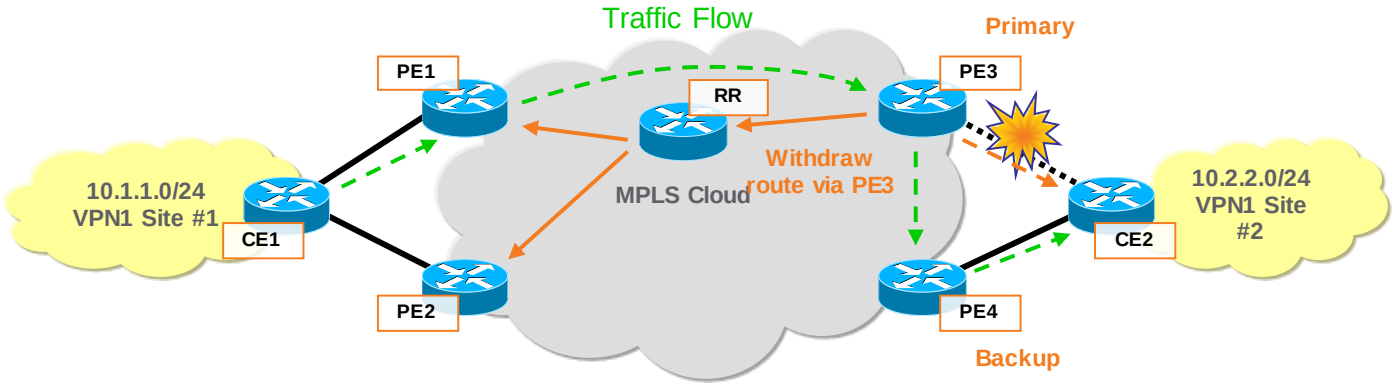
BGP Resiliency/HA Enhancement



- CEF (via BFD or link layer mechanism) detects PE3-CE2 link failure
  - CEF immediately swaps to repair path label
  - Traffic shunted to PE4 and across PE4-CE2 link.

PIC Edge: Link Protection

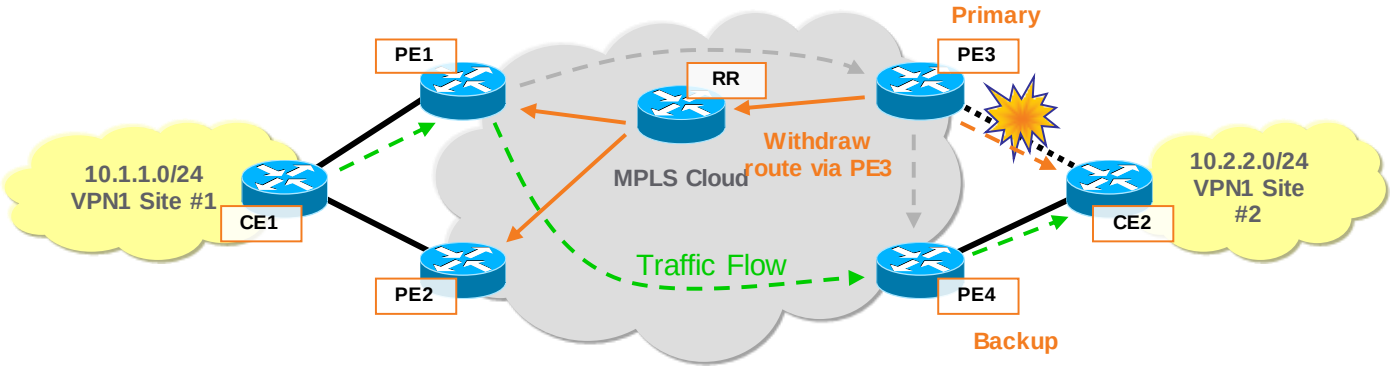
BGP Resiliency/HA Enhancement



- PE3 withdraws route via PE3-CE2 link
  - Update propagated to remote PE routers.

PIC Edge: Link Protection

BGP Resiliency/HA Enhancement



- BGP on remote PEs selects new bestpath
  - New bestpath is via PE4
  - Traffic flows directly to PE4 instead of via PE3.

Enabling BGP PIC Edge: IOS-XR

- Two BGP-PIC Edge Flavours: BGP PIC Edge Multipath and Unipath
- **Multipath**: Re-routing router load-balances across multiple next-hops, backup next-hops are actively taking traffic, are active in the routing/forwarding plane, commonly found in active/active redundancy scenarios.
  - No configuration, apart from enabling BGP multipath (maximum-paths ...)
- **Unipath**: Backup path(s) are NOT taking traffic, as found in active/standby scenarios

```
route-policy backup
 ! Currently, only a single backup path is supported
 set path-selection backup 1 install [multipath-protect] [advertise]
end-policy

router bgp ...
 address-family ipv4 unicast
  additional-paths selection route-policy backup
!
 address-family vpnv4 unicast
  additional-paths selection route-policy backup
!
```

Enabling BGP PIC Edge: IOS

- Just like IOS-XR, multipath requires no additional configuration on IOS
- PIC-Edge unipath needs to be enabled explicitly ...

```
router bgp ...
 address-family ipv4 [vrf ...]
   or
 address-family vpnv4
  bgp additional-paths install
```

... or implicitly when enabling best external

```
router bgp ...
 address-family ipv4 [vrf ...]
   or
 address-family vpnv4
  bgp advertise-best-external
```

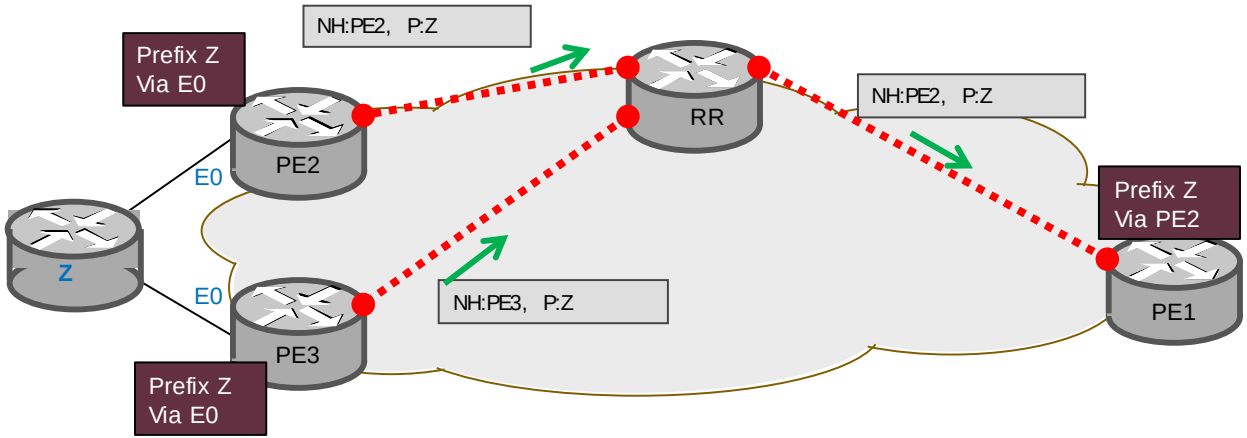
http://www.cisco.com/en/US/docs/ios/iproute_bgp/configuration/guide/irg_bgp_mp_pic.html

http://www.cisco.com/en/US/docs/ios/ios_xe/iproute_bgp/configuration/guide/irg_best_external_xe.html

How Will My PEs Learn About The Alternate Paths?

Problem: How Will My PEs Learn About The Alternate Paths?

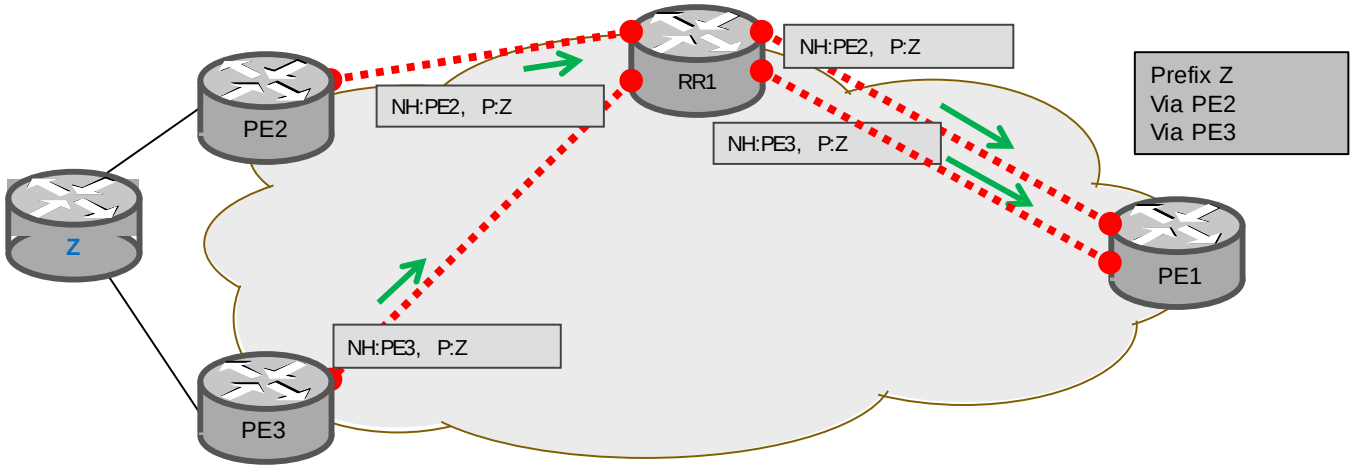
- By default my RR only reflects the Best-Route
- 2 Solutions available.



Solution 1: Diverse BGP Path Distribution Shadow Session

IOS: 15.2T
IOS-XE: 3.4

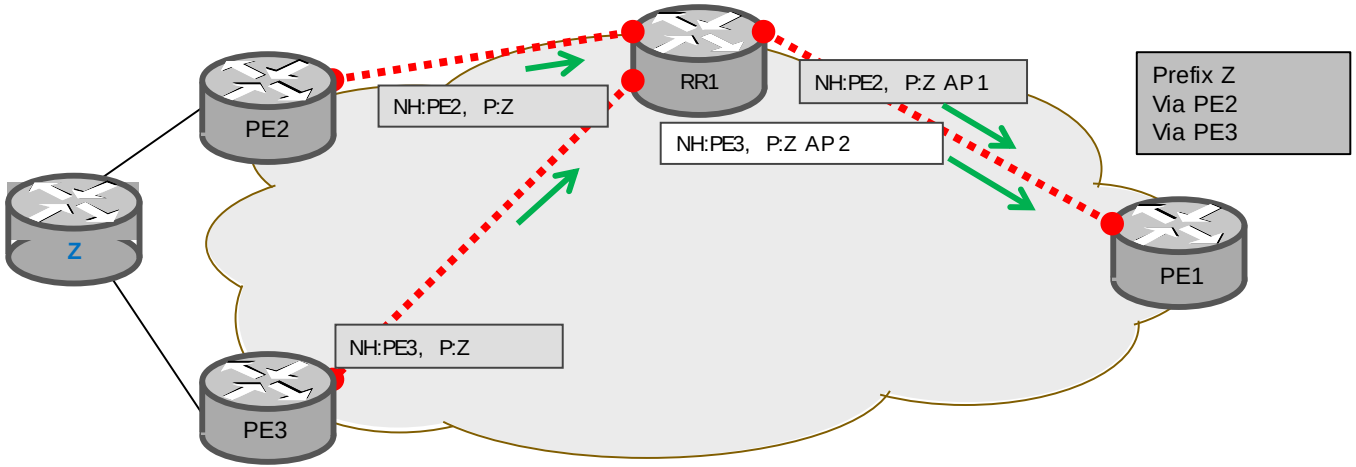
- Easy deployment – no upgrade of any existing PE is required, just new iBGP session per each extra path (CLI knob in RR1)
- Diverse iBGP session announces the 2nd best path
- "advertise diverse-path backup" command on second session on RR.



Solution 2: BGP Add-Path

IOS: 15.3T
IOS-XE: 3.7
IOS-XR: 4.0.0
NX-OS: 6.1

- Add-Path will signal diverse paths from 2 to X paths
- Required all PE and RR devices to support Add-Path capability
- RFC 6774.



BGP Add-path Flavours

- IETF defines 5 flavours of Add-x-Path. 2 are implemented by Cisco:

- Add-n-path: with add-n-path the route reflector will do best path computation for all paths and send n best to PE.
  - Use case: Primary + n Backup scenario

- Add-all-path: with add-all-path, the route reflector will do the primary best path computation (only on first path) and then send all path to BR/PE.

- Cisco innovation: Add-all-multipath and Add-all-multipath+backup in XR 4.3.1.

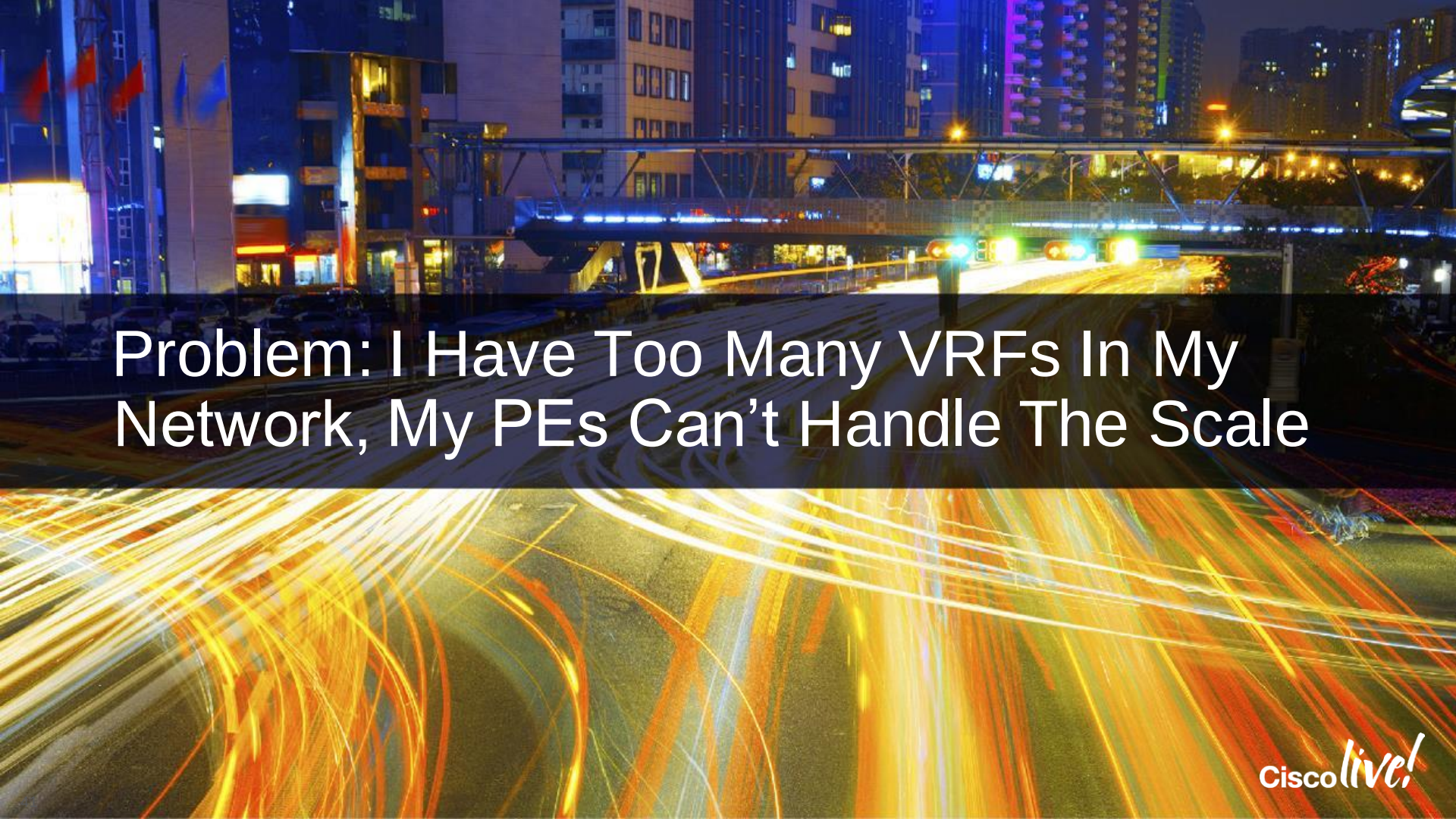
Add-Path Configuration – IOS-XR

- Enable in global address-family mode
  - Enables for all IBGP neighbours
- Enable/Disable in neighbour mode

```
router bgp 100
address-family ipv4 unicast
additional-paths send
!
address-family vpnv4 unicast
additional-paths send
!
neighbor 1.1.1.1
remote-as 100
address-family ipv4 unicast
!
address-family vpnv4 unicast
!
!
neighbor 2.2.2.2
remote-as 100
capability additional-paths send disable
address-family ipv4 unicast
!
```


Cisco live!


Scalability



Problem: I Have Too Many VRFs In My Network, My PEs Can’t Handle The Scale

Problem: Too Many VRFs, PEs Can't Scale

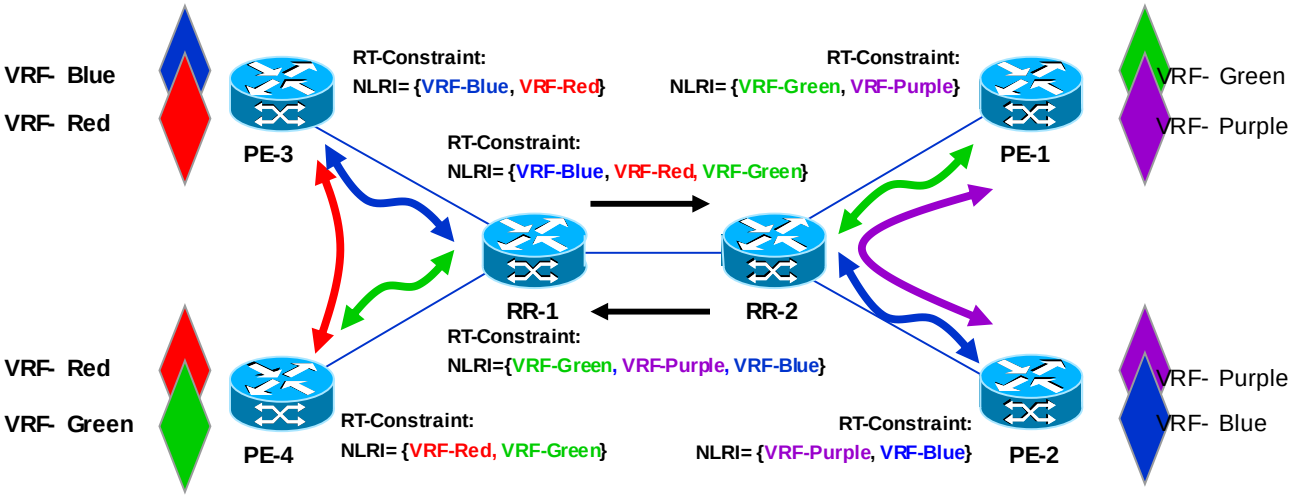
- Increased VPN service deployment increases load on PE routers
  - Each device can only maintain a fixed number of VRFs/routes
- PEs receive routes for all VRFs (even for which they don't need)
- Some features exist to filter routes on PE once they are received e.g. Selective VRF Download (SVD)
- Highly desirable to filter unwanted VPN routes before sending them to the PEs.

Solution: Automated Route Target Filtering

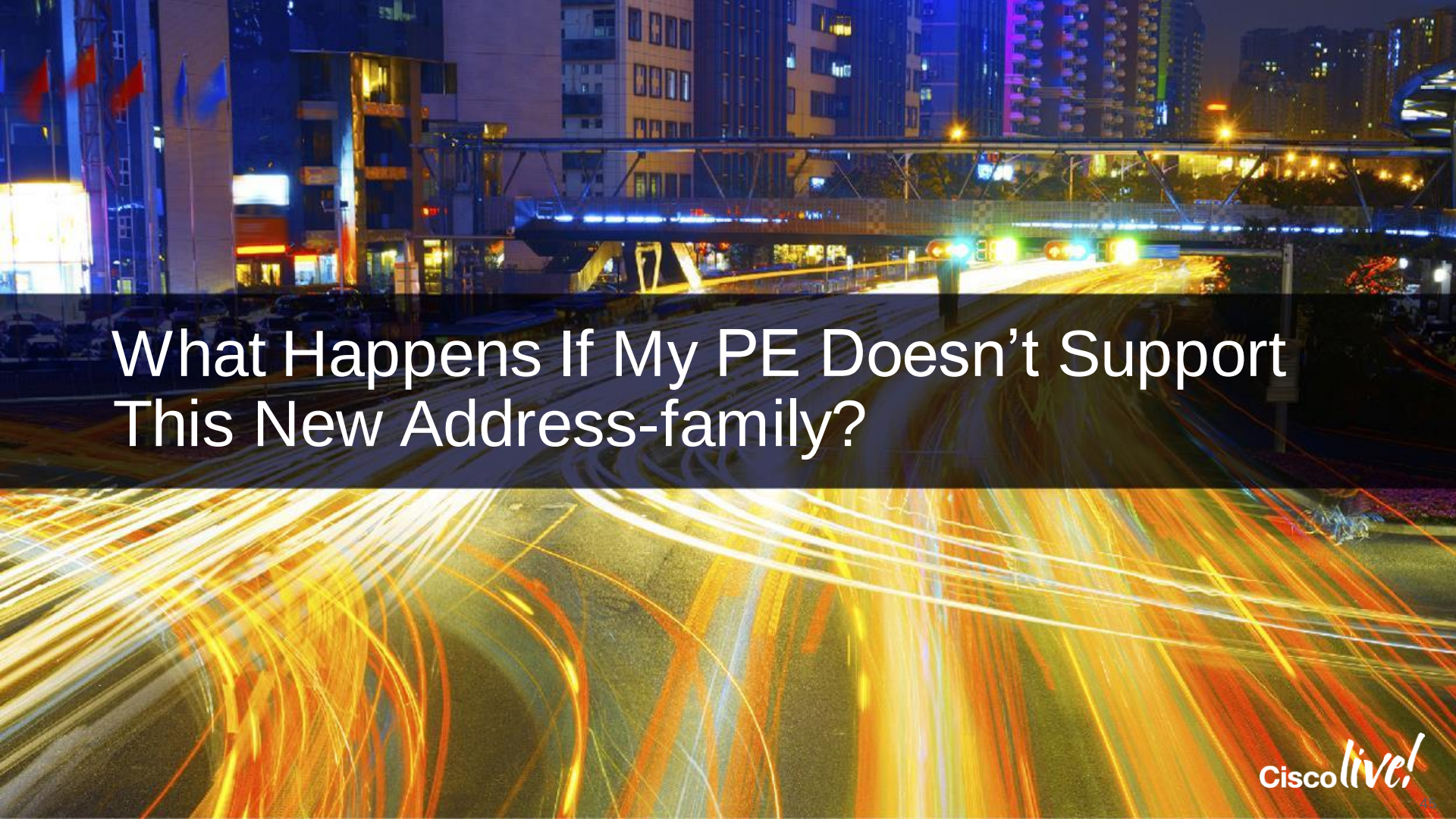
IOS: 15.2T
IOS-XE: 3.2
IOS-XR: 4.1.0

- Filter VPNs before they are sent to PE
- PE maintains less routes in Adj-RIB-in and RIB
- Improves PE and RR scaling and performance by sending only relevant VPN routes
- New “RT filter” address family.

Automated Route Target Filtering



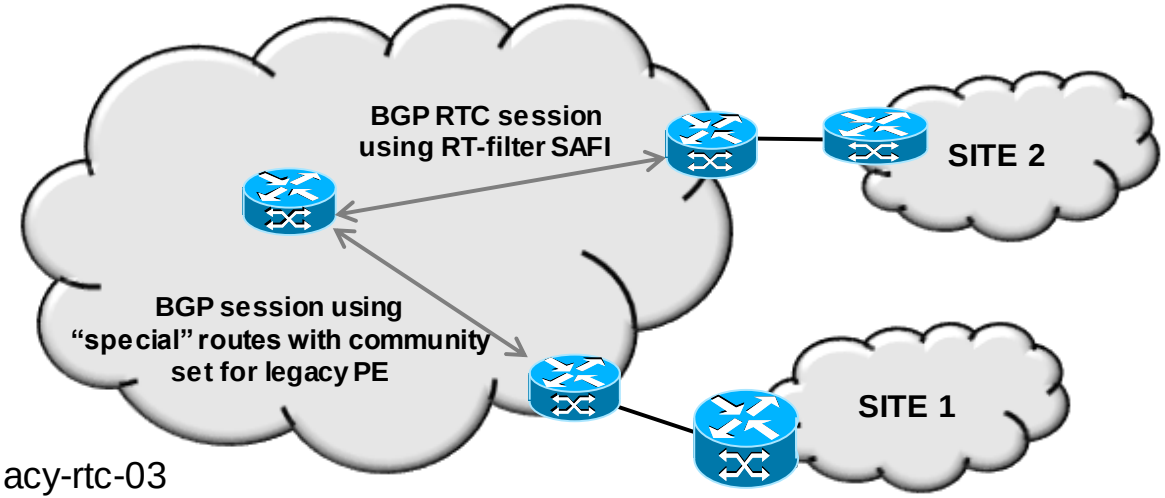

```
router bgp as-number
 address-family rtfilter unicast
  neighbor {ip-address | peer-group-name} activate
  neighbor {ip-address | peer-group-name} send-community extended
```



What Happens If My PE Doesn't Support This New Address-family?

Solution: Route Target Constraint for Legacy PE

- Feature enabled on RR only
- Legacy PEs advertise special routes with mapped RTs and ROUTE_FILTER community
- The presence of the community triggers the RR to extract the RTs and build RT membership information
- Allows co-existance of automated RTC and RTC for legacy PEs.



IOS: 15.4T
IOS-XE: 3.11
IOS-XR: 4.0.0

draft-ietf-idr-legacy-rtc-03

RTC for Legacy PE Configuration

Route-reflector

```
!
 router bgp 1
 address-family vpnv4 unicast
  neighbor 10.1.1.1 accept-route-legacy-rt
!
```

Legacy-PE

```
ip vrf route-filter
 rd 55:1111
 export map SET_RT

route-map SET_RT permit 10
 match ip address prefix-list RT_NET1
 set community 65535:2 (0xFFFF0002)
 set extcommunity rt 255.220.0.0:12241 255.220.0.0:12242 additive
 set extcommunity rt 255.220.0.0:12243 255.220.0.0:12244 additive
 set extcommunity rt 255.220.0.0:12245 255.220.0.0:12246 additive
 set extcommunity rt 255.220.0.0:12247 255.220.0.0:12248 additive
 set extcommunity rt 255.220.0.0:12249 255.220.0.0:12250 additive
!
route-map SET_RT permit 20
 match ip address prefix-list RT_NET2
 set community 65535:2 (0xFFFF0002)
 set extcommunity rt 255.220.0.0:12251 255.220.0.0:12252 additive
 set extcommunity rt 255.220.0.0:12253 255.220.0.0:12254 additive
 set extcommunity rt 255.220.0.0:12255 additive
!
ip route vrf route-filter 1.1.1.1 255.255.255.255 Null0 – (matching prefix-set RT_NET1)
ip route vrf route-filter 1.1.1.2 255.255.255.255 Null0 –(matching prefix-set RT_NET2)

route-map LEG_PE permit 10
 match ip address prefix-list RT_NET1 RT_NET2
 set community no-advertise additive

router bgp 55
 address-family vpnv4 unicast
 neighbor x.x.x.x route-map LEG_PE out
```

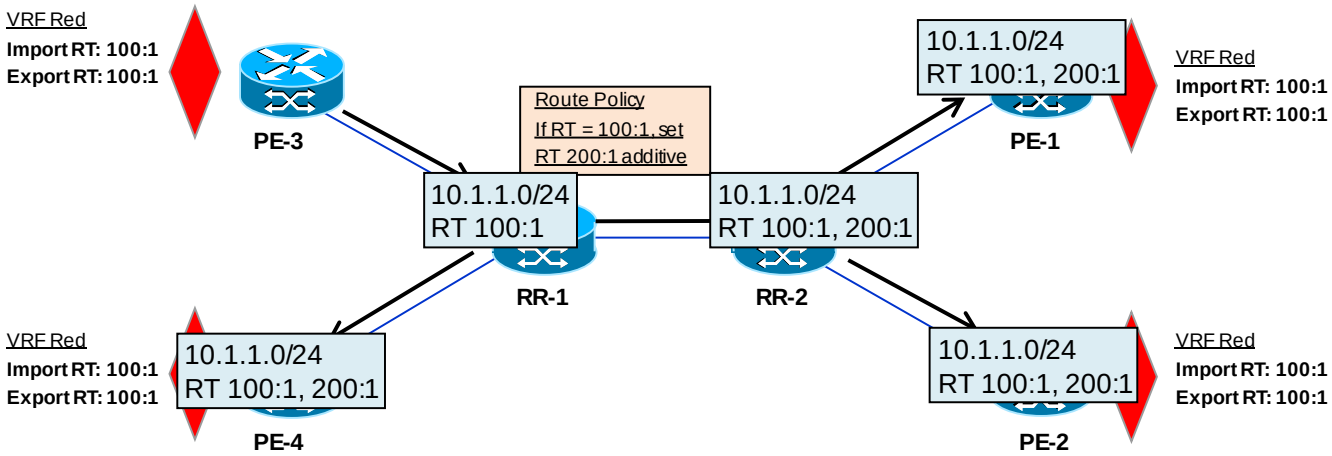
Problem: Route Leaking/RT Tagging In Large Networks

Problem: Route Leaking/RT Tagging In Large Networks

- Requirement to perform route leaking/tag certain routes with RT in a large network
- Increase overhead on PE devices
- Many devices to configure/manage
- Can we do this on a centralised device?

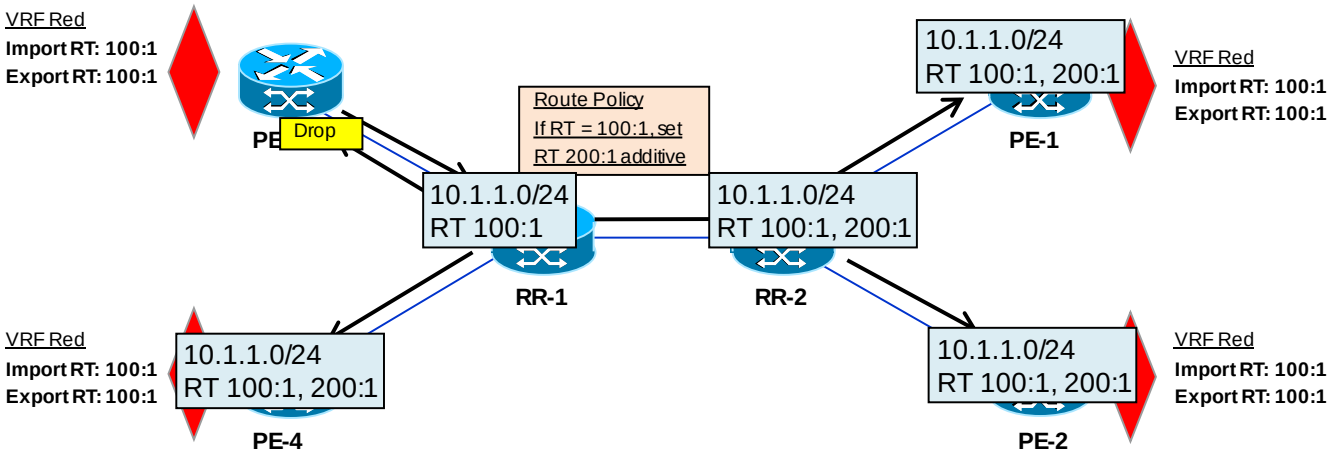
Solution: Centralised Route-Target Leaking/Extranetting

- Perform route-leaking/VRF import/export functions on centralised device (RR)
  - Reduce overhead on PEs
  - Less PEs to configure
  - Scales for large networks




What About The PE Who Originated The Route?
Cisco live!

What About the PE Who Originated the Route?



Solution: Accept Own

IOS-XR: 4.1.1

- Allows handling of self-originated VPN routes
- RR attaches ACCEPT-OWN extended community
- “accept-own” configured on PE signals to bypass ORIGINATOR_ID and NEXTHOP check
- Prefix tagged with “accept-own” community preferred over original route.

PE

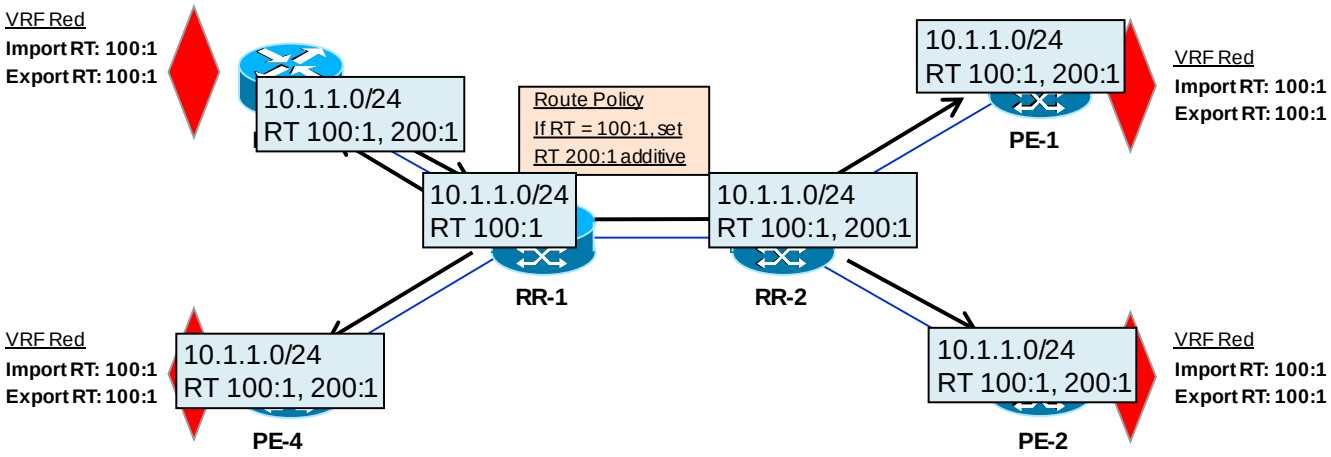
```
router bgp 1
 neighbor <route-reflector>
  accept-own
```

RR

```
route-policy rr-pe-out
 …
 if extcommunity rt matches-any CUSTOMERS then
   set extcommunity rt (xxx:yyyy) additive
   set extcommunity rt (‘accept-own’) additive
 endif
end-policy
```


Cisco live!

Accept Own - Centralised Route-Target Leaking



Problem: I Run A Large MPLS Network And My PEs Are Running Out Of Label Space

Solution: Per VRF/CE Label

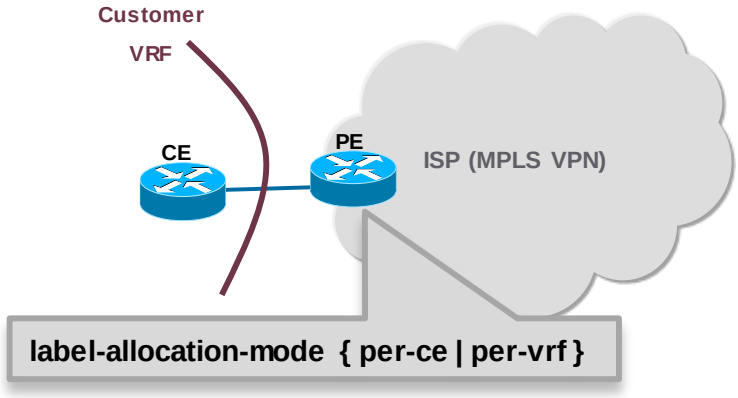
- Allows you to reduce VPN label allocation at the provider edge (PE) instead of per route/prefix
- Per CE
  - Allocates single label per CE device (save label space)
  - No VRF route lookup needed as next-hop is directly mapped to label
  - More efficient
  - Some caveats (Multipath, CsC, 6PE, etc). Check release notes.
- Per VRF
  - Allocates single label per VRF (save label space)
  - Single label allocation if multiple CEs (same VRF) are connected at PE
  - More scalable.

Per-CE	Per-VRF
IOS-XE: 3.10	IOS: 12.4(6)T
IOS-XR: 3.7.2	IOS-XE: 2.2
NX-OS: 6.0	IOS-XR: Pre 4.0.0
	NX-OS: 5.2.1

ASR9K – 1 Million labels

ASR1K – 1 Million labels

7600 – 500K labels

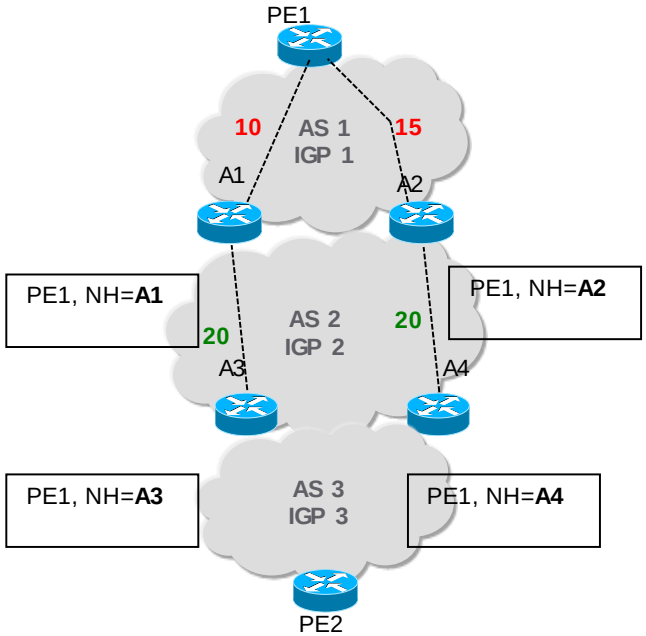


Functionality/Advanced Topologies

Problem: Routing Across Multiple AS In A Single Administrative Domain

Problem: Routing Across Multiple AS in a Single Administrative Domain

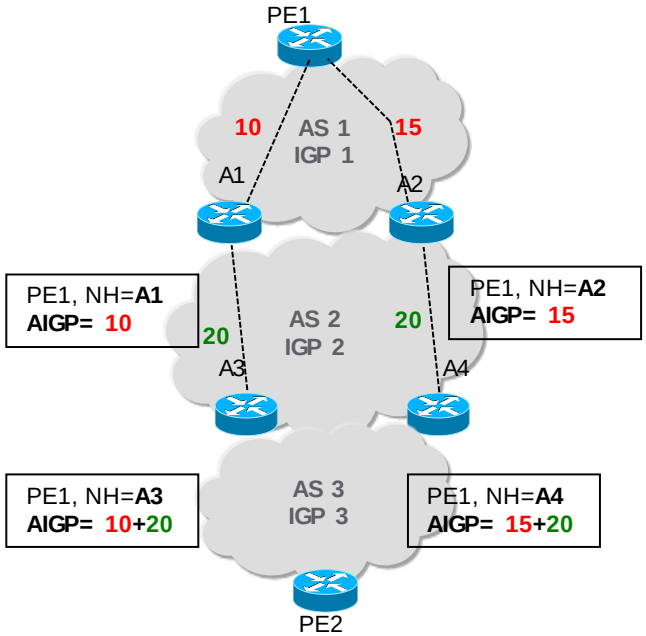
- The main driving force for this feature is to solve the IGP scale issue seen in some ISP core network (aka “Unified MPLS” architecture)
- Need to make routing decision based on the IGP metric, to choose the “shortest” path between two nodes across different AS
- MED is too low on best-path algorithm (AS Path preferred).



Solution: Accumulated IGP (AIGP)

IOS: 15.4S
IOS-XE: 3.12
IOS-XR: 4.0.0

- By default optional, non-transitive BGP path attribute
- Attribute to provide BGP a way to make its routing decisions based on the IGP metric across different AS
- The main driving force for this feature is to solve the IGP scale issue seen in some ISP core network (aka “Unified MPLS” architecture)
- Mainly to be deployed to carry next-hop prefixes/labels across different AS within the same administrative domain
- The remote ingress PE select its best path using the modified best path selection process using AIGP metric.



(draft-ietf-idr-aigp-09)

Cisco live!

AIGP: Originating AIGP

- AIGP is enabled between iBGP neighbours by default
- AIGP between eBGP neighbours need to be enabled (converted to cost community)
- AIGP can be originated by using redistribute ospf, redistribute isis, redistribute static or the BGP network command.
- AIGP can also be originated using neighbour address-family inbound or outbound policy to set AIGP to be the IGP cost or to a fixed value.

```
router bgp 1
   address-family ipv4 unicast
      redistribute ospf 1 route-policy set_aigp_1
```

```
route-policy set_aigp_1
  if destination in (…) then
    set aigp-metric 111
 elseif destination in (..) then
    set aigp-metric igp-cost
 endif
end-policy
```

Problem: I Want To Run Multiple Instances Of BGP On A Single Device

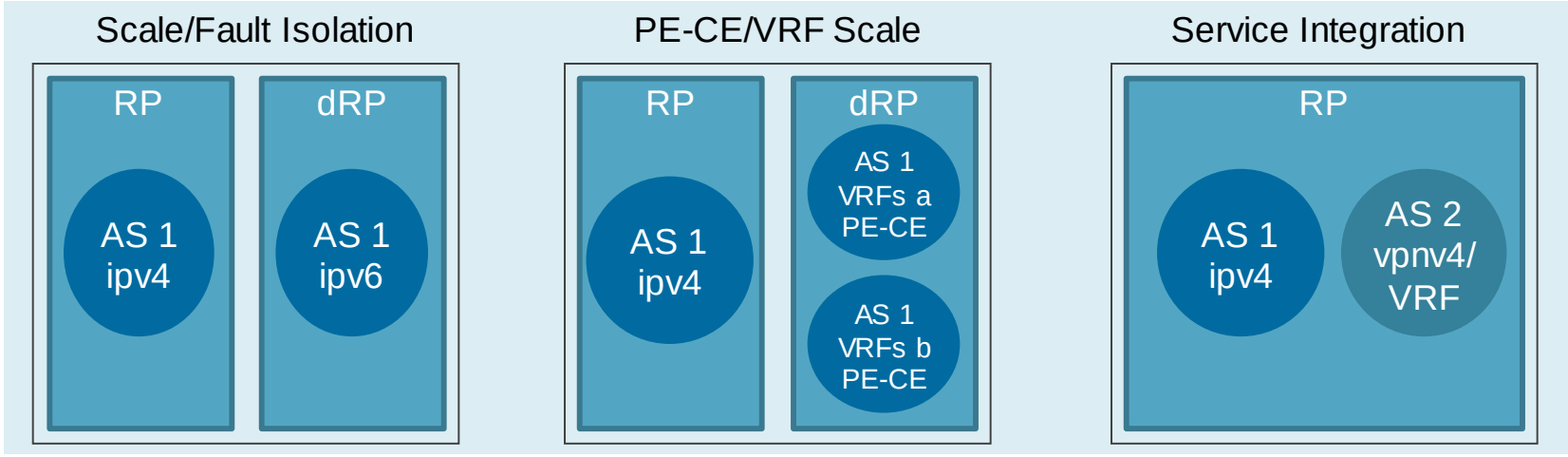
Problem: I Want to Run Multiple Instances of BGP on a Single Device

- Requirement to consolidate multiple devices to single physical device
- Provide logical separation of different address-families
- Need to achieve higher session scale across multiple instances.

Solution: Multi-Instance and Multi-AS BGP

- Run multiple instances of BGP on a router (possibly on different RP instances)
- Each instance of BGP can be configured with a different AS number
- Global address families can't be configured under more than one AS except vpnv4 and vpnv6
- VPN address-families may be configured under multiple AS instances that do not share any VRFs.

IOS-XR: 4.2.0



Configuration Example

```
router bgp 1 instance ipv4
   bgp router-id 10.0.0.1
   address-family ipv4 unicast
   neighbor 10.0.101.1
      remote-as 1
      address-family ipv4 unicast
         route-policy inbound in
         route-policy outbound out
      !
   !
  !
!
commit
!
router bgp 1 instance ipv6
   bgp router-id 10.0.0.2
   address-family ipv6 unicast
   neighbor 10.0.101.2
      remote-as 1
      address-family ipv6 unicast
         route-policy inbound in
         route-policy outbound out
      !
   !
  !
!
commit
```

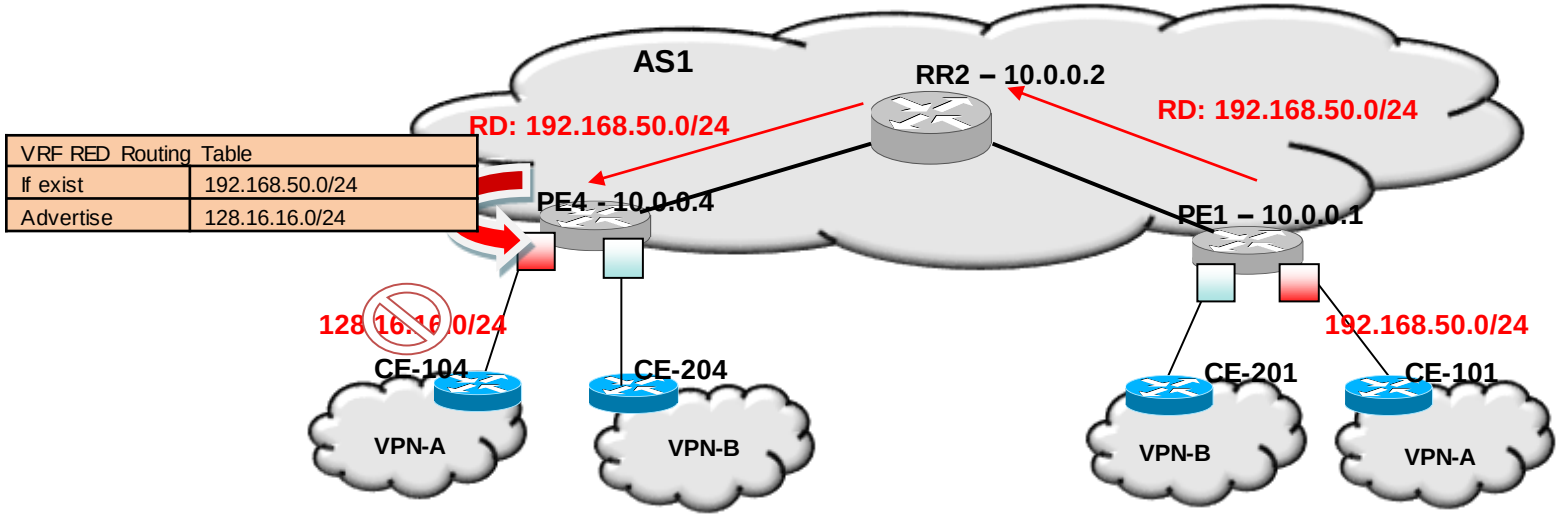
```
router bgp 3 instance vpn1
   bgp router-id 20.0.0.1
   address-family vpnv4 unicast
   neighbor 20.0.101.1
      remote-as 200
      address-family vpnv4 unicast
         route-policy inbound in
         route-policy outbound out
      !
   !
  !
  vrf foo
  !
!
commit
!
router bgp 3 instance vpn2
   bgp router-id 20.0.0.2
   address-family vpnv4 unicast
   neighbor 20.0.101.2
      remote-as 200
      address-family vpnv4 unicast
         route-policy inbound in
         route-policy outbound out
      !
   !
  !
  vrf bar
  !
!
commit
```

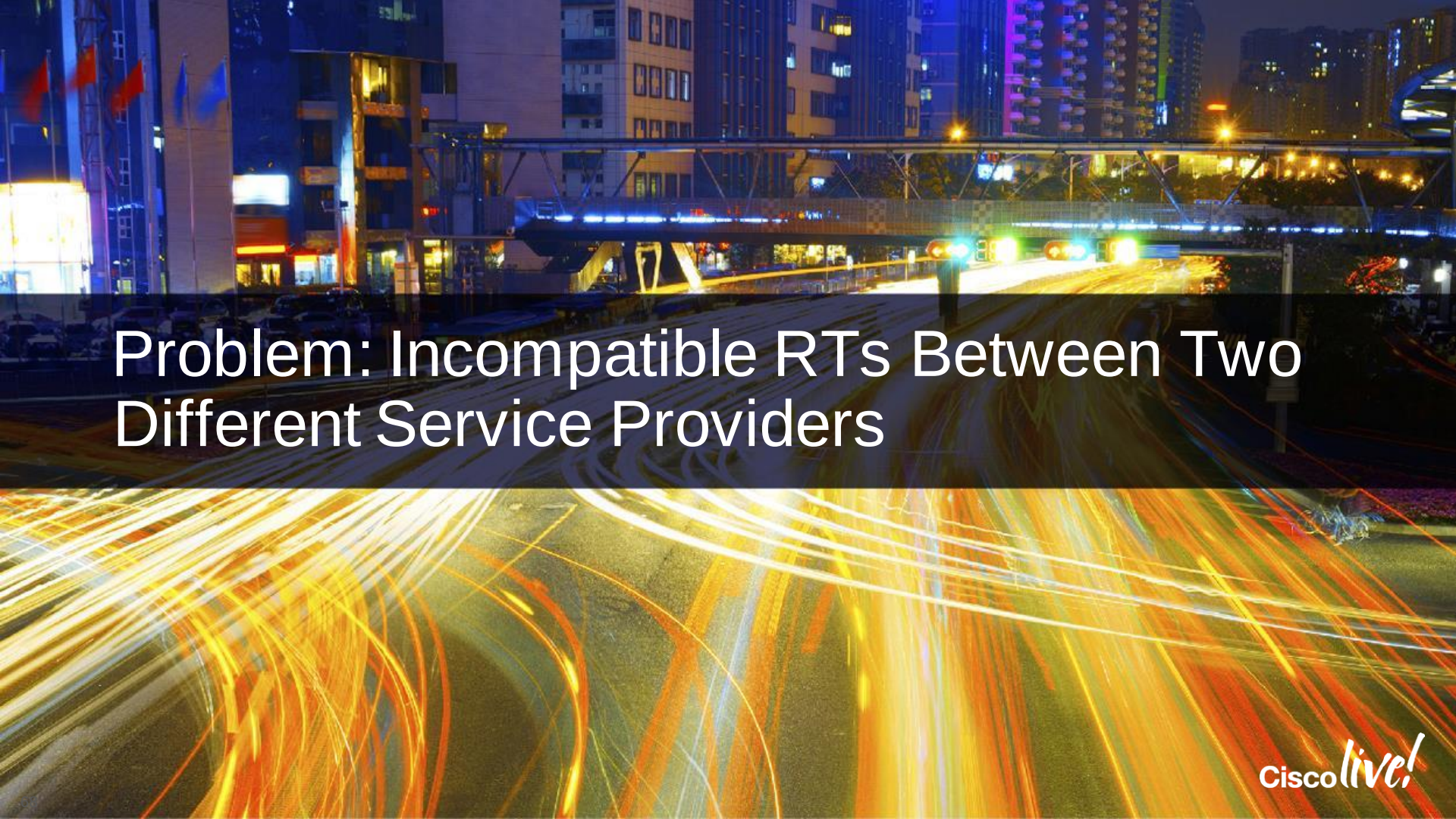
Problem: I Have Certain Routes In A VRF I Only Want To Advertise Under Certain Conditions

Solution: VRF Aware Conditional Advertisement

- Previously: Conditional advertisement supported in IPv4 Unicast/Multicast address-family
- New: Support for IPv4 VRF, IPv6 Unicast and IPv6 VRF
- Use case: Advertise backup link in network when primary link fails.

IOS: 15.4T
IOS-XE: 3.9
IOS-XR: 3.7.2

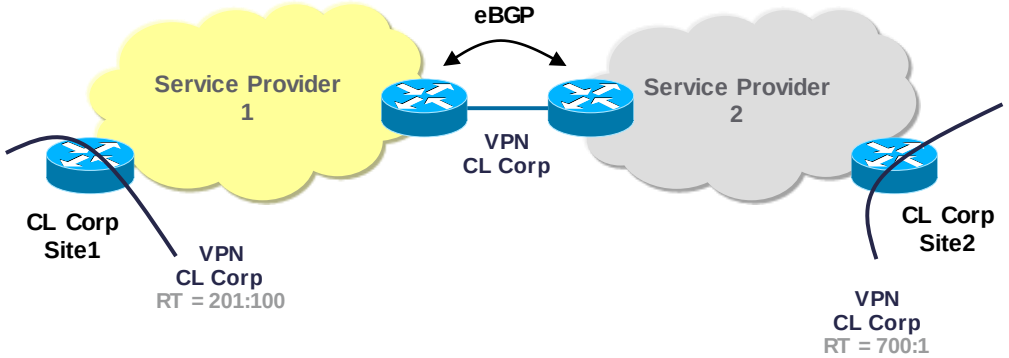




Problem: Incompatible RTs Between Two Different Service Providers

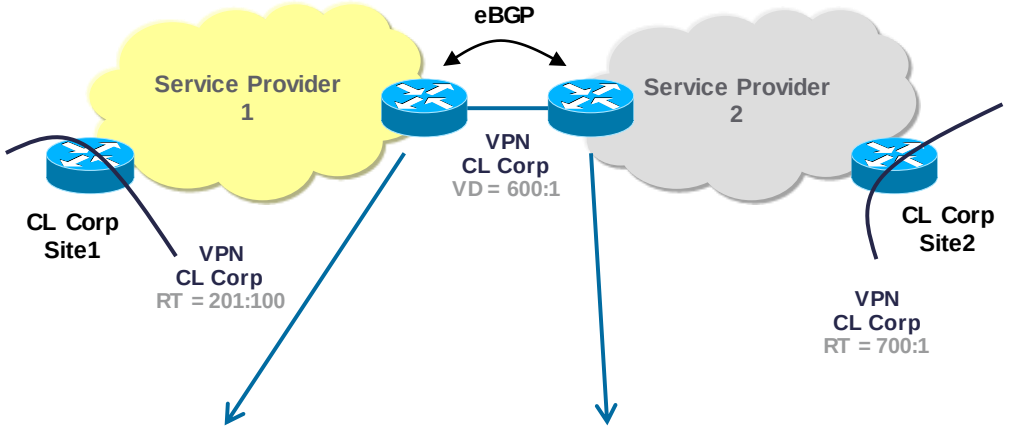
Problem: Incompatible RTs Between Two Different Service Providers

- Single VRF running across 2 SPs
- Route Target used in SP1 may be unsuitable in SP2
- Providers may want to keep Route Targets private with an AS.



Solution: RT / VPN Distinguisher Attribute Rewrite

- VPN Distinguisher (VD) Attribute exchanged via eBGP allows to keep RT private per AS
- Enhancement also allows RANGE statement for VD <-> RT mapping.



IOS: 15.3T
IOS-XE: 3.8
IOS-XR: 3.4.0

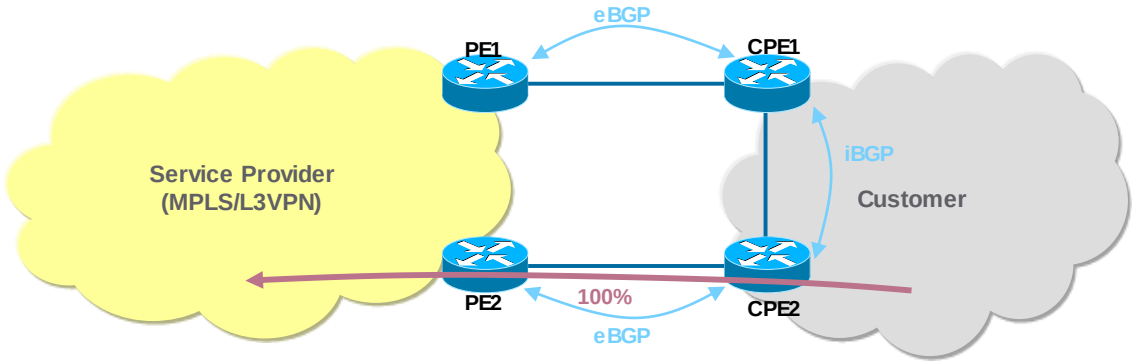
```
ip extcommunity-list 22 permit rt 201:100
!
route-map rt-mapping permit 10
 match extcommunity 22
 set extcomm-list 22 delete
 set extcommunity vpn-distinguisher 600:1
!
route-map rt-mapping permit 20
!
router bgp 3000
 neighbor 192.168.103.1 remote-as 3000
 address-family vpnv4
  neighbor 192.168.103.1 activate
  neighbor 192.168.103.1 route-map rt-mapping out
  exit-address-family
!
```

```
ip extcommunity-list 101 permit VD:600:1
!
route-map rtmap2 permit 10
 match extcommunity 101
 set extcomm-list 101 delete
 set extcommunity rt 700:1 additive
!
route-map rtmap2 permit 20
!
router bgp 4000
 neighbor 192.168.0.50 remote-as 4000
 address-family vpnv4
  neighbor 192.168.0.50 activate
  neighbor 192.168.0.50 route-map rtmap2 in
  exit-address-family
!
```

Problem: Load-Balancing Across Multiple Uplinks

Problem: Load-Balancing Across Multiple Uplinks

- 2 Uplink to Service Provider on different CEs
- BGP best path selection prefers eBGP over iBGP
- How can we load-balance across both paths?

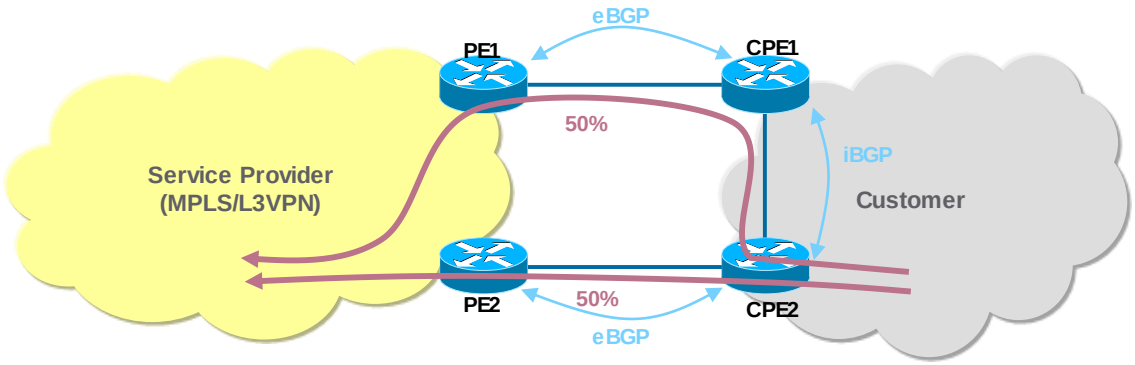


Solution: eiBGP Multipath for Non-VRF Interfaces

- eiBGP multipath removes this criteria in BGP path selection mechanism

- ECMP hash load balancing mechanism will forward 50% of traffic over iBGP and 50% over eBGP.

IOS: 15.4T
IOS-XE: 3.10
IOS-XR: 4.2.0
NX-OS: 6.0




```
!
router bgp 64496
 address-family ipv4 unicast
  maximum-paths eibgp 4
!
 address-family ipv6 unicast
  maximum-paths eibgp 4
!
```

Problem: Support for iBGP Between Customer and Service Provider

Problem: Support for iBGP Between Customer and Service Provider

- Require customer iBGP attributes to be retained across the VPN e.g. LOCAL_PREF, ORIGINATOR_ID, CLUSTER_ID and CLUSTER_LIST
- Need to configure remote sites within single AS
- No insertion of ISP BGP AS number in the AS_PATH.

Solution: L3VPN iBGP PE-CE (RFC 6368)

IOS: 15.4T
IOS-XE: 3.10
IOS-XR: 5.2.2

- Support for RFC 6368
- The PE will place the received iBGP attributes in a new attribute ATTR_SET and transport them over the ISP backbone
- Identify within the VPN the iBGP L3VPN characteristics
- Note: requires unique RDs per PE.

```
router bgp 100
 address-family ipv4 vrf blue
  neighbor 10.0.0.1 remote-as 200
  neighbor 10.0.0.1 local-as 200
  neighbor 10.0.0.1 internal-vpn-client
  neighbor 10.0.0.1 route-reflector-client
```

Problem: Attributes For An eBGP RR Client Need To Be Modified

Problem: Attributes for an eBGP RR Client need to be Modified

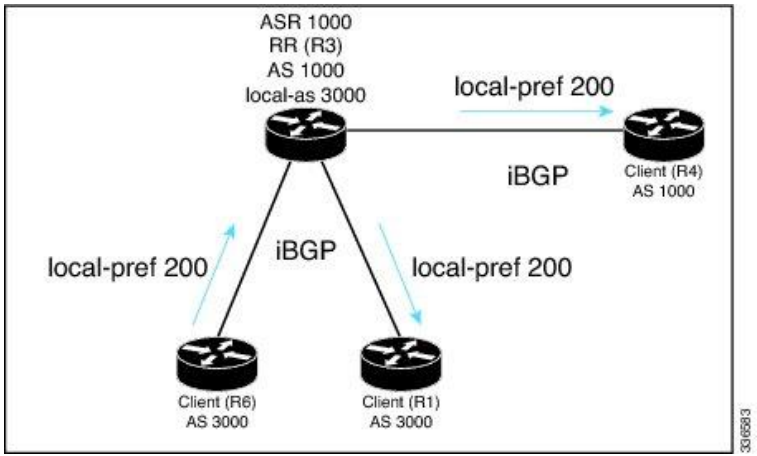
- How do we customise iBGP attributes such as LOCAL_PREF, ORIGINATOR_ID, CLUSTER_ID and CLUSTER_LIST?

- Useful for AS number migration.

Solution: Local-AS for iBGP and Allow-Policy

- **neighbour local-as** can now be used to enable sending of iBGP attributes (LOCAL_PREF, ORIGINATOR_ID, CLUSTER_ID and CLUSTER_LIST) over an iBGP local-AS session
- The flexibility to modify these attributes is achieved by configuring the **neighbour allow-policy** command on RR.

IOS: 15.4T
IOS-XE: 3.9



Problem: My RR Doesn’t Support EVPN

Problem: My RR Doesn't Support EVPN / PBB EVPN

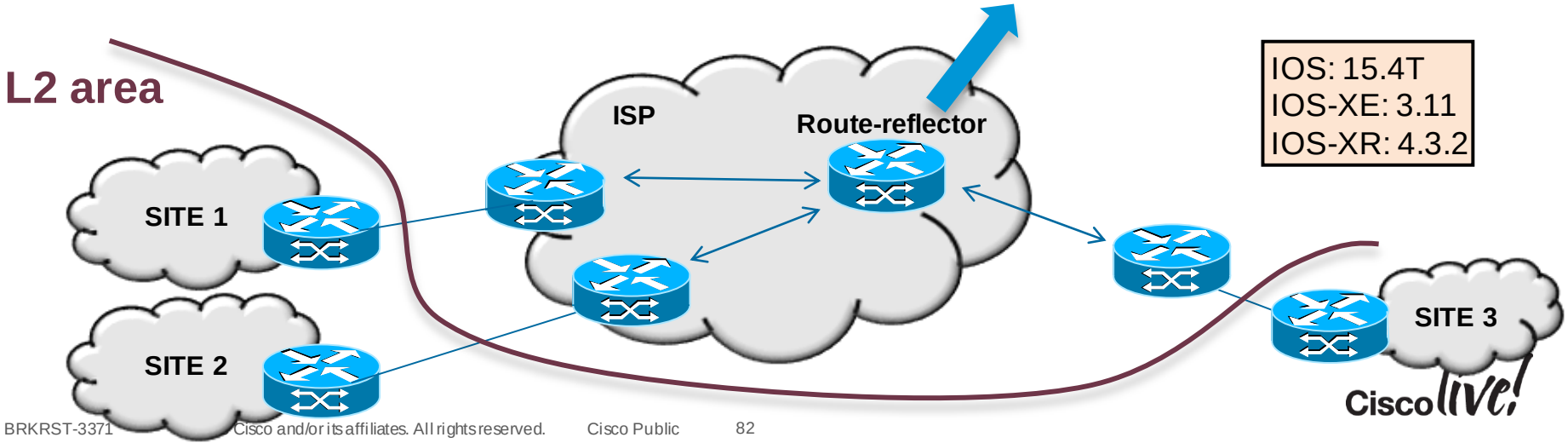
- EVPN is a feature to replace pseudowires for L2VPN
- MAC learning done in control-plane (BGP) rather than data-plane (flooding)
- Provider Backbone Bridging (PBB) allows logical separation of L2 domains (MAC-in-MAC)
- Uses route-targets to import/export Ethernet Virtual Instances (EVI)
- More scalable solution
- Previously no RR support for "l2vpn evpn" address-family.

Solution: EVPN / PBB-EVPN Route-Reflection

- EVPN Address Family is now allowed on both iBGP as well as eBGP neighbours under default VRF for both IPv4 and IPv6 neighbours.

Configuration on RR:

```
router bgp 1
 address-family l2vpn evpn
  neighbor 192.168.1.1 remote-as 1
  neighbor 192.168.1.1 route-reflector-client
  neighbor 192.168.1.1 send-community extended
  .....
```




Security/Operations

Problem: Malformed BGP Updates

Problem: Malformed BGP Updates

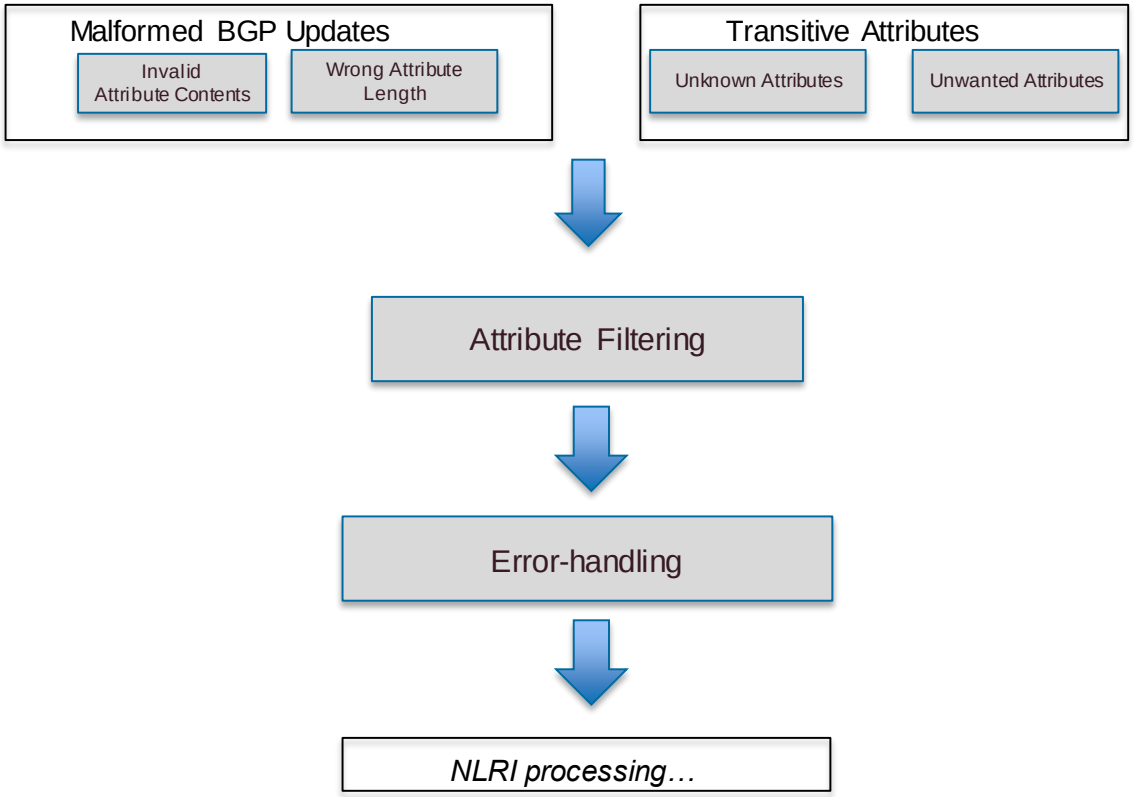
- May contain incorrect formatting or unknown attributes
- Default behaviour is to reset neighbour session resulting in potential network outage
- Caused by interoperability issue or DoS attack.

Solution: Attribute Filtering and Error-Handling for Malformed Updates

IOS: 15.3T
IOS-XE: 3.7
IOS-XR: 4.2.3

- Attribute filtering
  - Attributes can be filtered before NLRI is processed
  - Actions can be:
    - Treat as withdraw
    - Discard attribute
  - Update can be stored for further debugging and syslog generated
- Error-handling
  - Classifies errors based on various categories such as severity, likelihood of occurrence or type of attribute
  - Changes default behaviour to gracefully fix or ignore non-severe errors where possible
  - Avoid session resets for most cases.

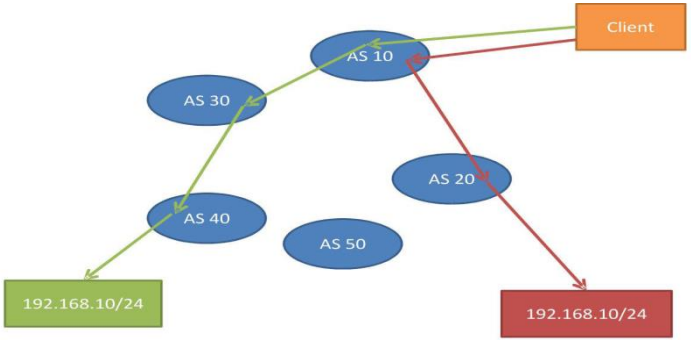
Architecture

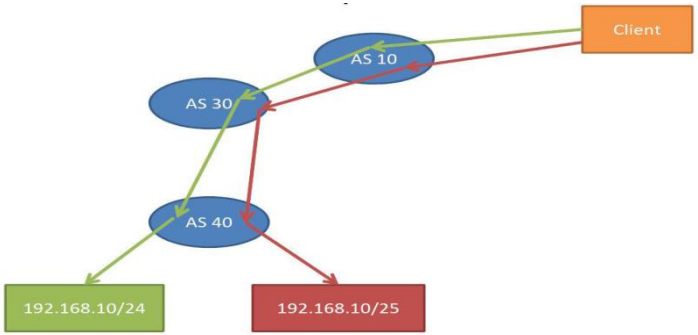


Problem: Prefix Hijacking

Problem: Prefix Hijacking

- Announce someone else's prefix
- Announce a more specific mask for someone else's prefix
- Either way, you are trying to "steal" someone else's traffic by getting it routed to you
  - Capture, sniff, redirect, manipulate traffic as you wish.





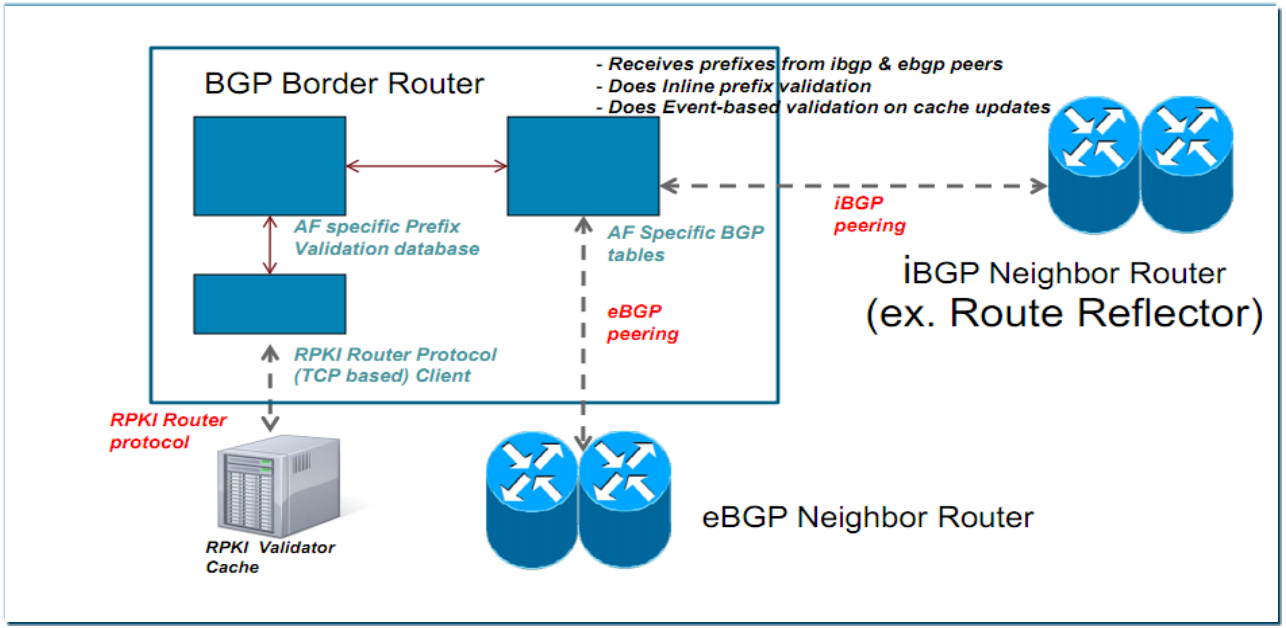
Source: nanog 46 preso

Solution: BGP Origin Validation

IOS: 15.2S
IOS-XE: 3.5
IOS-XR: 4.2.1

- Support client functionality of RPKI RTR protocol
  - Separate database to store record entries from the cache
- Announce path validation state to iBGP neighbours using a well known extended community. Paths can be:
  - Valid
  - Invalid
  - Unknown
- Can modify route policies to incorporate path validation states
- Must register prefixes with Internet Registry.

What Does the Solution Look Like?



RPKI Service information from APNIC at
https://www.apnic.net/services/services-apnic-provides/resource-certification/RPKI

Sample Configuration

```
router bgp 64726
 bgp rpki server tcp 10.1.2.3 port 30000 refresh 60
 bgp bestpath prefix-validate allow-invalid
 neighbor 10.9.9.9 remote-as 64209
 neighbor 10.9.9.9 route-map FOO in
!
route-map FOO permit 10
 match rpki invalid
 set local-preference 50
route-map FOO permit 20
 match rpki not-found
 set local-preference 100
route-map FOO permit 30
 match rpki valid
 set local-preference 200
route-map FOO permit 40
!
```

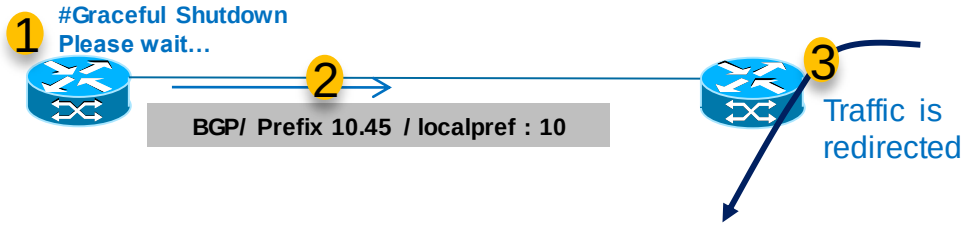
Problem: How Do I Minimise Outages During A Planned Maintenance Window?

Solution: BGP Graceful Shutdown

- RFC 6198 – April 2011
- Old Behaviour
  - If session drops then BGP will withdraw all prefixes learned over that session
  - BGP has no mechanism to signal prefix will soon be unreachable (for maintenance for example)
- Historically RR's have worsened the issue as they tend to hide the alternate path as they only forward the best path.

IOS: 15.4T
IOS-XE: 3.11

BGP Graceful Shutdown allows to do maintenance on router without service disruption.



This new knob allows a router to notify neighbour to redirect traffic to other paths and after some time will drop BGP sessions.

The notification could be done using Local Preference attribute or user community attribute

BGP Graceful Shutdown

- GSHUT well-known community

- The GSHUT community attribute is applied to a neighbour specified by the **neighbour shutdown graceful** command, thereby gracefully shutting down the link in an expected number of seconds.

```
neighbour {ipv4-address | ipv6-address | peer-group-name} shutdown graceful seconds {community value [local-preference value] | local-preference value}
```

http://www.cisco.com/en/US/docs/ios-xml/ios/iproute_bgp/configuration/15-s/irg-15-s-book.pdf

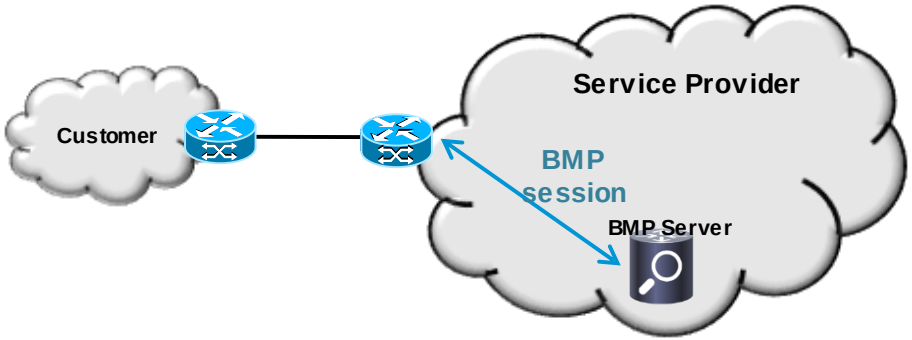
Cisco live!

Problem: How Do I Monitor My BGP Speakers?

Solution: BGP Monitoring Protocol (BMP)

IOS: 15.4T
IOS-XE: 3.11
IOS-XR: 5.2.2

- BMP is intended to be used for monitoring BGP sessions
- BMP is not impacting the routing decision process and is only used to provide monitoring information
- BMP provides access to the Adj-RIB-In of a BGP peer on an ongoing basis and provides a periodic dump of statistical information. A monitoring station can use this for further analysis.



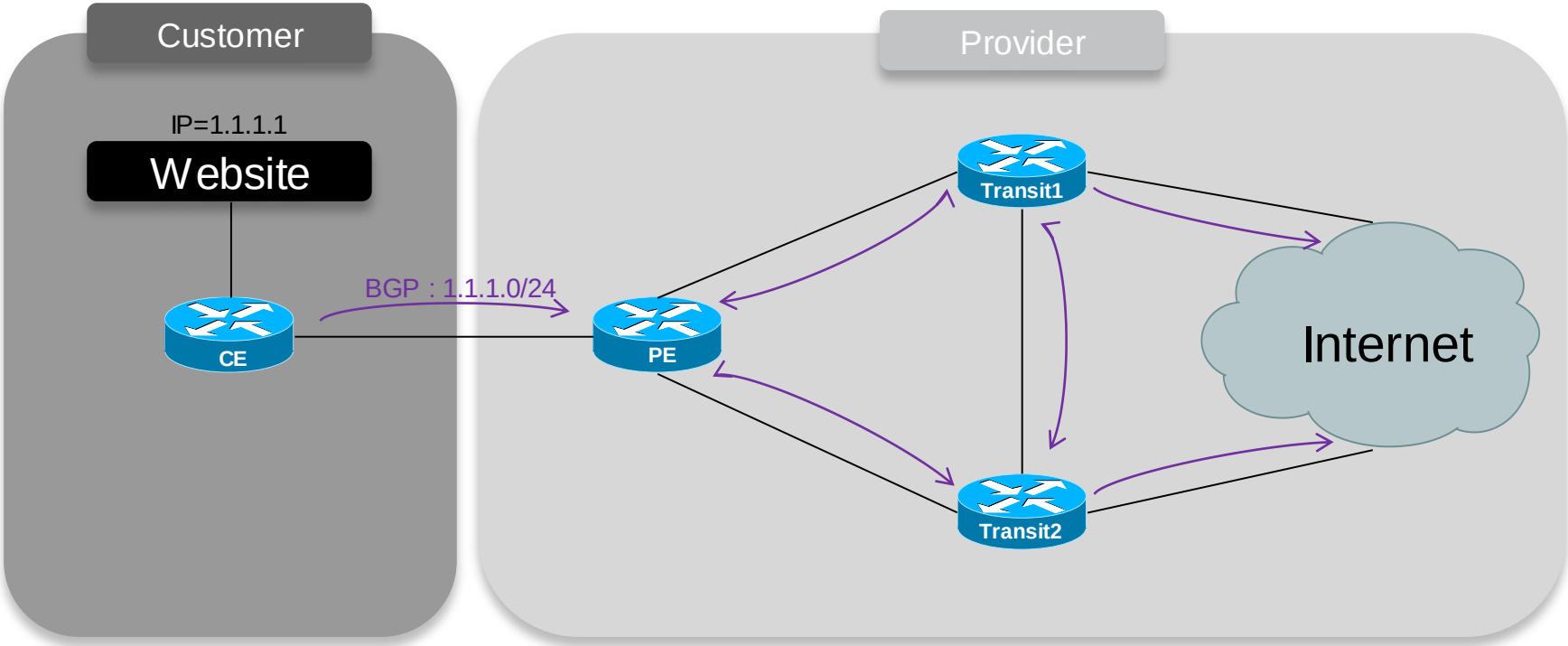
- Configuration
  - Enable monitoring per neighbour
  - Configure the BMP servers

```
!
router bgp 65000
 neighbor 30.1.1.1 bmp-activate server 1
!
bmp server 1
 activate
 address 10.1.1.1 port-number 8000
 description LINE SERVER1
 failure-retry-delay 40
 flapping-delay 120
 initial-delay 20
 set ip dscp 5
 stats-reporting-period 30
 update-source ethernet 0/0
 exit-bmp-server-mode
!
```

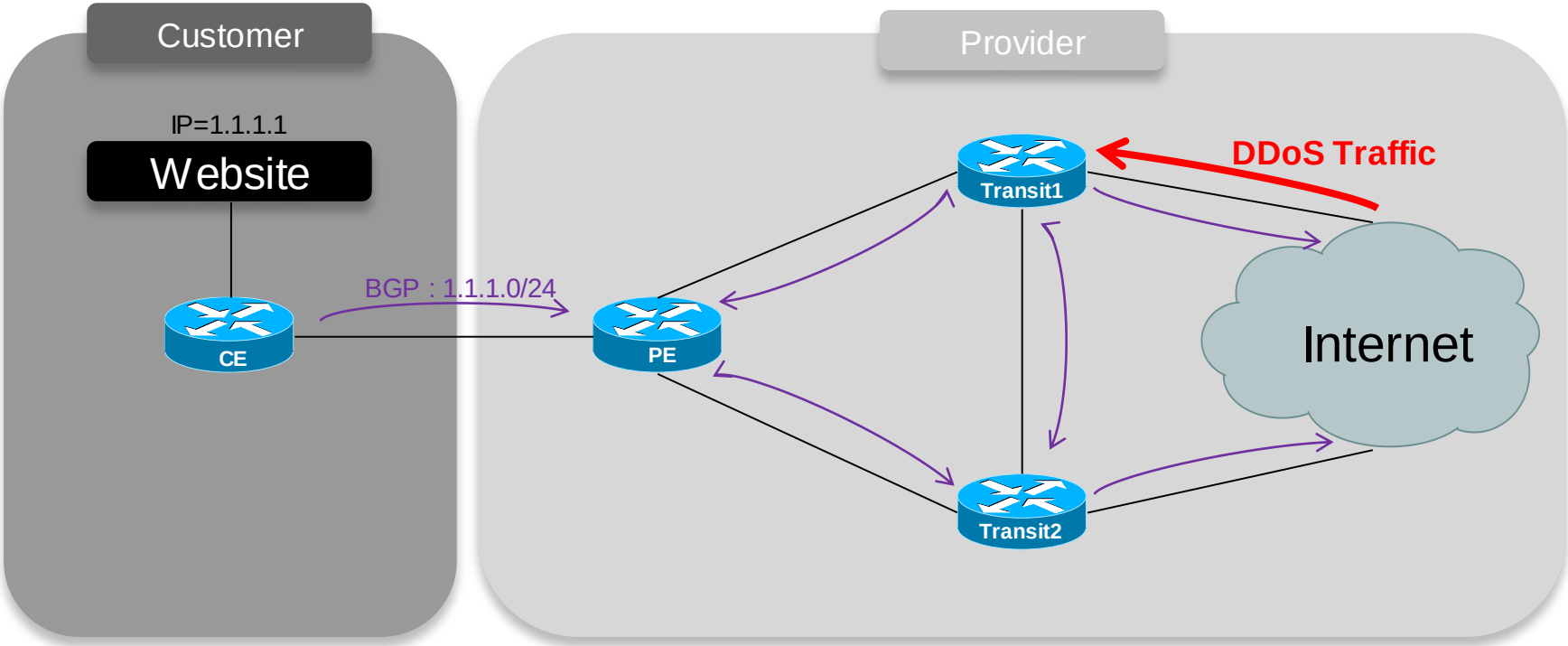
http://tools.ietf.org/html/draft-ietf-grow-bmp-07

Problem: How Do We Mitigate DDoS Attacks?

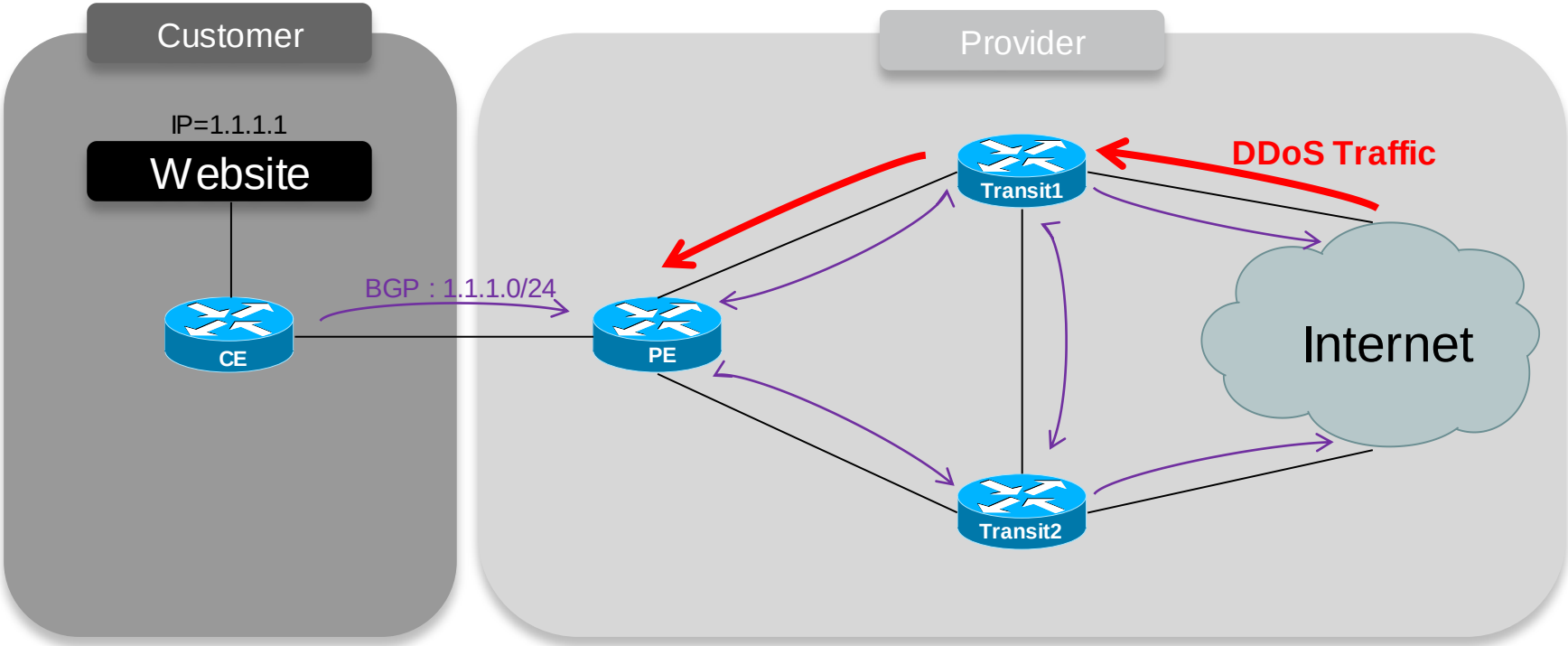
Problem: DDoS Attack



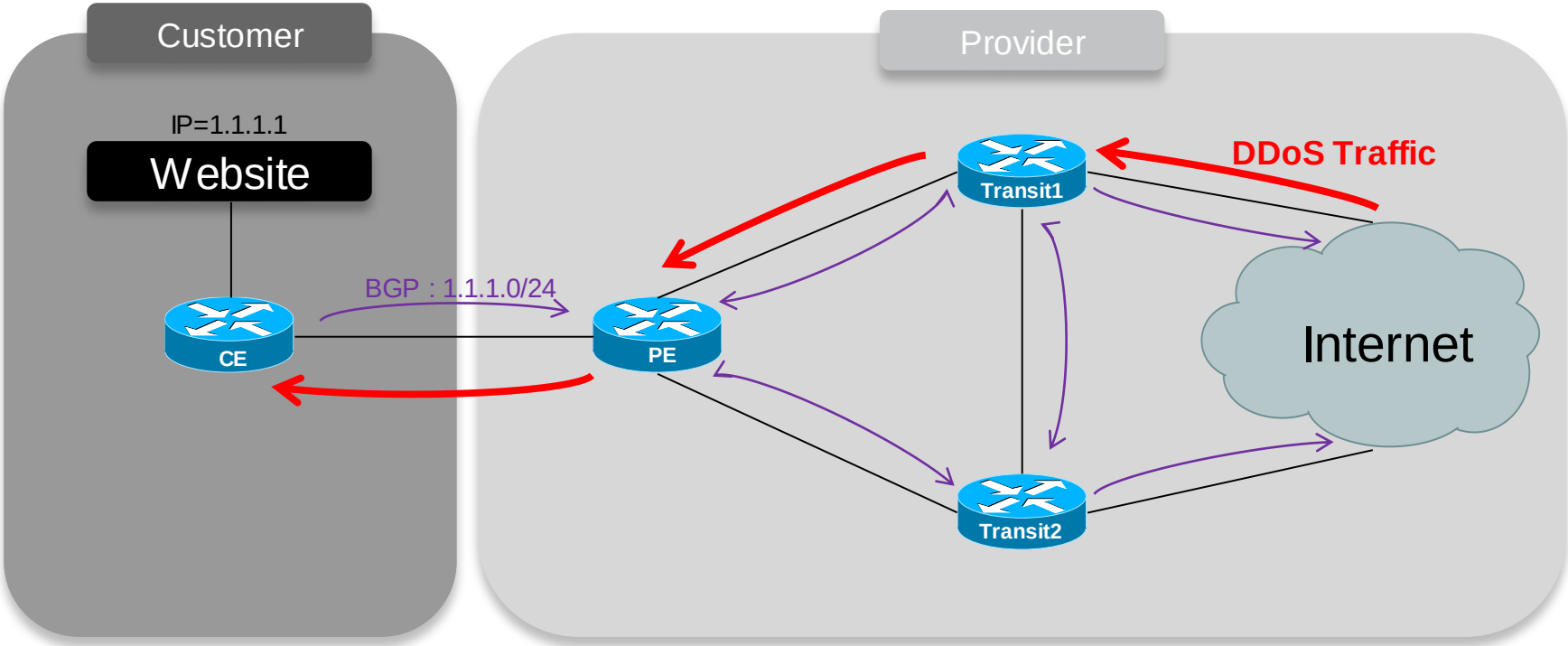
Problem: DDoS Attack



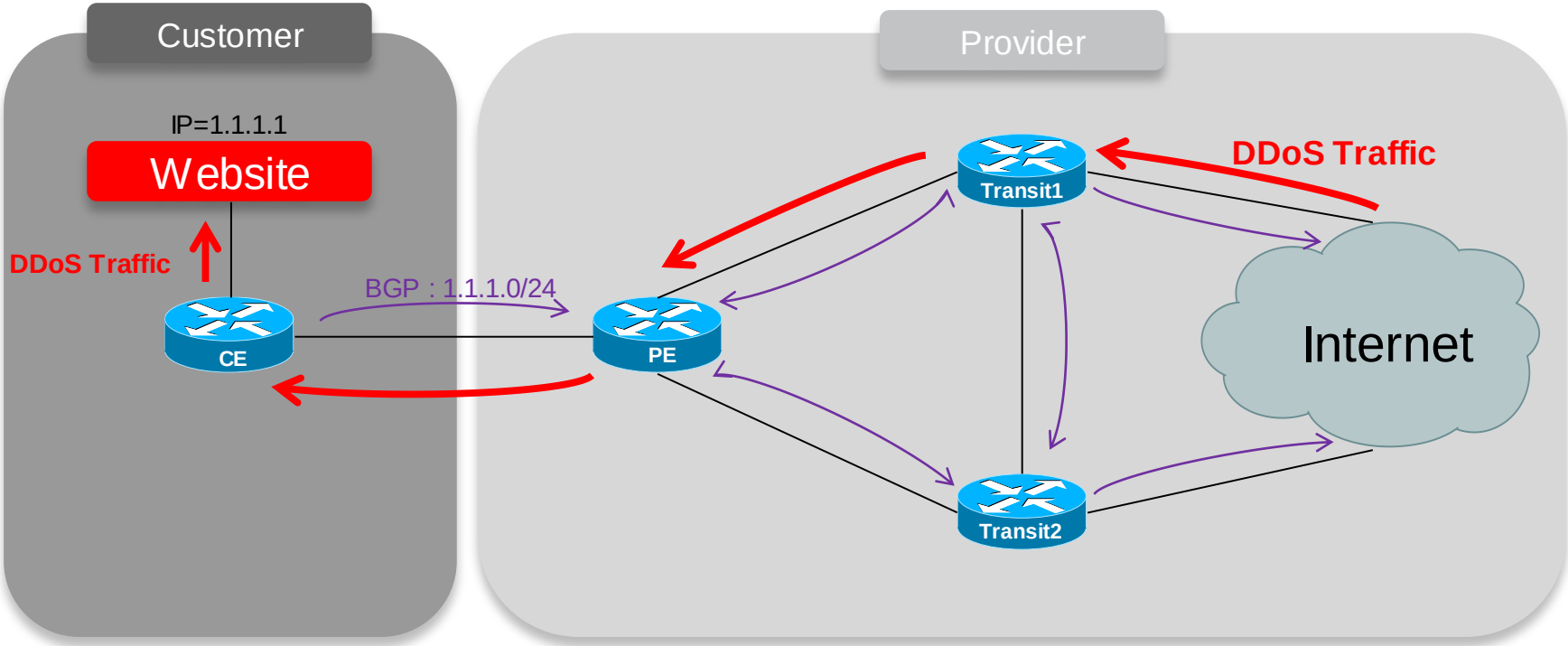
Problem: DDoS Attack



Problem: DDoS Attack

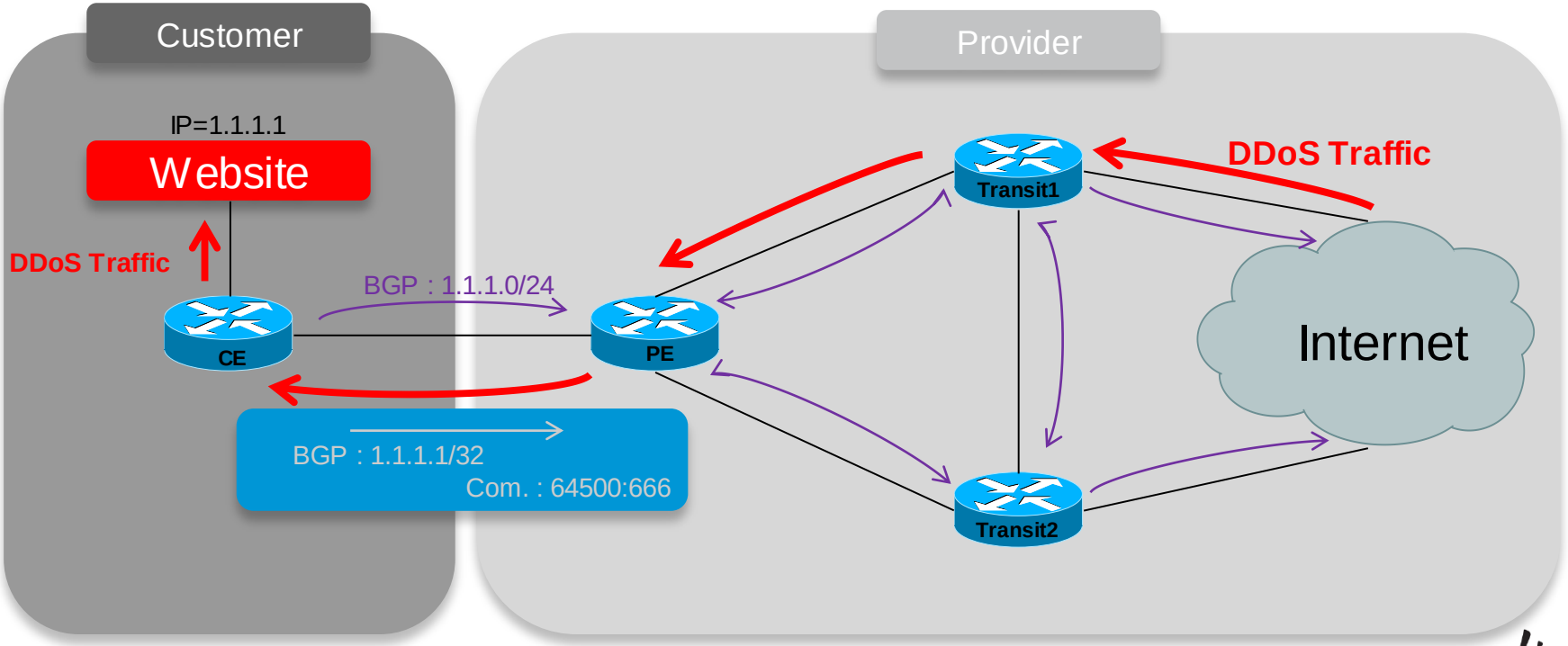


Problem: DDoS Attack



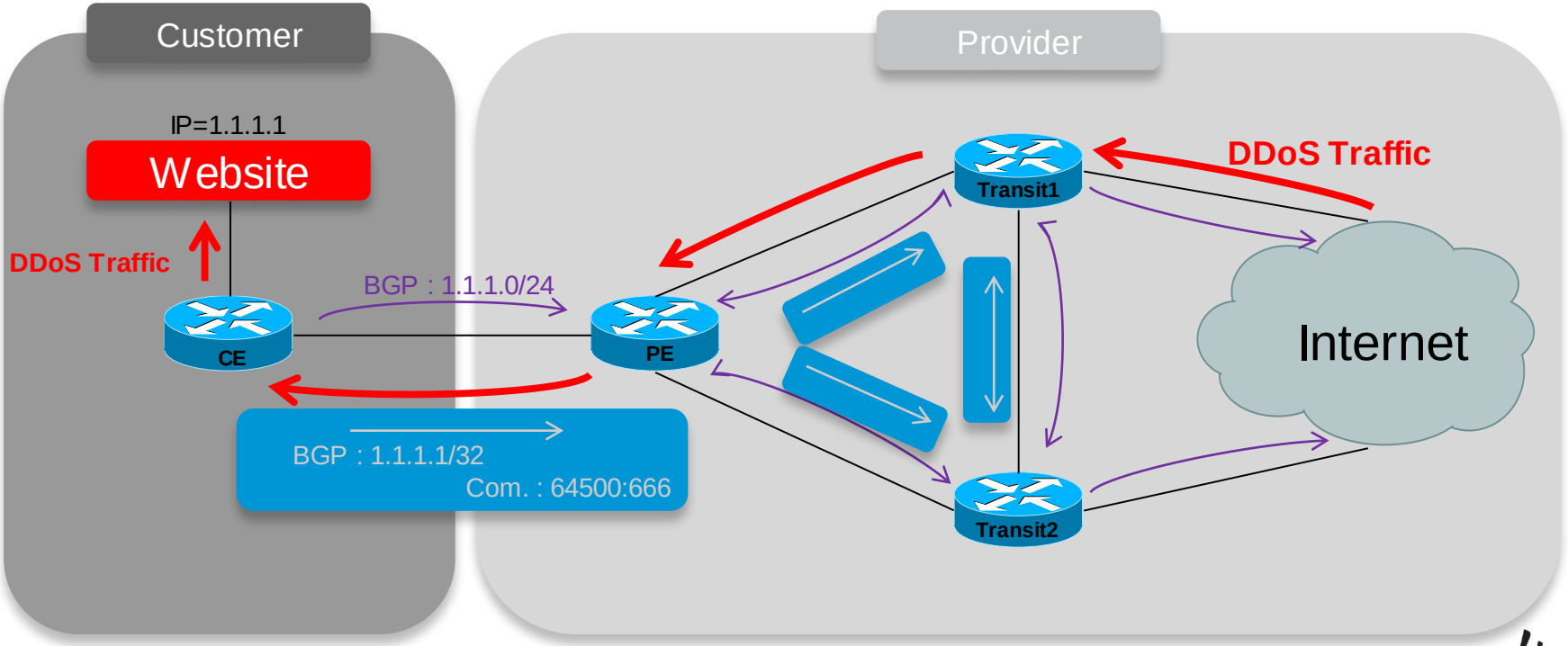
Solution 1: Remotely Triggered Black Hole

It is time to use the blackhole community given by the provider (i.e. 64500:666)



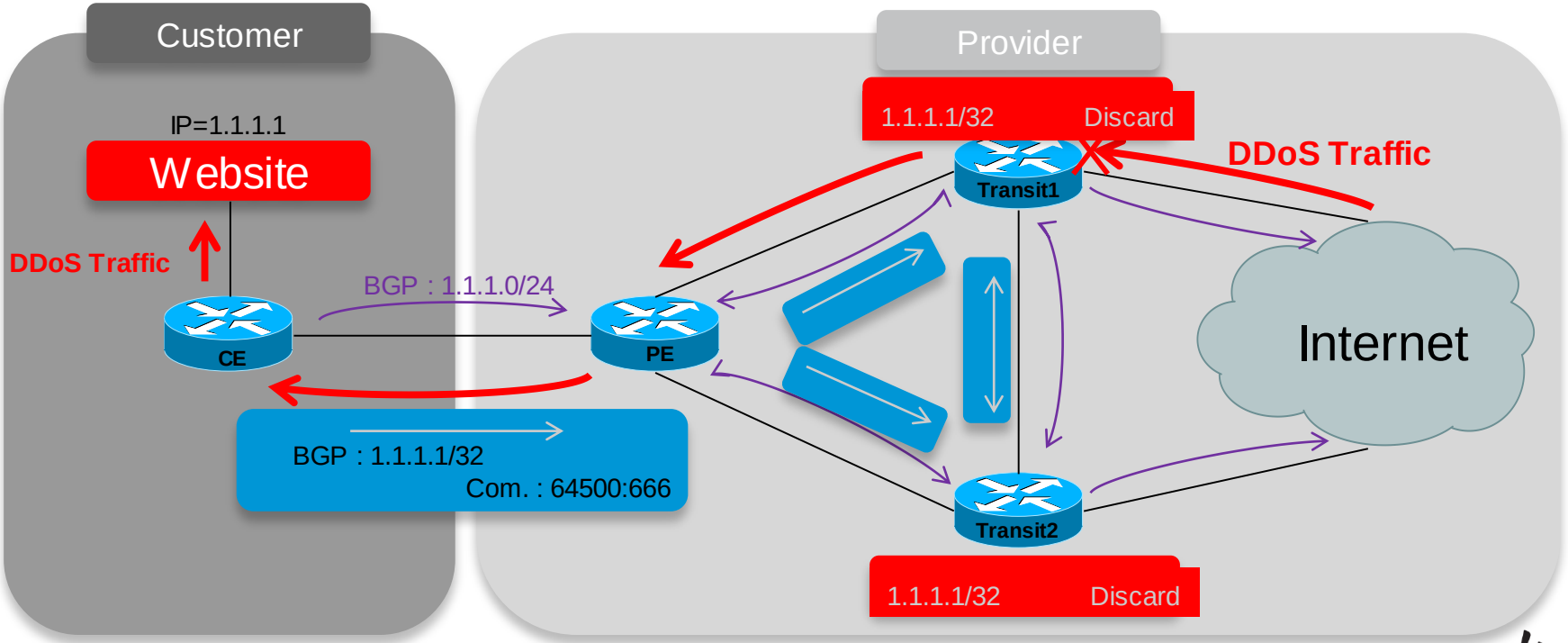
Solution: Remotely Triggered Black Hole

It is time to use the blackhole community given by the provider (i.e. 64500:666)



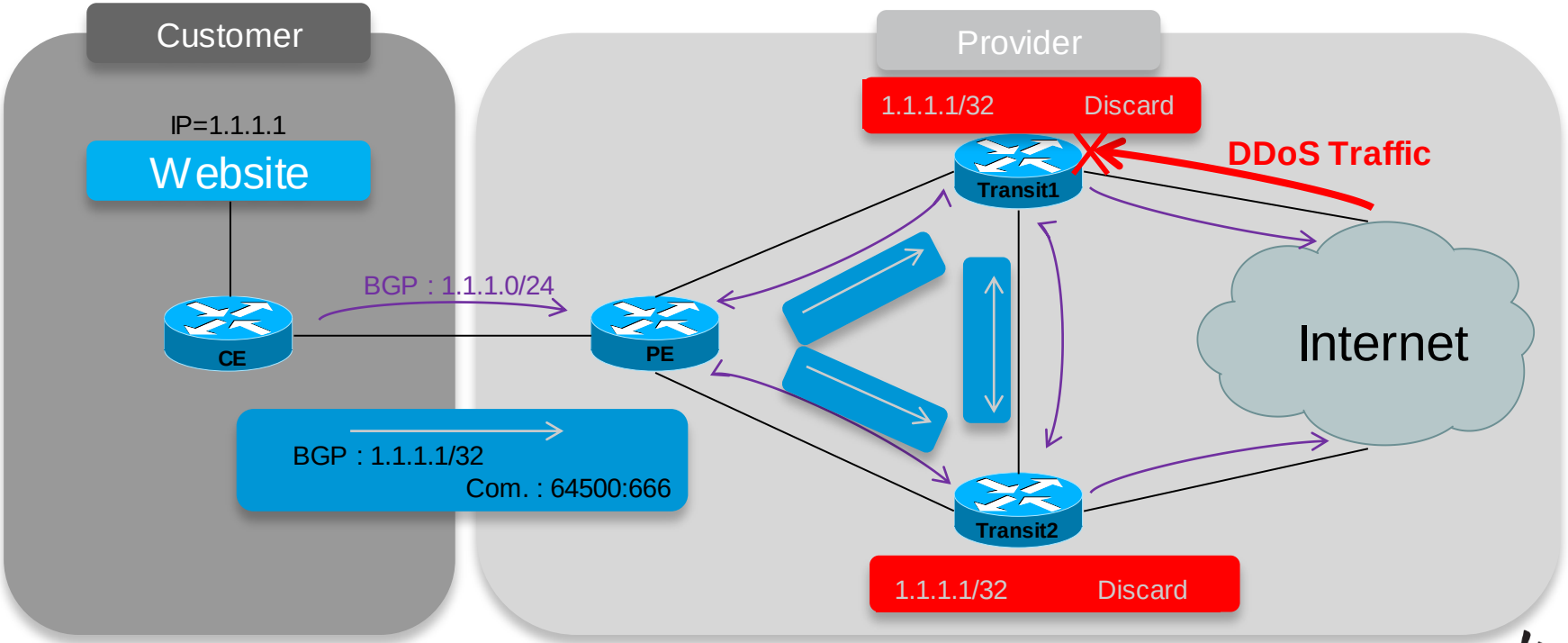
Solution 1: Remotely Triggered Black Hole

All prefixes with blackhole community get assigned a special nexthop which recurses to Null0



Solution 1: Remotely Triggered Black Hole

All prefixes with blackhole community get assigned a special nexthop which recurses to Null0



Solution 1: Remotely Triggered Black Hole

- Great, I have my server responding again!
  - No more DDoS traffic on my network
  - **But** no more traffic at all on my website…
- Well, maybe it was not the solution I was looking for…

Solution 2: Policy Based Routing

- Identification of DDoS traffic: based around a conditions regarding MATCH statements
  - Source/Destination address
  - Protocol
  - Packet size
  - Etc…
- Actions upon DDoS traffic
  - Discard
  - Logging
  - Rate-Limiting
  - Redirection
  - Etc…
- Doesn't this sound like a great solution?

Solution 2: Policy Based Routing

- Good solution for
  - Done with hardware acceleration on carrier grade routers
  - Can provide better precision of match statements and actions to impose.

- But…
  - Customer need to call my provider
  - Customer need the provider to accept and run this filter on each of their backbone/edge routers
  - Customer need to call the provider and remove the rule after!

- Reality: It won't happen…

Solution 3: BGP FlowSpec as a Better Alternative

- Comparison with the other solutions
  - Makes static PBR a dynamic solution!
  - Allows to propagate PBR rules
  - Existing control plane communication channel is used

- How?
  - By using your existing MP-BGP infrastructure

IOS: 15.5S (RR)
IOS-XE: 3.14 (RR)
IOS-XR: 5.2.0

BGP Flowspec: Introduction

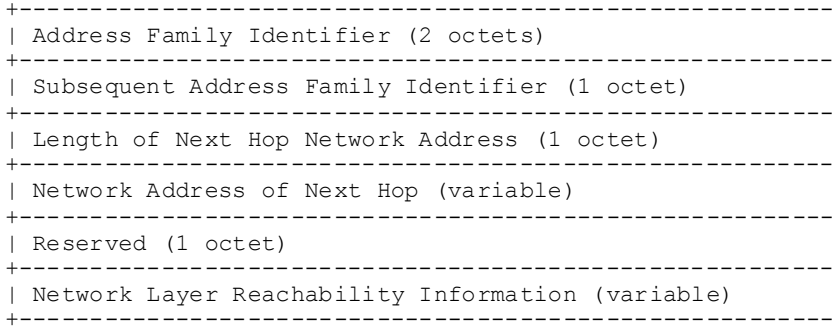
- BGP (like any other routing protocol) influences destination-based routing
- BGP routing information can be injected from a central place ("route server")
- Why not use it for more than just giving a destination address to route packets to?

- "Flow Specification Rules"
  - Application aware Filtering/redirect/mirroring
  - Dynamic and adaptive technology
  - Simple to configure

Dissemination of Flow Specification Rules
(RFC5575)

New NLRI defined (AFI=1, SAFI=133)

1. Destination IP Address
2. Source IP Address
3. IP Protocol
4. Port
5. Destination port
6. Source Port
7. ICMP Type
8. ICMP Code
9. TCP Flags
10. Packet length
11. DSCP
12. Fragment



The MP_REACH_NLRI – RFC 4760

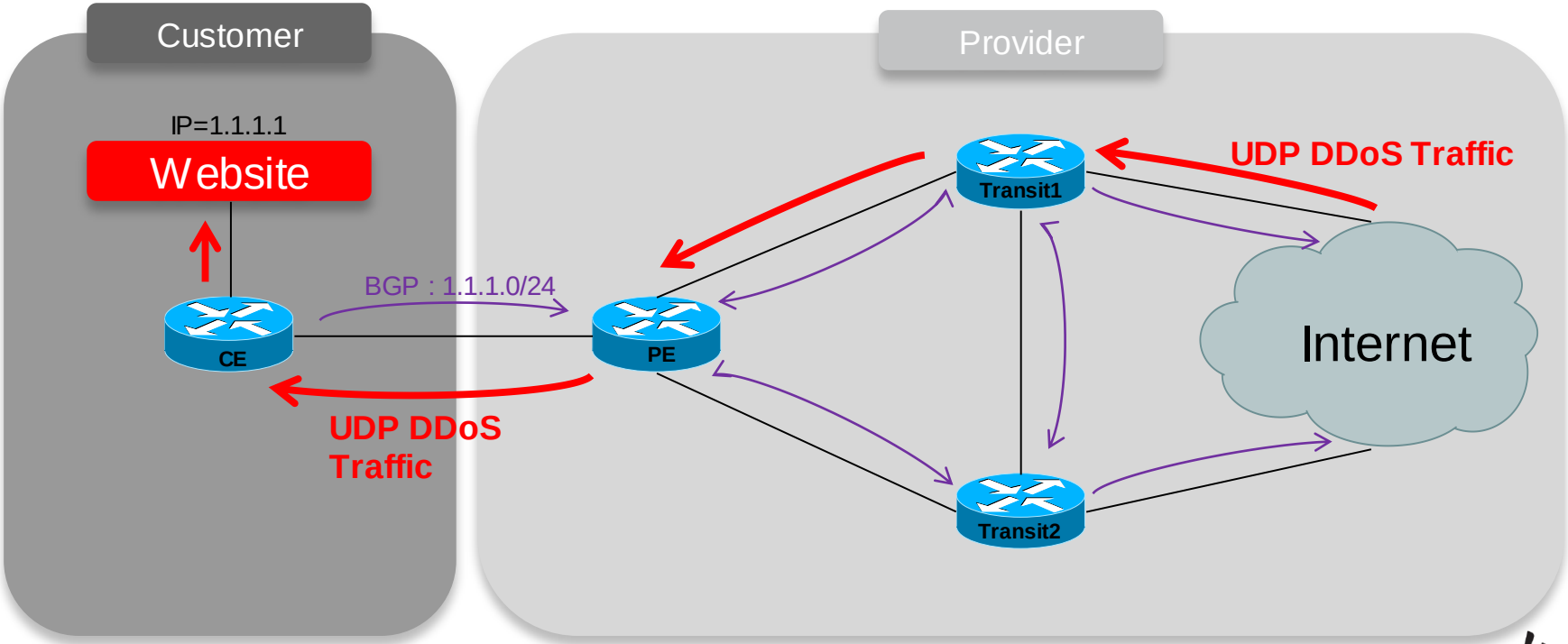
Notice from the RFC: “Flow specification components must follow strict type ordering. A given component type may or may not be present in the specification, but if present, it MUST precede any component of higher numeric type value.”

BGP Flowspec: Traffic Actions

Action	Description
Traffic-Rate	Ability to police flow to a given amount
Traffic-Marking	Rewrite DSCP value
Redirect VRF	Redirect to a VRF (using route-target) Ex: “cleaning” traffic
Redirect NH	Redirect to an alternate next-hop
Traffic-Action	Drop/Discard or Sample (not yet implemented)

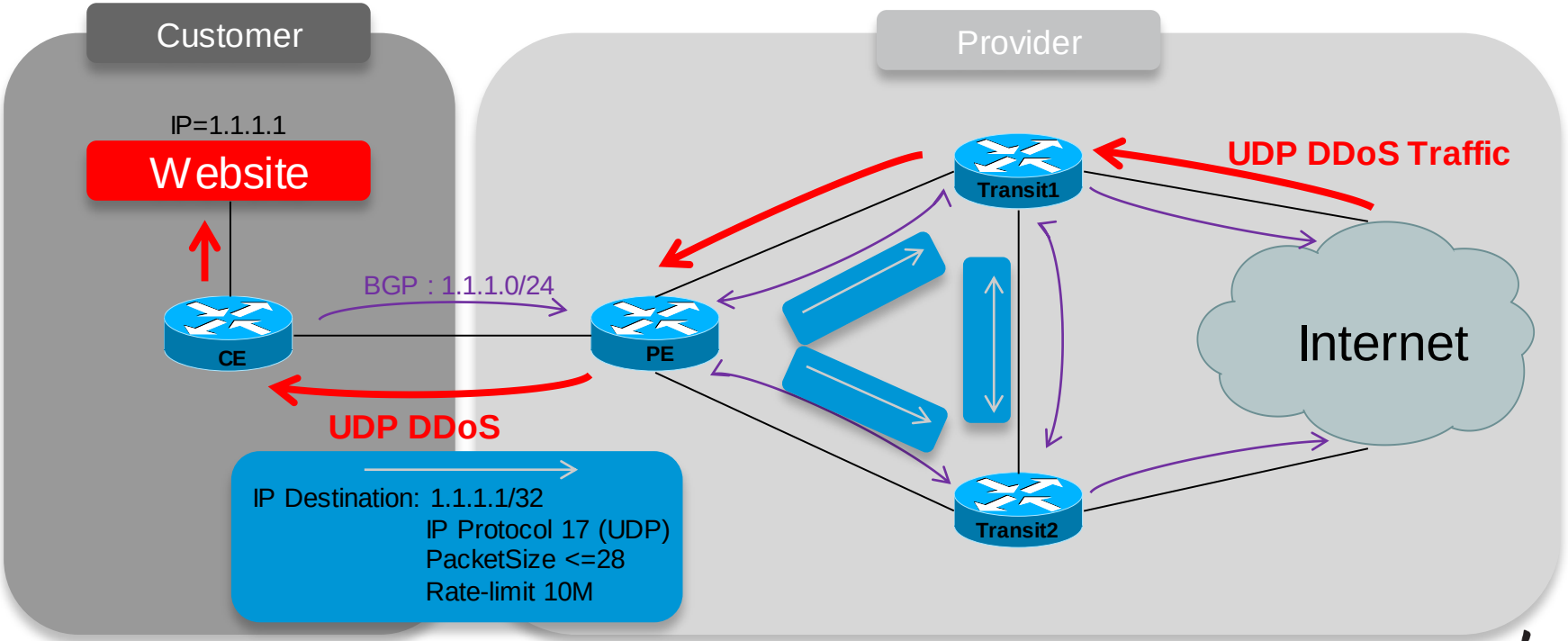
DDoS Mitigation using BGP FlowSpec

Let's do this better now with the new BGP FlowSpec functionality

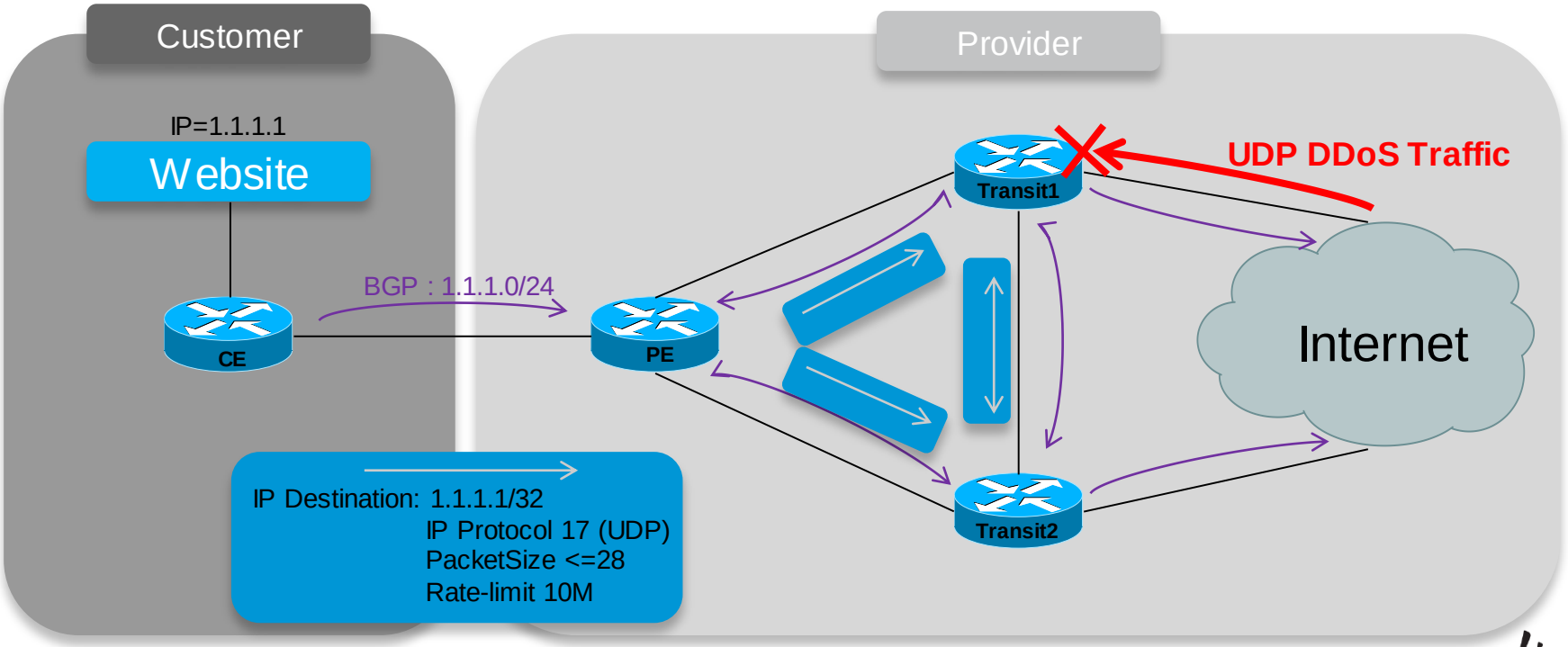


DDoS Mitigation using BGP FlowSpec

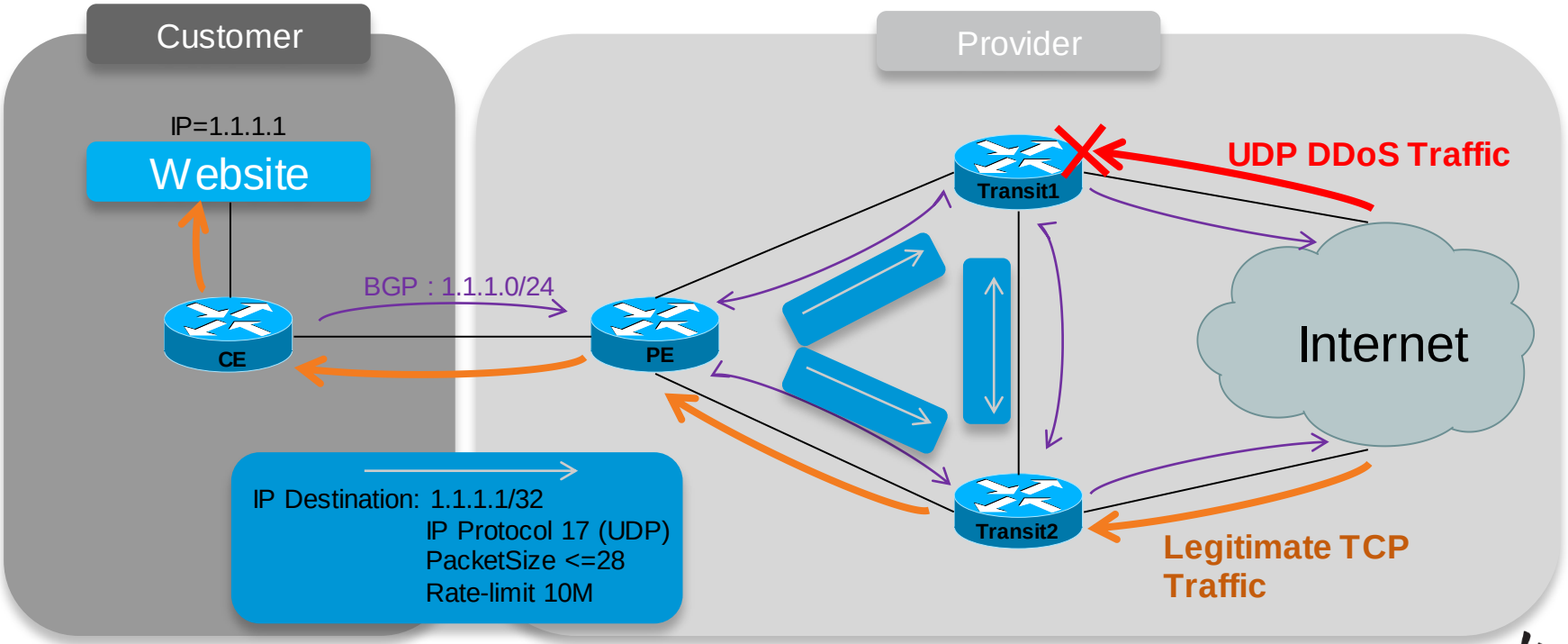
Customer advertises the web server's address with granular flow information



DDoS Mitigation using BGP FlowSpec



DDoS Mitigation using BGP FlowSpec

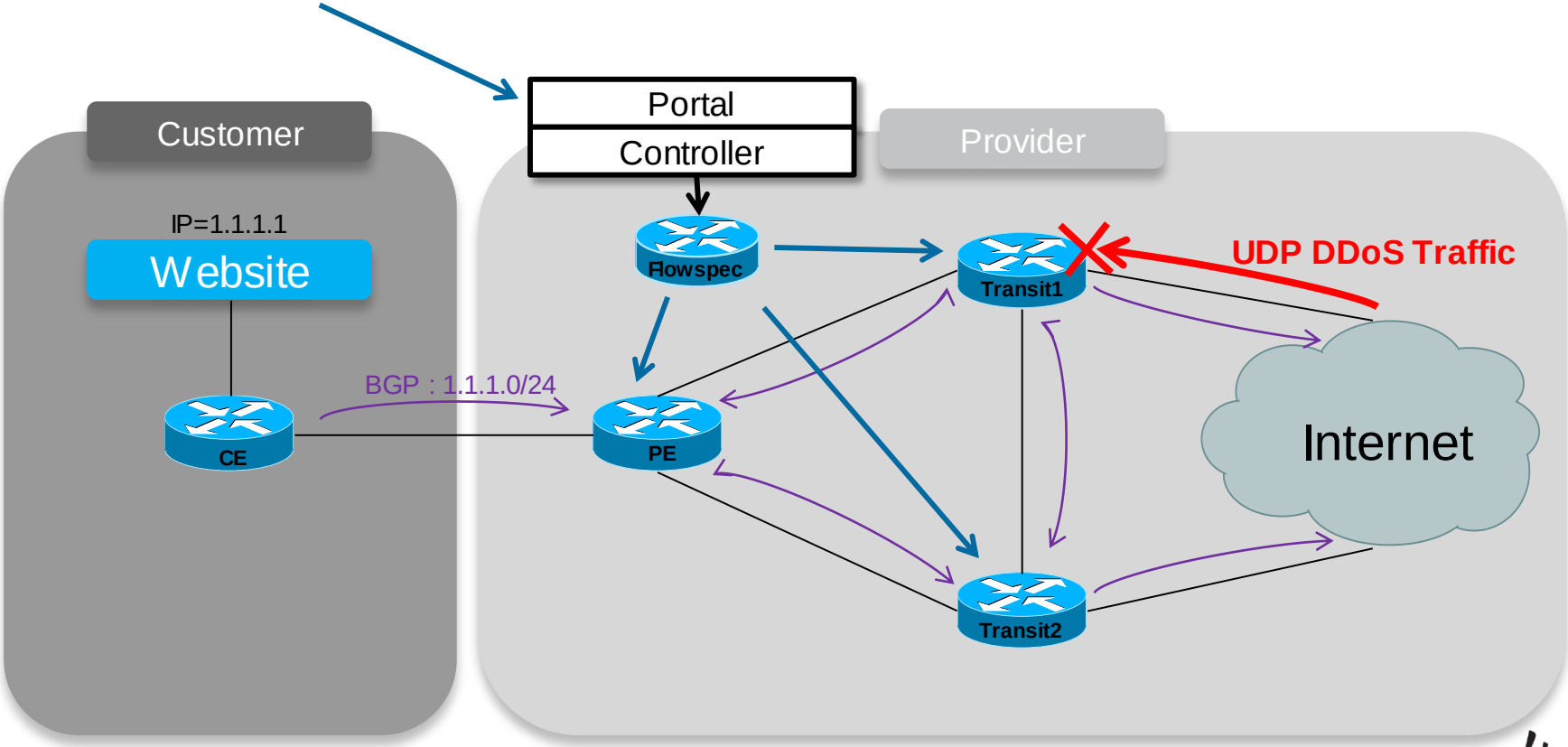


Real Life Architecture

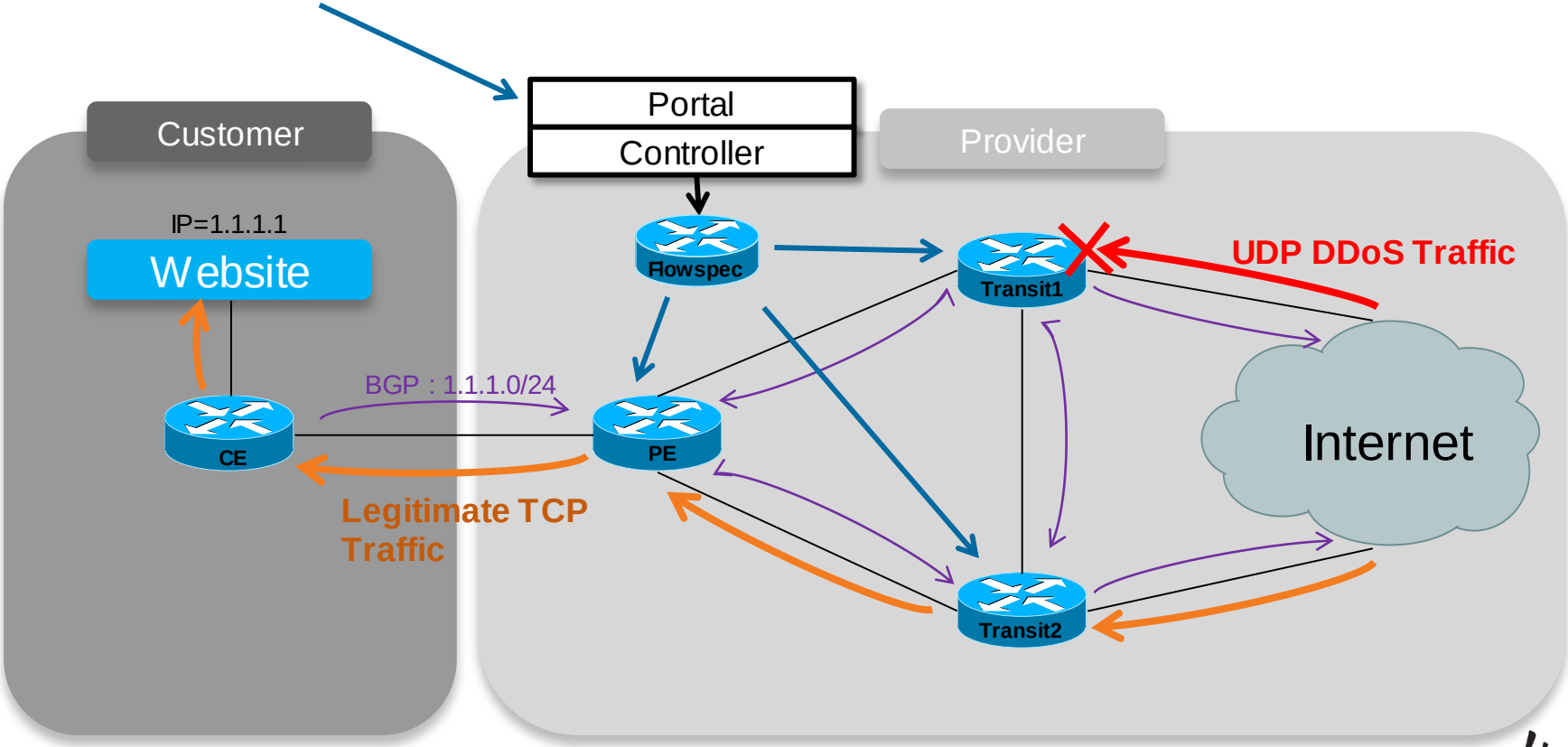
- In reality this architecture is not deployed
  - Service Provider DO NOT trust the Customer (at least not that much ;-)
  - It requires new BGP AFI/SAFI combination to be deployed between Customer and Service provider
  - Both these result in Flowspec not commonly being deployed between Customer and SP

- What is done instead?
  - SP utilise a central Flowspec speaker(s)
  - Have it BGP meshed within the Service Provider routers
  - Only the central Flowspec speaker is allowed to distribute Flowspec rules
  - Central Flowspec speaker is considered “trusted” by the network (no-validate)
  - Central Flowspec speaker is managed by the service provider.

Central FlowSpec Speaker



Central FlowSpec Speaker



Agenda

- Introduction
- Motivation to Enhance BGP
- What's happened in the BGP Landscape?
- Some new cool features that may interest you

Continue Your Education

- Demos in the Cisco Campus
- Walk-in Self-Paced Labs
- Table Topics
- Meet the Engineer 1:1 meetings

Cisco live!


Q & A

Complete Your Online Session Evaluation

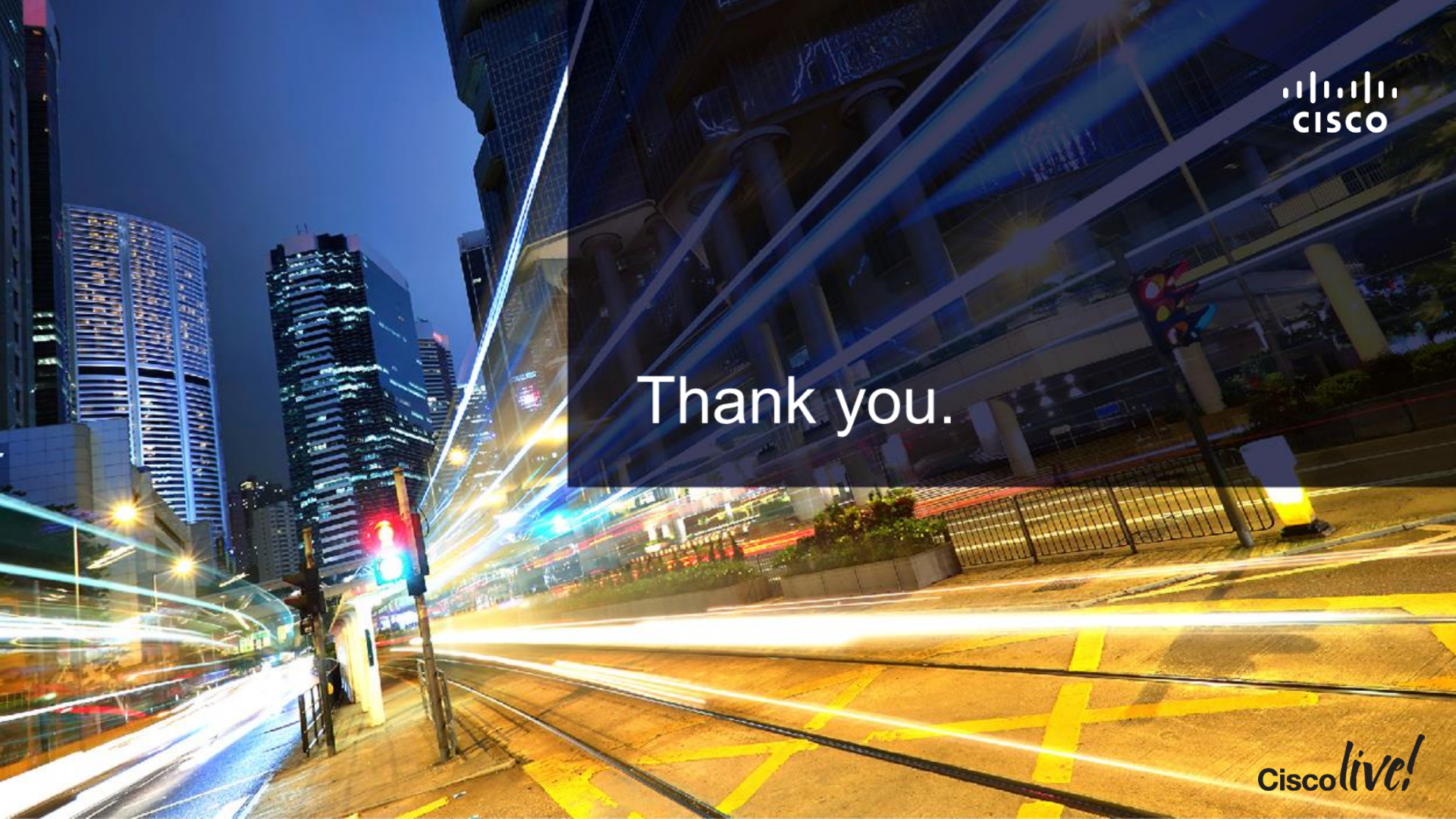
Give us your feedback and receive a Cisco Live 2015 T-Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site http://showcase.genie-connect.com/clmelbourne2015
- Visit any Cisco Live Internet Station located throughout the venue

T-Shirts can be collected in the World of Solutions on Friday 20 March 12:00pm - 2:00pm

Learn online with Cisco Live!
Visit us online after the conference for full access to session videos and presentations. www.CiscoLiveAPAC.com


Thank you.


CISCO