Cisco live!
What You Make Possible
Deploying BGP Fast Convergence / BGP PIC

Oliver Böhmer
Housekeeping

• We value your feedback- don't forget to complete your online session evaluations after each session & the Overall Conference Evaluation which will be available online from Thursday
• Visit the World of Solutions and Meet the Engineer
• Visit the Cisco Store to purchase your recommended readings
• Please switch off your mobile phones
• After the event don’t forget to visit Cisco Live Virtual:  
  www.ciscolivevirtual.com
Agenda

• Introduction / Motivation
• BGP Data Plane Convergence – how does it work?
• Designing for BGP PIC
• Configuring and Monitoring BGP PIC
• Advanced BGP PIC Topics
  Dealing with Summaries/Default Routes
  BGP Label Allocation and Effect on BGP PIC
  MPLS-VPN InterAS and BGP PIC
Loss of Connectivity – LoC

• Link and/or Node Failures cause packet loss until routing has converged
  Loss can be caused by black-holes and/or micro-loops, until all relevant
  nodes in the path have converged
• Time between failure and restoration is called LoC
• How much LoC is acceptable?
  Minutes?
  Seconds?
 Milliseconds?
• How often is LoC acceptable?
LoC Impact on sample applications

- Signalling: 50 msec
- Video: 500 msec
- VoIP call: 1 sec
- Routing Protocols: 5 secs
- L1/L2 Transport Convergence: 30 secs
- Tunnels go down: 1 Minute
- TCP Session dies: 1 Minute
Routing Convergence Building Blocks

- Detection (link or node aliveness, routing updates received)
  - Update topology database
  - Compute optimal path
  - Download to HW FIB
  - Switch to newer path

- State propagation (routing updates sent)
Routing Fast Convergence – IS-IS Test Results

- c12000, IOS

- Sub-second IGP (OSPF/ISIS) convergence can be conservatively achieved
IGP vs. BGP Convergence

• IGP (OSPF/ISIS) deals with hundreds routes
  Max a few thousands, but only a few hundreds are really important/relevant

• BGP is designed to carry millions of routes
  and a few large customers carry that amount of prefixes!

• So? Isn’t that a routing problem in the global Internet, where we don’t really have SLAs for?

• Well, no …
BGP Fast Convergence – Two Sample Use Cases

- Larg(er) Layer 3/MPLS VPN Deployments offering critical Enterprise voice or video services
  - > Couple thousand routes

- Internet-Facing Datacenter/Hosting Facility using Full Internet Routing table
  - n * 350000 (!!) routes received
BGP Control-Plane Convergence Components I: Core failure

1. Core Link or node goes down
2. IGP notices failure, computes new paths to PE1/PE2
3. IGP notifies BGP that a path to a next-hop has changed
4. PE3 identifies affected paths, runs best path, path to PE2 no longer as good as the one to PE1
5. Updates RIB/FIB, traffic continues
BGP Control-Plane Convergence Components II: Edge Node Failure

Edge Node (PE1) goes down

2. IGP notices failure, update RIB, remove path to PE1

3. IGP notifies BGP that path to PE1 is now longer valid

4. PE3 identifies affected paths, runs best path, removes paths via PE1

5. Updates RIB/FIB, traffic continues
BGP Control-Plane Convergence Components III: Edge Neighbour Failure (with next-hop-self)

1. Edge link on PE1 goes down
2. eBGP session goes down
3. PE1 identifies affected paths, runs best path
4. PE1 sends withdraws to other PEs
5. PE2 & 3 run best path, update RIB/FIB, traffic continues

Withdraw: 3/8, NH <PE1>

3/8, NH 1.1.1.1
NH <PE2>

3/8, NH 1.1.1.5
NH <PE1>
BGP Control-Plane Convergence – Summary

- Two out of the three scenarios depend on IGP convergence, including IGP failure detection.
  Can react to this in much less than a second.
  Fast ReRoute techniques can even recover core failures in sub-50 milliseconds.
- BGP is notified of the change, and needs to identify affected paths by scanning the BGP table(s).
- Scanning implementation can be improved by only scanning affected table(s) or parts of tables (ex: scoped walks in IOS-XR), but scanning effort will still grow with table size.
- Hence: We need to find another approach to achieve predictable fast BGP convergence!
Control vs. Data Plane Convergence

• Control Plane Convergence
  For the topology after the failure, the optimal path is known and installed in the dataplane
  May be extremely long (table size)

• Data Plane Convergence
  Once IGP convergence has detected the failure, the packets are rerouted onto a valid path to the BGP destination
  While valid, this path may not be the most optimum one from a control plane convergence viewpoint
  We want this behaviour, in a prefix-independent way – that’s what this session is all about
BGP Data Plane Convergence
how does it work
Prefix Independent Convergence (PIC)

- PIC is the ability to restore forwarding without resorting to per prefix operations.
- Loss Of Connectivity does not increase as my network grows (one problem less to worry about)
Prefixes, path lists and paths

- Advertised with next hop 10.0.0.3/32
- Group of prefixes: 110.0.0.0/24, 110.1.0.0/24, 110.2.0.0/24, ...

Path List
- Path 1
  - Via 10.0.0.3

BGP Entry
- 110.0.0.0/24
- 110.1.0.0/24
- 110.2.0.0/24

IGB Entry
- 10.0.0.3/32
- GigE0/0

Path List
- Path 1
  - GigE0/0
Non Optimal: Flat FIB

• Each BGP FIB entry has its own local Outgoing Interface (OIF) information (ex: Gig0/0)
• Forwarding Plane must directly recurse on local OIF Information
• Original Cisco implementation and still in use both by us and competition
• FIB changes can take long, dependent on number of prefixes affected
Right Architecture: Hierarchical FIB

- Pointer Indirection between BGP and IGP entries allow for immediate update of the multipath BGP pathlist at IGP convergence
- Only the parts of FIB actually affected by a change needs to be touched
- Used in newer IOS and IOS-XR (all platforms), enables Prefix Independent Convergence
**BGP PIC Core**

- **Core Failure**: a failure within the analyzed AS
- **IGP convergence on PE3 leads to a modification** of the RIB path to PE1
  
  BGP Dataplane Convergence is finished assuming the new path to the BGP Next Hop is leveraged immediately (*BGP PIC Core*)

  BGP NHT sends a “modify” notification to BGP which may trigger BGP Control-Plane Convergence. This may be long without impacting Tight-SLA experience
Characterization
BGP PIC Core

• As soon as IGP converges, the IGP pathlist memory is updated, and hence all children BGP PL’s leverage the new path immediately
• Optimum convergence, Optimum Load-Balancing, Excellent Robustness
BGP PIC Edge

- **Edge Failure**: a failure at the edge of the analyzed AS
- IGP convergence on PE3 leads to a deletion of the RIB path to BGP Next-Hop PE1

BGP Dataplane Convergence is kicked in on PE3 (BGP PIC Edge) and immediately redirects the packets via PE2.

BGP NHT sends a “delete” notification to BGP which triggers BGP Control-Plane Convergence. This may be long, but without impacting Tight-SLA experience.
• At IGP Convergence time, the complete IGP pathlist to PE1 is deleted.
• SW FIB walks the linked list of parent BGP path lists and in-place modify them to use alternate next hops (ECMP or backup).
  BGP Path lists are shared, so this is efficient.
BGP PIC Edge and Next-Hop Self

PE1 does set next-hop-self

- With the edge device setting next-hop to its loopback (next-hop-self), edge link going down does not change next-hop as seen on other routers
  - Failure goes unnoticed by others!
- However: Next-hop on edge device with failing link goes away, so this device can react in PIC-Edge fashion
  - Traffic re-routed via core to alternate edge
Key Take-Aways

• PIC Core and PIC-Edge leverage hierarchical forwarding structure
  PIC Core: Path towards Next-Hop changes – IGP LoadInfo changed
  PIC Edge: Next-Hop goes away – BGP Path list changed

• ➔ All BGP prefixes (no matter how many!!) converge as quickly as their next-hop

• Generally, IGP is responsible for next-hop convergence ➔ BGP convergence depends on IGP convergence

• So? What do I need to do to speed up my BGP convergence with BGP PIC???
Designing for BGP PIC
Designing (for) BGP PIC

BGP PIC is a forwarding / data plane layer feature, so what’s there to design???

Well, there is a bit:

• BGP data plane convergence depends on how quickly the next-hop(s) converges (or is deleted), which boils down to
  Fast failure detection
  Fast IGP convergence

• For PIC Edge, we need some form of tunnelling/encapsulation between edge devices

• For BGP PIC-Edge, we need to have an alternative/backup next-hop
BGP PIC Design I: Fast IGP Convergence

IGP Convergence is outside the scope of this session, but in principle

1. Provide for fast failure detection (BFD, carrier-delay)
   Avoid tuning down BGP hello/hold timers for failure detection

2. For IOS, tune OSPF/ISIS LSA- and SPF-throttle timers
   Reduces convergence from ~5 down to <1 sec
   NX-OS and IOS-XR already tuned reasonably fast

3. Prioritize next-hop prefixes (i.e. PE loopback addresses) to make sure they converge before less important prefixes (i.e. link addresses or the like)

4. Keep the IGP slim and clean, use point-to-point adjacencies, carry customer routes in BGP

5. Evaluate Fast ReRoute techniques (like LFA-FRR) to further improve convergence

- Please see past Fast Convergence sessions/techtorials for details, visit us in the Design Clinics or meet using “Meet the Engineer”
BGP PIC Design II: PE – PE Encapsulation

• Some BGP-PIC Edge convergence scenarios lead to edge devices forwarding packets on to alternate edges, back via the core

• Core routers might be unaware of the failure (yet) and send packets back to the previous edge device, causing a loop

• Solution: Ensure there is no intermediate IP destination lookup, via means of encapsulation:
  Direct adjacency between edges (physical link or GRE tunnel)
  Using MPLS LSPs/Tunnels in the core
BGP PIC Design III: More than one path

• When a PE/next-hop goes away, the re-routing node needs already a backup/alternate path in its FIB

• This sounds rather obvious, but can be non-trivial in some scenarios:

• Scenario 1: Route Reflectors

• Scenario 2: Active/Standby Routing Policies

let’s start with this one first...
Scenario 2: BGP Active/Standby – The Problem (1/3)

Initial State: CE just comes up
BGP Active/Standby – The Problem (2/3)

Initial State: CE just comes up

Update: NetA, via PE1, LP: 200
NetA, via PE2, LP: 100

Update: NetA, via PE2, LP: 100

NetA, via e0, LP: 200
via PE2, LP: 100

NetA, via s0, LP: 100
via PE1, LP: 200
BGP Active/Standby – The Problem (3/3)

- PE2 withdraws its eBGP path as it is no longer best
- But now all other PEs are left with a single path – no alternate path for PIC-Edge to converge to!!!
BGP Active/Standby – The Solution: BGP Best-External

- PE2 keeps advertising his best external path
- Requires no BGP protocol change (local change only)
- Availability: 12.2SRE/15.0S/MR, XE3.1 and XR3.9

Extra question: What if PE1/PE2 acted as RR towards PE3?

router bgp ...
  address-family { vpv4 | ipv4 [vrf ...] }
  bgp advertise-best-external
Scenario 1: iBGP Path Propagation and Reflection

- Regular BGP BestPath algorithm leads to an RR only reflecting one path, namely its best path
- Without explicit policy either hot-potato policy prevails (rule 8) or the lowest neighbor address (rule 13) is used as tiebreaker
- What can we do to propagate both paths to PE3/4?
Path Diversity in L3VPN Environments

- Important sites are dual homed
- Unique RD allocation ensures both paths are learned, even through route reflectors
- For active/backup scenarios, “best-external” is required

RD1:1/8 via PE2, Label L1
RD2:1/8 via PE2, Label L2
Solution 1: Internet in a VRF

- Unique RD allocation ensures both paths are learned, even through route reflectors
- Consider per-vrf or per-CE label allocation when advertising full Internet routing table within a VRF
Solution 2: No RR

- Full iBGP mesh
- Yes, I am serious 😊, at least for reasonably sized and static environments
Solution 3: RR + partial iBGP mesh

- Done in practice by several operators
- Very specific and hence difficult to abstract any rule. Just know it exists and analyses the possibility.
Solution 4: Engineered RR

- Some operators place their RR’s in the topology to ensure they select different paths. Their PE’s are clients to multiple RR’s.
  - the top RR is closer to PE1 and selects the black path
  - the bottom RR is closer to PE2 and selects the blue path
Above behaviour can also be achieved using specific BGP policies
Solution 5: AddPath

- New Capability to allow a BGP speaker to advertise more than one path (“The holy grail”)
  
  Available in IOS-XR 4.0, IOS-XE 3.7, 15.2(4)S, 15.3T
  
  Requires support for this functionality on RR and PEs

Solution 6: BGP Diverse Paths (aka “Shadow RR”)

- New feature (IOS-XE 3.4S) allows a RR to advertise a 2nd best path
- Two deployment models:
  1. RR maintains two iBGP session to PEs (shown below)
     “Primary” connection advertises best path (PE1)
     “Secondary” connection (or secondary RR) advertises next-best-path (PE2)
  2. Dual RRs, first RR advertises best, the other RR the blue/2nd best path
- No update on PEs/RR-clients required

Designing (for) BGP PIC – Summary

- Work on the baseline – Improve your IGP convergence
  Use BFD for speedy eBGP failure detection
- Consider enabling MPLS to provide PE-PE tunnelling
- Ensure you have multiple paths – and keep this **always** in the back of your mind, whatever BGP designs/services you are coming up with

  This one is the hardest one to fix once solutions/etc. are deployed
Coming back to: BGP Active/Standby – BGP Best-External

Extra question: What if PE1/PE2 acted as RR towards PE3? Let’s look at PE2

- A BGP RR always needs to reflect the best path to its clients (PE3) – this is the one via PE1!!
- We need to enable AddPath to advertise the best-external in addition to the best one
Configuring BGP PIC
Enabling BGP PIC – Enabling IP Routing Fast Convergence

• BGP PIC leverages IGP convergence ➞ Make sure IGP converges quickly

• IOS-XR: IGP Timers pretty-much tuned by default

• IOS: Sample OSPF config:

```plaintext
process-max-time 50
ip routing protocol purge interface

interface ...
  carrier-delay msec 0
  negotiation auto
  ip ospf network point-to-point
  bfd interval 100 min_rx 100 mul 3
router ospf 1
  timers throttle spf 50 100 5000
  timers throttle lsa all 0 20 1000
  timers lsa arrival 20
  timers pacing flood 15
  passive-interface Loopback 0
  bfd all-interfaces
```
Enabling BGP PIC Core

• PIC Core is a pure data plane feature, leveraging a hierarchical forwarding/CEF plane
• On IOS devices (C7600, ASR1k), hierarchy needs to be enabled via:

```plaintext
cef table output-chain build favor convergence-speed
```

• All other platforms supporting PIC core (CRS, XR12k ASR9k, NX-OS) do this natively, there is nothing to configure here
Enabling BGP PIC Edge: IOS

• Two BGP-PIC Edge Flavors: 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

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

... or implicitly when enabling best external
```

```
router bgp ...
  address-family ...
  bgp advertise-best-external
```
Enabling BGP PIC Edge: IOS-XR

- As in IOS, PIC-Edge in XR w/ multipath requires no additional configuration
- PIC-Edge unipath needs to be enabled explicitly:

```plaintext
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
  !
```
Monitoring BGP-PIC: IOS

PE3#sh bgp vpnv4 unicast vrf red
BGP table version is 43, local router ID is 10.0.0.3
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
    r RIB-failure, S Stale, m multipath, b backup-path, x best-external, f RT-Filter, a additional-path
Origin codes: i - IGP, e - EGP, ? - incomplete

<table>
<thead>
<tr>
<th>Network</th>
<th>Next Hop</th>
<th>Metric</th>
<th>LocPrf</th>
<th>Weight</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>*&gt; 100.0.0.10/32</td>
<td>100.1.10.10</td>
<td>0</td>
<td>0</td>
<td>100 i</td>
<td></td>
</tr>
<tr>
<td>*bi 110.0.0.11/32</td>
<td>10.0.0.2</td>
<td>0</td>
<td>100</td>
<td>0 110 i</td>
<td></td>
</tr>
<tr>
<td>*&gt;i</td>
<td>10.0.0.1</td>
<td>0</td>
<td>2000</td>
<td>0 110 I</td>
<td></td>
</tr>
</tbody>
</table>

PE3#

Backup/Repair Path

Use “show ip route [vrf ..] repair-paths …” to view backup paths in RIB..
Monitoring BGP-PIC: IOS

PE3#sh bgp vpnv4 unicast vrf red 110.0.0.11
BGP routing table entry for 1:3:110.0.0.11/32, version 43
Paths: (2 available, best #2, table red)
  Additional-path-install
  Advertised to update-groups:
    1
  Refresh Epoch 1
110, imported path from 1:1:110.0.0.11/32
  10.0.0.2 (metric 21) from 10.0.0.6 (10.0.0.6)
    Origin IGP, metric 0, localpref 100, valid, internal, backup/repair
    Extended Community: RT:1:100
    Originator: 10.0.0.2, Cluster list: 10.0.0.6, recursive-via-host
    mpls labels in/out nolabel/16
  Refresh Epoch 1
110, imported path from 1:2:110.0.0.11/32
  10.0.0.1 (metric 21) from 10.0.0.6 (10.0.0.6)
    Origin IGP, metric 0, localpref 2000, valid, internal, best
    Extended Community: RT:1:100
    Originator: 10.0.0.1, Cluster list: 10.0.0.6, recursive-via-host
    mpls labels in/out nolabel/17

PE3#
Monitoring BGP-PIC: IOS

PE3#sh ip cef vrf red 110.0.0.11/32 internal | i list|lock|adj|buck|choic|chain
contains path extension list
path DD944670, path list DD60E02C, share 1/1, type recursive, for IPv4, flags must-be-labelled, recursive-via-host
path DD9449F0, path list DD60E2AC, share 1/1, type attached nexthop, for IPv4 nexthop 10.1.2.2 Ethernet0/1 label 17, adjacency IP adj out of 10.1.2.2 DE6DF558
path DD944A60, path list DD60E2AC, share 1/1, type attached nexthop, for IPv4 nexthop 10.1.5.5 Ethernet0/2 label 18, adjacency IP adj out of Ethernet0/2, addr 10.1.5.5 DE6DF6B8
path DD9446E0, path list DD60E02C, share 1/1, type recursive, for IPv4, flags must-be-labelled, repair, recursive-via-host
path DD944AD0, path list DD60E2FC, share 1/1, type attached nexthop, for IPv4 nexthop 10.1.5.5 Ethernet0/2 label 19, adjacency IP adj out of 10.1.5.5 DE6DF6B8
...
PE3#

BGP Path List
Primary Entry

IGP Path List

BGP Path List
Backup (repair) Entry

IGP Path List
(to backup next-hop)
Monitoring BGP-PIC: IOS-XR

```
RP/0/5/CPU0:C3#show bgp vrf VPN_1 10.1.0.0/22

BGP routing table entry for 10.1.0.0/22, Route Distinguisher: 100:100
...
Paths: (2 available, best #1)
   Not advertised to any peer
   Path #1: Received by speaker 0
      Not advertised to any peer
      101
      10.1.1.1 (metric 120) from 172.16.1.7 (10.1.1.1)
         Received Label 66
         Origin EGP, localpref 100, valid, internal, best, group-best, import-candidate, imported
            Received Path ID 0, Local Path ID 1, version 28702
            Extended community: RT:100:1
            Originator: 10.1.1.1, Cluster list: 172.16.1.7
      Path #2: Received by speaker 0
      Not advertised to any peer
      101
      10.2.1.1 (metric 145) from 172.16.1.7 (10.2.1.1)
         Received Label 66
         Origin EGP, localpref 100, valid, internal, backup, add-path, import-candidate, imported
            Received Path ID 0, Local Path ID 2, version 155502
            Extended community: RT:100:1
            Originator: 1.2.1.1, Cluster list: 172.16.1.7
```
Monitoring BGP-PIC: IOS-XR

RP/0/5/CPU0:C3#show route vrf VPN_1 10.1.0.0/22

Routing entry for 10.1.0.0/22
  Known via "bgp 100", distance 200, metric 0
  Tag 101
  Number of pic paths 1, type internal
  Installed Jan 25 03:21:36.026 for 1d19h
  Routing Descriptor Blocks
    10.1.1.1, from 172.16.1.7
      Nexthop in Vrf: "default", Table: "default", IPv4 Unicast, Table Id: 0xe0000000
      Route metric is 0
    10.2.1.1, from 172.16.1.7, BGP backup path
      Nexthop in Vrf: "default", Table: "default", IPv4 Unicast, Table Id: 0xe0000000
      Route metric is 0
      No advertising protos.
RP/0/5/CPU0:C3#

Backup/Repair Path shows up in routing table as well!
RP/0/5/CPU0:C3#show cef vrf VPN_1 10.1.0.0/22

10.1.0.0/22, version 4, internal 0x40040001 (ptr 0x9e1923b8) [1], 0x0 (0x0), 0x4100 (0x9ed2b0d4)
Updated Jan 25 03:21:36.793
Prefix len 22, traffic index 0, precedence routine (0)
  via 10.1.1.1, 3 dependencies, recursive [flags 0x10]
    path-idx 0
    next hop VRF - 'default', table - 0xe0000000
    next hop 10.1.1.1 via 16393/0/21
    next hop 10.10.13.1/32 PO0/0/1/0 labels imposed {16393 66}
  via 10.2.1.1, 2 dependencies, recursive, backup [flags 0x110]
    path-idx 1
    next hop VRF - 'default', table - 0xe0000000
    next hop 10.2.1.1 via 16032/0/21
    next hop 10.12.16.2/32 PO0/0/0/0 labels imposed {16032 66}
RP/0/5/CPU0:C3#
Advanced BGP PIC Topics
FIB Recursion and Summaries/Default Routes

- Default behaviour is for the FIB to recurse through the longest match prefix
- This could include a summary or a default route, which could be a valid path to the destination
- However: In most designs, a /32 next-hop going away ➔ next-hop router is gone for good, there is no point to find a longer match
- Similar problem happens on the control plane when tracking the BGP next-hop

BGP Net 172.16.0.0/16

Next Hops:
- 10.0.0.3
- b: 10.0.0.4

Hmm, should I resolve my primary 10.0.0.3 via the /24 summary, or do I need to invoke the backup????

IP Routing Table
- 10.0.0.3/32
- 10.0.0.4/32
- 10.0.0.0/24
- 0.0.0.0/0
- ...

10.0.0.3/32
10.0.0.4/32
10.0.0.0/24
0.0.0.0/0
FIB Recursion and Summaries/Default Routes

IOS doesn’t recurse on summaries/default …

- when PIC Edge is enabled or explicitly via:
- when next-hop is directly connected (eBGP)
  - recurse-through-connected flag

IOS-XR

- No issue for labeled paths (i.e. vpnv4/6/rfc3107 next-hops), always only recursed via /32s
- New in 4.2: For unlabelled paths, including VRF prefixes: Does not recurse via default route or via BGP summaries, or when candidate prefix length is less than configured:

```
routing bgp ...
  address-family ...
  bgp recursion host
```

```
routing bgp ...
  address-family ...
  next-hop resolution prefix-length minimum {32|128}
```

```
BGP Label Allocation Modes

- BGP Layer3 VPN generally support multiple MPLS Label Allocation modes on the PE
  - One label **per prefix** (default)
  - One label **per CE**
  - One label for all routes **per VRF**
    Requires aggregate lookup

- The label allocation mode has an impact to BGP PIC…
BGP Label Allocation Modes and BGP PIC: Per-Prefix

- Default label allocation mode
- Normal label switched path to primary PE
BGP Label Allocation Modes and BGP PIC: Per-Prefix

- Default label allocation mode
- Following PE-CE failure, primary PE1 re-routes traffic to standby PE2 (PIC-Edge) while control plane converges

**FIB:**
- p1 via CE1, local label 111
  - backup: PE2, L=222

**LFIB:**
- L 111: untagged via <ce1>
  - backup: swap with <pe2>, 222
- L 222: untagged via <ce1>

**Diagram:**
- PE1 (active) connects to CE1 via L1 111 and PE2 (standby) via L2 222.
- PE1 re-routes p1 via PE1, L=111 backup: CE1, local label 222.
- PE2 standby uses CE1, local label 222 for backup.
BGP Label Allocation Modes and BGP PIC: Per-VRF

- Per-VRF requires additional IP lookup to find the final destination
- This creates a transient loop in an active/standby environment with “best-external” ...

**Diagram:***

- PE1 allocates per-vrf label 50
- PE1 allocates per-vrf label 30
- L 50: strip label, L3 lookup in <vrf>
  - backup: swap with <pe2>,30
- L 30: strip label, L3 lookup in <vrf>
  - backup: via CE1
- p1 via PE1, L=50
  - backup: via CE1
  - backup: via PE1, L=30
- p1 via CE1, backup: PE2, L=30
  - backup: PE2, L=30
BGP Label Allocation Modes and BGP PIC: Per-VRF

- Per-VRF requires additional IP lookup to find the final destination.
- This creates a transient loop in an active/standby environment with “best-external”
  Following PE-CE link failure until PE2 has converged via BGP control plane convergence.

Diagram:
- P1 via CE1
  - backup: PE2, L=30
- P2 via CE1
  - backup: PE2, L=30

PE1
- (active)
- L 50: strip label, L3 lookup in <vrf>
- backup: swap with <pe2>,30

Loop!
- P1 via PE1, L=50
  - backup: via CE1
- P2 via PE1, L=50
  - backup: via CE1

CE1
- L 30: strip label, L3 lookup in <vrf>
- PE2
- (standby)
BGP Label Allocation Modes and BGP PIC: Per-VRF

- Per-VRF requires **additional IP lookup** to find the final destination
- This creates a transient loop in an active/standby environment with “best-external”

Following PE-CE link failure until PE2 has converged via BGP control plane convergence
BGP Label Allocation Modes and BGP PIC: Per-VRF

- Per-VRF requires additional IP lookup to find the final destination.
- This creates a transient loop in an active/standby environment with “best-external”
  Following PE-CE link failure until PE2 has converged via BGP control plane convergence.

```
L 30: strip label, L3 lookup in <vrf>
```
BGP Label Allocation Modes and BGP PIC: Per-CE

- Per-CE does not require a 2nd lookup, so no issues with best-external, but:
- BGP PIC with Per-CE label allocation would require that
  a) All PEs hosting redundant CE connections are configured with per-CE
  b) All prefixes are advertised over all redundant connections
- Example: In below situation: which label (123 or 456) should PE1 swap for label 50 when PE1-CE1 link goes down? PE1 can’t be sure!
BGP Label Allocation Modes and BGP PIC: Per-Prefix

- Per-prefix has a 1:1 correspondence between label and prefix, so no ambiguities when it comes to swapping backup labels.
- Also, there are also no issues with aggregate lookups, so no problem with transient loops.
- Full BGP PIC Edge Support!!
# BGP Label Allocation Modes: Summary

<table>
<thead>
<tr>
<th></th>
<th>Per-Prefix</th>
<th>Per-CE</th>
<th>Per-VRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGP PIC Edge</td>
<td>Fully supported</td>
<td>Not supported*)</td>
<td>Possible transient loop</td>
</tr>
<tr>
<td>BGP Best-External</td>
<td>Yes</td>
<td>Yes</td>
<td>Possible transient loop</td>
</tr>
<tr>
<td>eiBGP Multipath</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ECMP PE-CE</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

*) Future: “Resilient Per-CE Label” compatible with PIC-Edge
Inter-AS MPLS-VPN Deployments

• Essentially three different methods to inter-connect two MPLS-VPN networks:
  
  ▪ Option A: Individual (sub)interfaces per VPN, each SP treats the other as “CE”
    
      ✓ No special PIC considerations

  ▪ Option B: One interface for all VPNs, direct eBGP sessions exchanging vpnv4/6 routes
    
      One problem, see next slide

  ▪ Option C: One interface, eBGP session exchanging vpnv4/6 routes between RRs
    
    next-hop information (remote PEs) exchanged via direct eBGP or via IGP+LDP
    
    ✓ No special PIC considerations
Inter-AS Option B and BGP PIC Edge

- Using unique RDs on PEs ensures path diversity by making the two updates distinguishable while they are sent through the vpnv4 iBGP “cloud”
- But this also prevents the ASBRs in the local AS to make a connection between these updates for PE backup purposes
  - ASBR1 doesn’t know that it could use rd2:p1 to back up PE1
- We’re working on a solution to this problem
  - Common RD and “Add-Path” could be used

Hmm, how can I protect CE1 against PE1 failure?
Wrapping Up
Summary

• Routing Fast Convergence essential to deliver today’s critical applications
• IGP conservatively meet sub-second convergence requirements
  Timer tuning required in IOS, IOS-XR and NX-OS optimized by default
  IGP needs to be kept slim
• BGP PIC enables BGP to achieve the same level of convergence, no matter how many prefixes we deal with
• Availability of alternate path needs to be kept in mind whenever designing BGP services
BGP PIC Implementation – IOS (7600, ASR1k), NX-OS

• PIC-Core
  7600
  12.2(33)SRB: IPv4, non-ECMP
  12.2(33)SRC: IPv4, non-ECMP + ECMP / vpnv4, non-ECMP
  15.0(1)S: IPv4+vpnv4, non-ECMP and ECMP

ASR1k: XE2.5.0
NX-OS: 5.2

• PIC-Edge
  7600, 7200: 12.2(33)SRE
  ASR1k: XE3.2.0 (v4), 3.3.0(v6)
  NX-OS: Radar
BGP PIC Implementation – IOS-XR

• BGP PIC Core
  3.4 CRS, 3.3 12k, 3.7 ASR9k
• BGP PIC Edge
  Multipath: 3.5
  Unipath: 3.9

• Development continues – several PIC Core & Edge enhancements have recently been released in 4.2
  Like 3107 + Label, additional PIC triggers and many more
Practice BGP PIC in LABRST-2007

Visit the Labs in Word of Solutions to get Hands On with BGP PIC
- BGP Prefix Independent Convergence (PIC)
- BGP Best-External
- BGP Diverse Path
Related Sessions

At CiscoLive London 2013

• BRKSPG-2402 - Best Practices to Deploy High-Availability in Service Provider Edge and Aggregation Architectures
• BRKRST-2333 - Network Failure Detection

Previous CiscoLive (Check CiscoLive365.com)

• BRKRST-3363 - Routed Fast Convergence and High Availability
• BRKIPM-2000 - Routing Resiliency - Latest Enhancements
• TECRST-3190 - Fast Convergence Techtorial
• ...
Call to Action

• **Visit** the Cisco Campus at the World of Solutions to experience Cisco innovations in action

• **Get** hands-on experience attending one of the Walk-in Labs

• **Schedule** face to face meeting with one of Cisco’s engineers at the Meet the Engineer center

• **Discuss** your project’s challenges at the Technical Solutions Clinics
Your feedback is important to us.

Complete the session survey at: www.ciscolivelondon.com/onsite or via the Cisco Live Mobile App