BGP Prefix Independent Convergence

draft-bashandy-rtgwg-bgp-pic-01

Authors:
Ahmed Bashandy, Cisco Systems
Clarence Filsfils, Cisco Systems
Pradosh Mohapatra, Sproute Networks

Presenter:
Ahmed Bashandy

IETF94, Nov/2015
Yokohama, Japan
Agenda

• Problem
• Solution
• Handling Limited Hardware Capability
• Recovery from various failure scenarios
Agenda

- Problem
- Solution
- Handling Limited Hardware Capability
- Recovery from various failure scenarios
What do we want to Achieve?

• *Serial* nature of reachability propagation
  ⇒ BGP convergence is *inherently* slow

• Can we adjust the data structure in time complexity that does *not* depend on BGP prefixes?
  – If a path is lost or gained, modifications should be independent of the number of prefixes
  – If there is a recovery, recover as soon as possible independent of the number of prefixes

**Objective**

Make re-convergence after topology change independent of the number of BGP prefixes
Terminology

- **Leaf**: A prefix or local label container datastructure
- **Path**: a recursive or non-recursive path
  - May be primary or backup
- **Pathlist**: an array of paths
  - Each path carries its own pathindex
- **OutLabel-Array**: Array of outgoing labels and/or label actions*
  - A possibly different outlabel-array attached to each leaf
  - Each entry represents an outgoing label and/or label action for a path in the pathlist**
- **Adjacency**:
  - The L2 information to send a packet to a directly connected router
Agenda

• Problem
• Solution
• Handling Limited Hardware Capability
• Recovery from various failure scenarios
Basic Idea

• BGP-PIC is really a **FIB** feature
• BGP-PIC need not involve BGP at all: e.g. works for pseudo-wires
• *Hierarchical* and *shared* forwarding chains
• On topology changes
  - FIB modifies pathlists immediately without modifying leaves
  - Restore traffic very quickly
  - BGP modifies leaves later at a slower pace
• Behavior is internal to the router
  - Completely *transparent* to operator
  - *Incrementally* deployable
Sample Topology: VPN in SR/LDP Core

SR/LDP Core

PE1

PE2

PE3

Two IGP paths

Two BGP paths*

N/n

M/m
Forwarding Chain on *Ingress* PE

**PE2**

- **VPN-L21**
- **VPN-L31**

- **PE2** IGP leaf

**PE3**

- **VPN-L22**
- **VPN-L32**

- **PE3** IGP leaf

**Shared BGP Pathlist** (even among VRFs)

**Shared IGP Pathlist**

**LDP-L11**

- **LDP-L12**
- **LDP-L21**
- **LDP-L22**

**IGP OutLabel Array**

**BGP OutLabel Array**

SR/LDP Core

PE1

PE2

PE3

N/n

M/m

Adj1

Adj2
Agenda

• Problem
• Solution
• Handling Limited Hardware capability
• Recovery from various failure scenarios
Flattening a Forwarding Chain: Sample Topology

- Inter-AS Option C, but
  - ABRs advertise remote domain core using **BGP** instead of **IGP**

![Sample Topology Diagram](image-url)

- Advertise PE2x Using **iBGP-LU**
- Advertise PE2x using **eBGP-LU**
- Redistribute IGP into BGP
Forwarding Chain on PE11

OutLabel Array to reach ASBRs (advertised by iBGP)

VPN OutLabel Array

VPN-PE21(N/n)
VPN-PE22(N/n)

N/n

M/m

VPN-PE22(M/m)
VPN-PE32(M/m)

PE11

ASBR11
ASBR12
ASBR13

ASBR21
ASBR22
ASBR23

PE21

PE22

PE23

ASBR11
ASBR12
ASBR13

LASBR11(PE21)
LASBR12(PE21)

LASBR11(PE22)
LASBR12(PE22)

LASBR13(PE23)

LDP-L11
LDP-L12
LDP-L13

Adj1
Adj2
Adj3
Flattened* Forwarding Chain on PE11

VPN OutLabel Array

VPN-PE21(N/n)  VPN-PE22(N/n)

N/n

M/m

VPN-PE22(M/m)  VPN-PE32(M/m)

LASBR11(PE21)  LASBR12(PE21)

LASBR11(PE22)  LASBR12(PE22)

LASBR11(PE23)  LASBR12(PE23)

LDP-L11  LDP-L12  LDP-L13

Adj1  Adj2  Adj3

ASBR11  ASBR12  ASBR13

ASBR11**  ASBR12**  ASBR13

LASBR11(PE21)  LASBR12(PE21)

LASBR11(PE22)  LASBR12(PE22)

LASBR11(PE23)  LASBR12(PE23)
Agenda

- Problem
- Solution
- Handling Limited Hardware Capability
- Recovery from various failure scenarios
Core Link/Node Failure

Works whether BGP is running In the core or not

A

Failure

B

IGP/LDP Convergence

C
Core Link/Node Failure: FIB on PE1

Core Link Failure

IGP/LDP Convergence

BGP pathlist **NOT** modified

Only IGP pathlist is modified
Unipath PE Node Failure: Protection at *Ingress PE*

Works whether BGP is running in the core or not*

PE3 must be pre-programmed as a backup path

---

Core or Edge node Failure

FIB Correction

Works whether BGP is running in the core or not*

PE3 must be pre-programmed as a backup path

---

*This text is a placeholder and needs to be replaced with actual content.*
Unipath PE Node Failure: Protection at Ingress PE

Failure

Backup PE

BGP Pathlist recalculation

Outlabel-array does NOT change

Path keeps its original index

Use backup PE and backwalk to recalculate higher level LDIs

PE3 must be pre-programmed as a backup path

IGP withdraws

The leaf for PE2

Unipath PE Node Failure: Protection at Ingress PE

Failure

Backup PE

BGP Pathlist recalculation

Outlabel-array does NOT change

Path keeps its original index

Use backup PE and backwalk to recalculate higher level LDIs

PE3 must be pre-programmed as a backup path

IGP withdraws

The leaf for PE2
Unopath PE-CE Link Failure: Protection at *Egress PE*

More useful for BGP-free core

PE3 must be *pre-programmed* as a backup path

Failure

FIB Correction

More useful for BGP-free core

PE3 must be *pre-programmed* as a backup path

Failure

FIB Correction
Unipath PE-CE Link failure: FIB on PE2

**Failure**

- **Locally** detect failure
- 50ms recovery!!

**Backup path** Going to PE3

**BGP Pathlist recalculation**

- Swap incoming PE2 VPN label with PE3 VPN label then push transport label for PE3
- **Outlabel-array does **NOT** change**

- Path keeps its **original index**

**PE3 must be pre-programmed as a backup path**
Effect of Flattening: ASBR12 fails

• Suppose ASBR12 is no longer reachable in domain 1
Unflattened Chain After Failure

ASBR11

PE21

LASBR11(PE21)

LASBR12(PE21)

PE22

LASBR11(PE22)

LASBR12(PE22)

PE23

LASBR13(PE23)

ASBR13

LDP-L11

Adj1

LDP-L13

Adj3

VPN OutLabel Array

VPN-PE1(N/n)

VPN-PE21(N/n)

VPN-PE22(N/n)

VPN-PE23(M/m)

VPN-PE32(M/m)

N/n

M/m
Flattened Chain After Failure

VPN OutLabel Array

VPN-PE21(N/n)
VPN-PE22(N/n)
VPN-PE22(M/m)
VPN-PE32(M/m)

ASBR11
ASBR13

LASBR11(PE21)
LASBR12(PE21)
LASBR11(PE22)
LASBR12(PE22)
LASBR13(PE23)

ASBR11
ASBR12

LDP-L11
LDP-L12
LDP-L13

Adj1
Adj2
Adj3

ASBR11
ASBR13

VPN OutLabel Array

VPN-PE21(N/n)
VPN-PE22(N/n)
VPN-PE22(M/m)
VPN-PE32(M/m)
Conclusion

• Single elegant design to handle many convergence/protection cases
• BGP prefix (as well as other) independent convergence
• Support for any hierarchy depth with graceful PIC degradation for shallower hierarchy
Request to become a Working Group Document

!!

😊😊😊😊😊😊😊😊😊