Fast Convergence Techniques

SANOG 25, Kandy, Sri Lanka
January 16, 2015

Sumon Ahmed Sabir
sumon@fiberathome.net

Md. Abdullah-Al-Mamun
mamun@fiberathome.net

Need for Fast Convergence

Its not only browsing, mail and watching videos any more.

Internet and Networks carrying Voice/Video calls.

Carrying business and mission critical data.

No option for outage or interruption.

Need for Fast Convergence

Few years before in Ethernet network Convergence time was about 2 minutes.

At present it takes few seconds without any fast convergence techniques applied in Interface and protocol configuration.

But many critical services demand < 50ms convergence time in a carrier grade network.

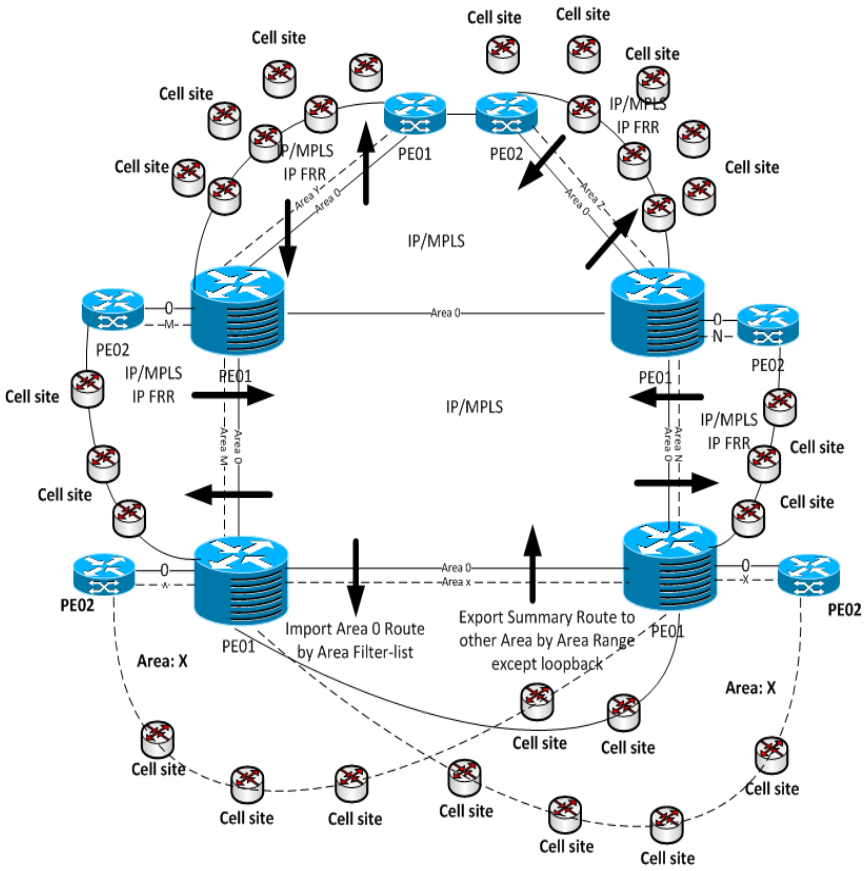
Design Consideration

- Network Topology
- IP Planning
- IGP Fine Tuning
- Type of Service Delivery
- Service Takeover and handover Point

SANOG

Better IP Plan Better Convergence

- Domain/Area Based IP Plan must be taking place to minimize the prefixes
- Prefix Summery or Area summery is very effective to aggregate individual small prefixes within the Area



IGP Fast Convergence

- Failure Detection
- Event Propagation
- SPF Run
- RIB FIB Update

- Time to detect the network failure, e.g. interface down condition.
- Time to propagate the event, i.e. flood the LSA across the topology.
- Time to perform SPF calculations on all routers upon reception of the new information.
- Time to update the forwarding tables for all routers in the area.

Purging the RIB on link failure

- Routing protocols are more efficient than RIB process in detecting link failure to delete the associate next-hop routes of the failed interface. Enabling this feature reduces convergence time significantly specially in case of a large routing table.

```
ip routing protocol purge interface
```

Link Failure Detection Process

Here is few methods to detect the link failure

1. IGP keepalive times/ fast hellos with the dead/hold interval of one second and sub-second hello intervals. It is CPU hungry
2. carrier-delay msec 0, Physical Layer
3. BFD, Open Standard more reliable rather than IGP Keepalive fast hello

Link Failure Detection

- Set Carrier-delay to 0 ms to change the link state instantly. If you are using any other transport services like SDH or DWDM set the value according to your transport network

```
int gi0/0/1
    carrier-delay msec 0
```

Link Failure Detection

- Enable BFD to notify routing protocols about the link failure in sub second interval. Without BFD it will take at least 1 second

```
int gi0/0/1
ip ospf bfd
bfd interval 50 min_rx 50 multiplier 3
```

SANOG

Link Failure Detection

- In Ethernet interface, ISIS/OSPF will attempt to elect a DIS/DR when it forms an adjacency
  - As it is running as a point-to-point link, configuring ISIS/OSPF to operate in "point-to-point mode" reduces link failure detection time

```
int gi0/0/1
     isis network point-to-point


int gi0/0/1
     ip ospf network point-to-point
```

SPF Calculation

- The use of Incremental SPF (iSPF) allows to further minimize the amount of calculations needed when partial changes occur in the network
- Need to enable ispf under ospf/isis process

```
router ospf 10
  ispf
```

Set Overload bit

- Wait until iBGP is running before providing transit path

```
router isis isp
    set-overload-bit on-startup wait-for-
bgp

router ospf 10
    max-metric router-lsa on-startup wait-
for-bgp
```

- Avoids blackholing traffic on router restart
- Causes OSPF/ISIS to announce its prefixes with highest possible metric until iBGP is up and running

Non Stop Forwarding

- Cisco NSF with SSO or Juniper Non Stop Active Routing for systems with dual route processor allows a router that has experienced a hardware of software failure of an active route processor to maintain data link layer connections and to continue forwarding packets during the switchover to the standby route processor

Event Propagation

After Link Down Event	Remarks	Command
LSA generation delay	timers throttle lsa initial hold max_wait	timers throttle lsa 0 20 1000
LSA reception delay	This delay is a sum of the ingress queuing delay and LSA arrival delay	timers pacing retransmission 100
Processing Delay	timers pacing flood (ms) with the default value of 55ms	timers pacing flood 15
Packet Propagation Delay	12usec for 1500 bytes packet over a 1Gbps link	N/A

RIB/FIB Update

Lesser Number of Prefixes lesser time to converge the RIB and FIB

RIB/FIB Update

- After completing SPF computation, OSPF/ISIS performs sequential RIB update to reflect the changed topology. The RIB updates are further propagated to the FIB table
- The RIB/FIB update process may contribute the most to the convergence time in the topologies with large amount of prefixes, e.g. thousands or tens of thousands
- Platform what you are using, higher capacity CPU and RAM will cater better performance.

Configuration Template

```
router ospf 10
  max-metric router-lsa on-startup
  wait-for-bgp
  timers lsa arrival 50
  timers throttle lsa all 10 100 1000
  timers throttle spf 10 100 1000
  timers pacing flood 5
  timers pacing retransmission 60
  ispf
  bfd all interfaces
```

SANOG

Configuration Template

```
router isis ISP
  set-overload-bit on-startup wait-for-
  bgp
  spf-interval 5 1 20
  lsp-gen-interval 5 1 20
  prc-interval 5 1 20
  fast-flood 10
  bfd all-interfaces
  ispf level-1-2 60
```

SANOG

Final Calculation

Event	Time(ms)	Remarks
Failure Detection Delay: Carrier-delay msec 0	0	about 5-10ms worst case to detect
In BFD Case	150	Multiplayer 3 is last count: 50ms interval
Maximum SPF runtime	64	doubling for safety makes it 64ms
Maximum RIB update	20	doubling for safety makes it 20ms
OSPF interface flood pacing timer	5	does not apply to the initial LSA flooded
LSA Generation Initial Delay	10	enough to detect multiple link failures resulting from SRLG failure
SPF Initial Delay	10	enough to hold SPF to allow two consecutive LSAs to be flooded
Network geographical size/Physical Media(Fiber)	0	signal propagation is negligible

Final FIB UPDATE Time: Maximum 500ms. It is sub-second convergence

SANOG

Beyond Sub second Convergence

But if you need < 50 ms Convergence time, Need to do more…….

i. RSVP Based link/node protection route
ii. LDP Based LFA-FRR

50-ms Convergence: Do we really need this?

- Most of the applications and services we are using today are fine with sub second(500ms) convergence.
- Few applications like stock trading, mobile phone recharge, few other poorly written apps people using asks for 50ms convergence.
- L2Circuit emulation over IP some times breaks over 100ms
- `http://www.ethernetacademy.net/Ethernet-Academy-Articles/putting-50-milliseconds-in-perspective`

LFA-FRR

- Provide local sub-100ms convergence times and complement any other fast convergence tuning techniques that have been employed
- LFA-FRR is easily configured on a router by a single command, calculates everything automatically
- Easy and lesser complex than RSVP Based Traffic Engineering.

Prerequisite

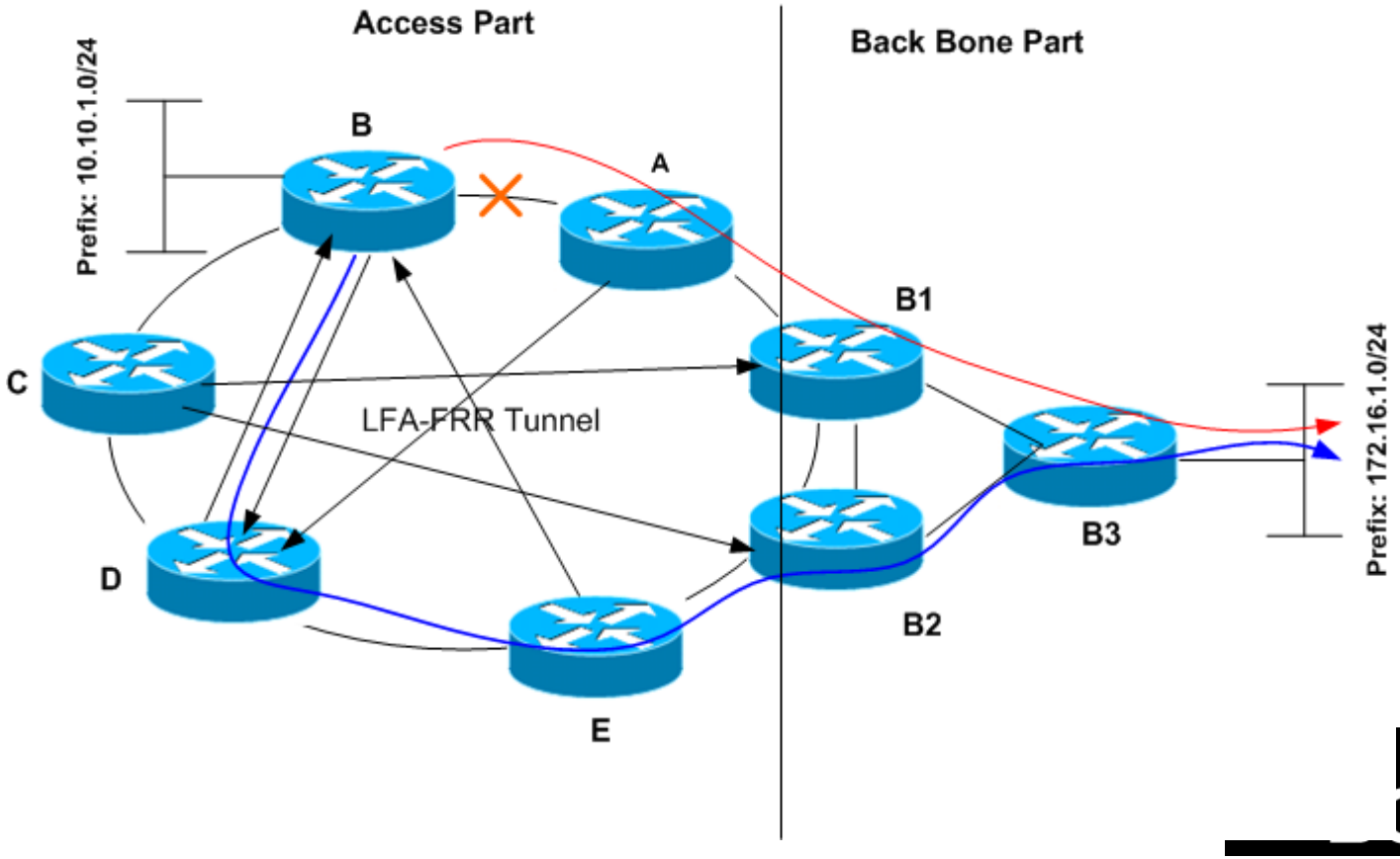
- Need MPLS LDP Configuration
- Need BFD Configuration to trigger Fast Reroute
- Need some Fast Reroute configuration under OSPF Process
- Need some special configuration based on platform

```
mpls ldp discovery targeted-hello
   accept
router ospf Y
router-id xxxxx
ispf
prefix-priority high route-map
   TE_PREFIX
 fast-reroute per-prefix enable
   area y prefix-priority high
 fast-reroute per-prefix remote-
   lfa tunnel mpls-ldp

ip prefix-list TE_PREFIX seq 5
   permit a.b.c.d/32
!
route-map TE_PREFIX permit 10
 match ip address prefix-list
   TE_PREFIX
```

How it works

1. **Initially best path for the prefix 172.16.1.0/24 is B-A-B1-B3**
2. **Once the link fails between B-A then prior computed LFA Tunnel Triggered by BFD**
3. **Immediate Target Prefix(es) are passed through B-D LFA Tunnel**
4. **Pack drop does not observe because B router does not wait for IGP convergence**



LFA-FRR Design Consideration

- In a Ring Topology
- Lesser Prefix make quicker convergence
- Specific Prefix with higher priority will show best performance without any service interruption and packet drop.

```
show ip cef 10.255.255.29
10.255.255.29/32
  nexthop 10.10.202.65 Vlan10 label [166|1209]
    repair: attached-nexthop 10.253.51.94 MPLS-Remote-Lfa124
```

```
ROBI39-DHKTL25#sh ip int brief
Loopback1              10.253.51.91            YES NVRAM    up       up
MPLS-Remote-Lfa124       10.10.202.69     YES unset up       up
```

Before/After LFA FRR

```
Xshell:\> ping 10.252.51.111 -t
Reply from 10.252.51.111: bytes=32 time=2ms TTL=253
Reply from 10.252.51.111: bytes=32 time=4ms TTL=253
Reply from 10.252.51.111: bytes=32 time=2ms TTL=253
Reply from 10.252.51.111: bytes=32 time=2ms TTL=253
Request timed out.
Reply from 10.252.51.111: bytes=32 time=61ms TTL=253
Reply from 10.252.51.111: bytes=32 time=86ms TTL=253
Reply from 10.252.51.111: bytes=32 time=70ms TTL=253
Reply from 10.252.51.111: bytes=32 time=147ms TTL=253
```

```
Reply from 10.252.51.111: bytes=32 time=2ms TTL=253
Reply from 10.252.51.111: bytes=32 time=2ms TTL=253
Reply from 10.252.51.111: bytes=32 time=1ms TTL=253
Reply from 10.252.51.111: bytes=32 time=1ms TTL=253
Reply from 10.252.51.111: bytes=32 time=27ms TTL=253
Reply from 10.252.51.111: bytes=32 time=32ms TTL=253
Reply from 10.252.51.111: bytes=32 time=1ms TTL=253
Reply from 10.252.51.111: bytes=32 time=2ms TTL=253
Reply from 10.252.51.111: bytes=32 time=2ms TTL=253
Reply from 10.252.51.111: bytes=32 time=1ms TTL=253
```

SANOG

BGP Fast Convergence

LFA-FRR or RSVP can improve L2-VPN and Intra-AS Convergence but can’t do much for External prefixes learn via EBGP

BGP Fast Convergence

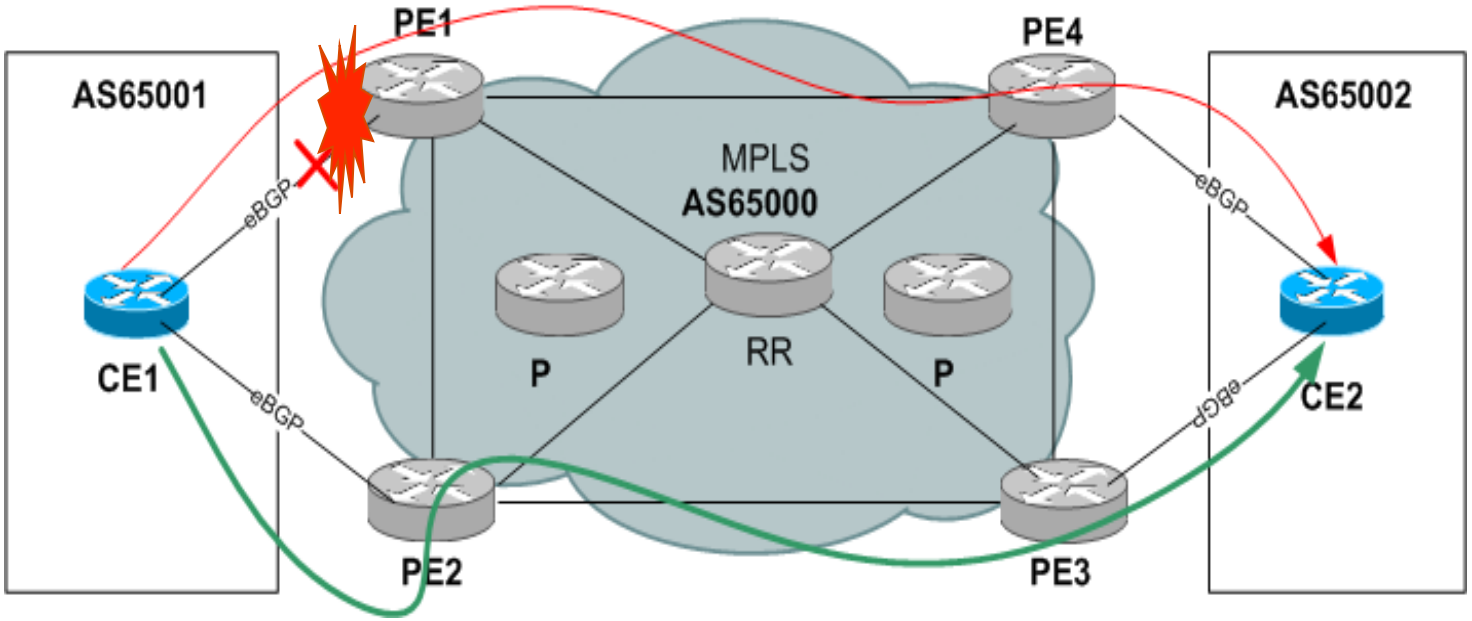
The BGP PIC Edge for IP and MPLS-VPN feature improves BGP convergence once a network failure.

SANOG

Prerequisites

- BGP and the IP or Multiprotocol Label Switching (MPLS) network is up and running with the customer site connected to the provider site by more than one path (multihomed).
- Ensure that the backup/alternate path has a unique next hop that is not the same as the next hop of the best path.
- Enable the Bidirectional Forwarding Detection (BFD) protocol to quickly detect link failures of directly connected neighbors.

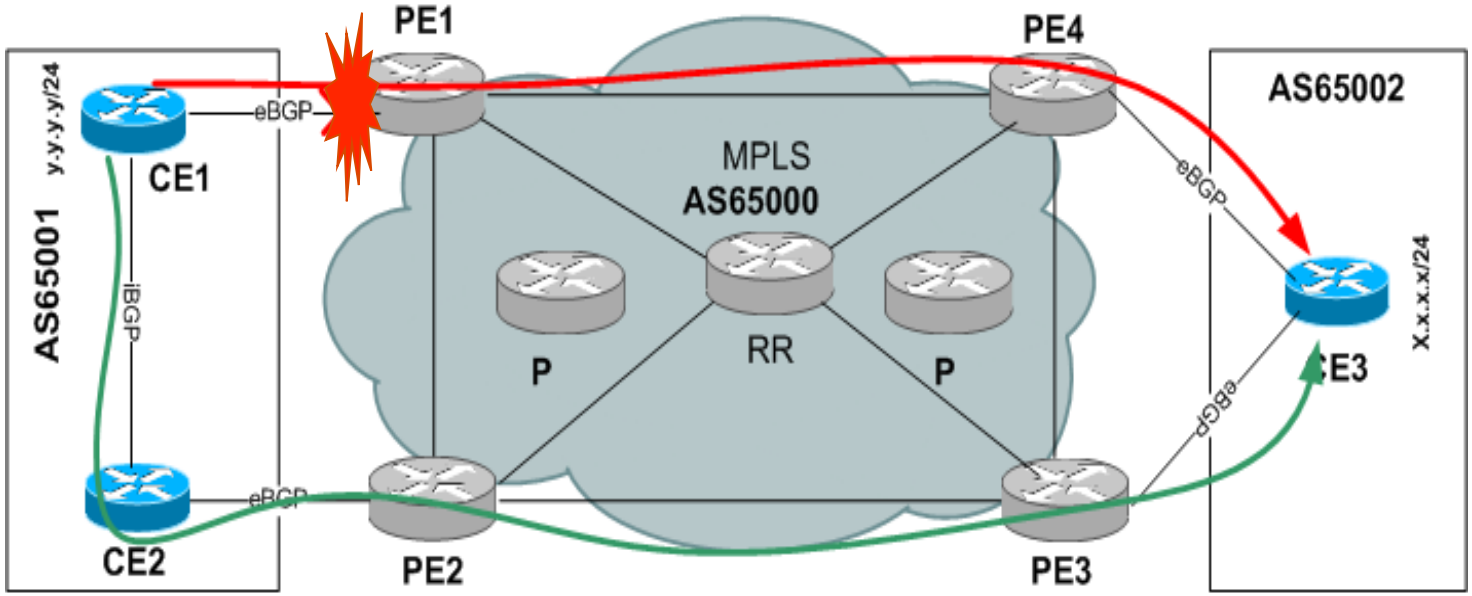
How To Work: PE-CE Link/PE Failure



- eBGP sessions exist between the PE and CE routers.
- Traffic from CE1 uses PE1 to reach network x.x.x.x/24 towards the router CE2. CE1 has two paths:
- PE1 as the primary path and 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 and CEF plane. When the CE1-PE1 link/PE goes down, CEF detects the link failure and points the forwarding object to the backup/alternate path. Traffic is quickly rerouted due to local fast convergence in CEF.

SANOG

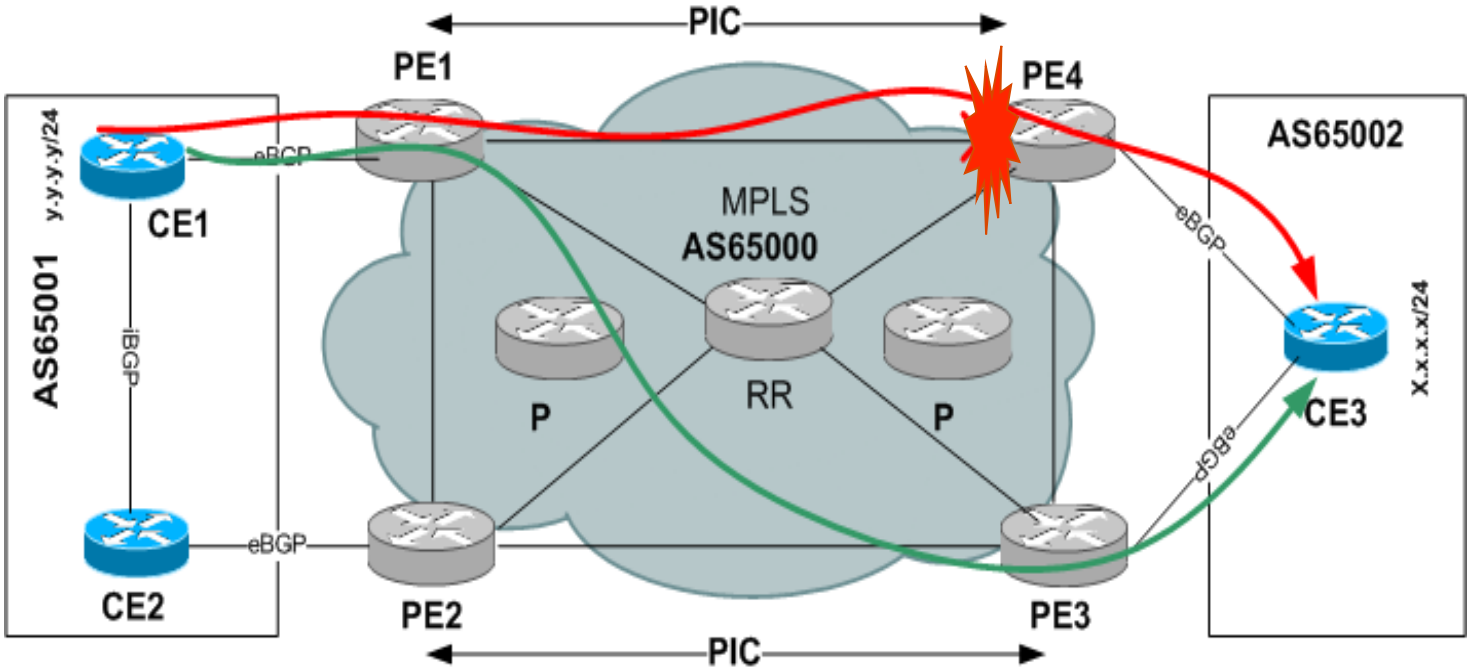
How to Work: Dual CE-PE Line/Node Failure



- eBGP sessions exist between the PE and CE routers. Traffic from CE1 uses PE1 to reach network x.x.x.x/24 through router CE3.
- CE1 has two paths: PE1 as the primary path and PE2 as the backup/alternate path.
- An iBGP session exists between the CE1 and CE2 routers.
- If the CE1-PE1 link or PE1 goes down and BGP PIC is enabled on CE1, BGP recomputes the best path, removing the next hop PE1 from RIB and reinstalling CE2 as the next hop into the RIB and Cisco Express Forwarding. CE1 automatically gets a backup/alternate repair path into Cisco Express Forwarding and the traffic loss during forwarding is now in subseconds, thereby achieving fast convergence.

SANOG

How to Work: IP MPLS PE Down



- The PE routers are VPNv4 iBGP peers with reflect routers in the MPLS network.
- Traffic from CE1 uses PE1 to reach network x.x.x.x/24 towards router CE3. CE3 is dual-homed with PE3 and PE4. PE1 has two paths to reach CE3 from the reflect routers: PE4 is the primary path with the next hop as a PE4 address.
- PE3 is the backup/alternate path with the next hop as a PE3 address.
- When PE4 goes down, PE1 knows about the removal of the host prefix by IGPs in subseconds, recomputes the best path, selects PE3 as the best path, and installs the routes into the RIB and Cisco Express Forwarding plane. Normal BGP convergence will happen while BGP PIC is redirecting the traffic towards PE3, and packets are not lost.

SANOG

Configuration Template

```
router bgp 65000
 no synchronization
neighbor 10.0.0.10 remote-as 65000
 neighbor 10.0.0.10 update-source Loopback0
no auto-summary
 !
 address-family vpnv4
  bgp additional-paths install
  neighbor 10.0.0.10 activate
  neighbor 10.0.0.10 send-community both
exit-address-family
 !
 address-family ipv4 vrf abc
  import path selection all
  neighbor 10.10.10.20 remote-as 65534
  neighbor 10.10.10.20 activate
 exit-address-family
```

Conclusion

IGP Fine tuning

- 100% Dynamic and simplified can reach sub second convergence time

LFA-FRR

- LFA Tunnel Pre-computed, pre-installed
- Prefix-independent
- Simple, deployment friendly, good scaling
- Can reach < 50 ms convergence time suitable for Intra-AS and L2-VPN traffic
- But
  - Topology dependant
  - IPFRR IGP computation is very CPU-intensive task

BGP PIC

- Can achieve < 50 ms convergence time for Inter-AS and L3-VPN traffic

Thank You