BGP Convergence in much less than a second

Clarence Filsfils - cf@cisco.com

 Cisco Public

Down Convergence

Default metric = 1

20

Src

R

R

F

Link L

Dst
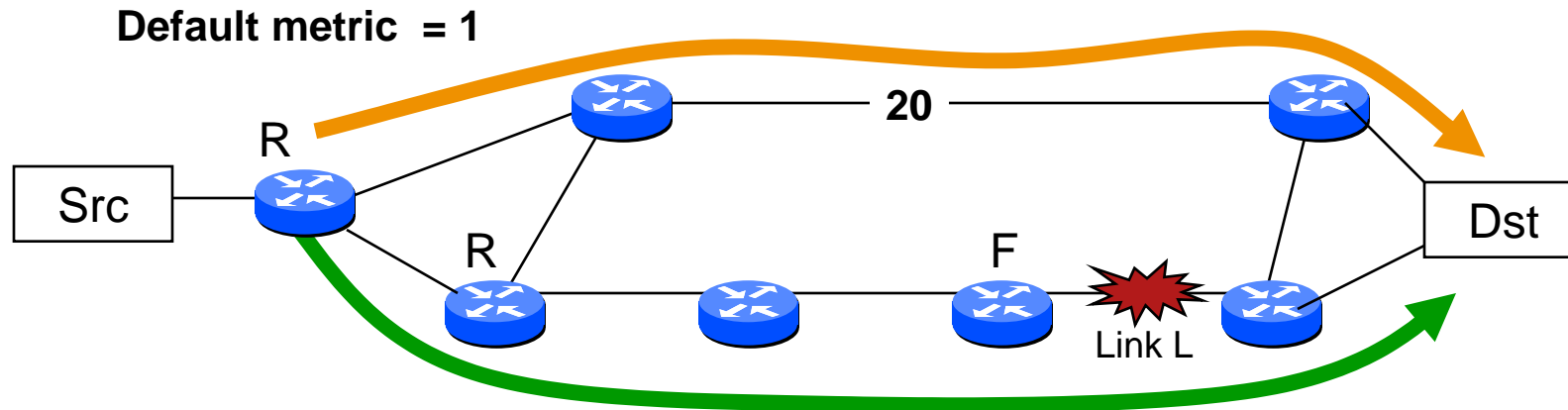
- Assume a flow from Src to Dest

- T1: when L dies, the best path is impacted
    - loss of traffic

- T2: when the network converges, a next best path is computed
    - traffic reaches the destination again

- Loss of Connectivity: T2 – T1, called "Down convergence" hereafter

- Analyzed for streams going to IGP and BGP learned prefixes

Up Convergence

Default metric = 1
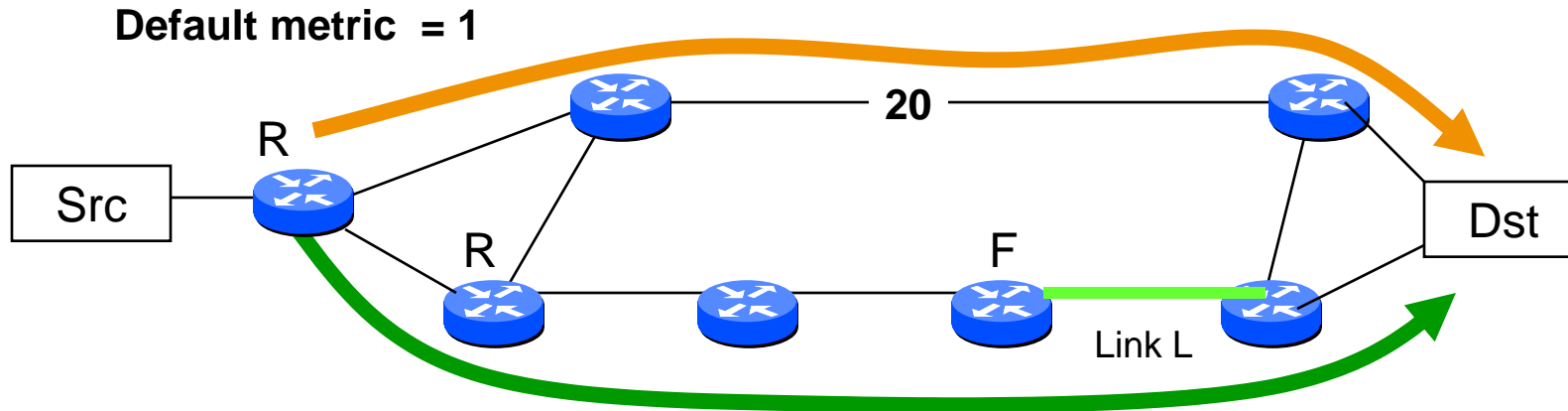
- Assume a flow from Src to Dest
- T1: when L comes up, a better best path is available
    - there is no traffic loss
- T2: the new bestpath is applied
    - there is no traffic loss
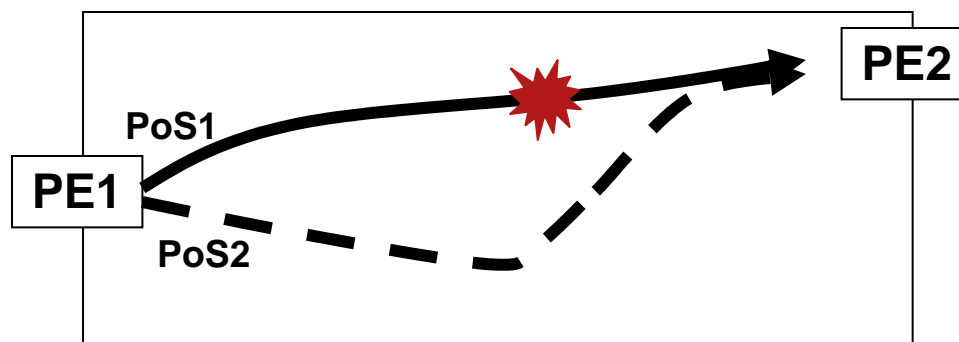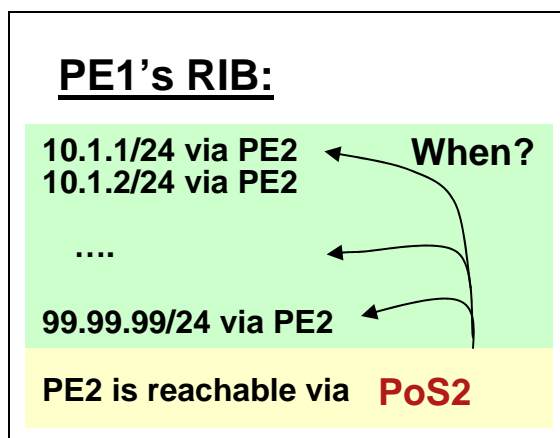- "T2-T1" is called "UP convergence". This does not imply any loss of connectivity

Convergence – Down versus UP

- Down convergence

  - impacts customer service (there is loss between T2 and T1)

- UP Convergence

  - does not impact customer service (there is no loss between T2 and T1 because modern router implementation supports lossless switch-over from valid to valid path)

- **This paper focuses on "BGP Convergence upon Down transition"**

BGP PIC Core

        Cisco Public

BGP Prefix-Independent Convergence Core failure

PE1's RIB:

10.1.1/24 via PE2 ← **When?**
10.1.2/24 via PE2

....

99.99.99/24 via PE2

PE2 is reachable via **PoS2**

PoS1
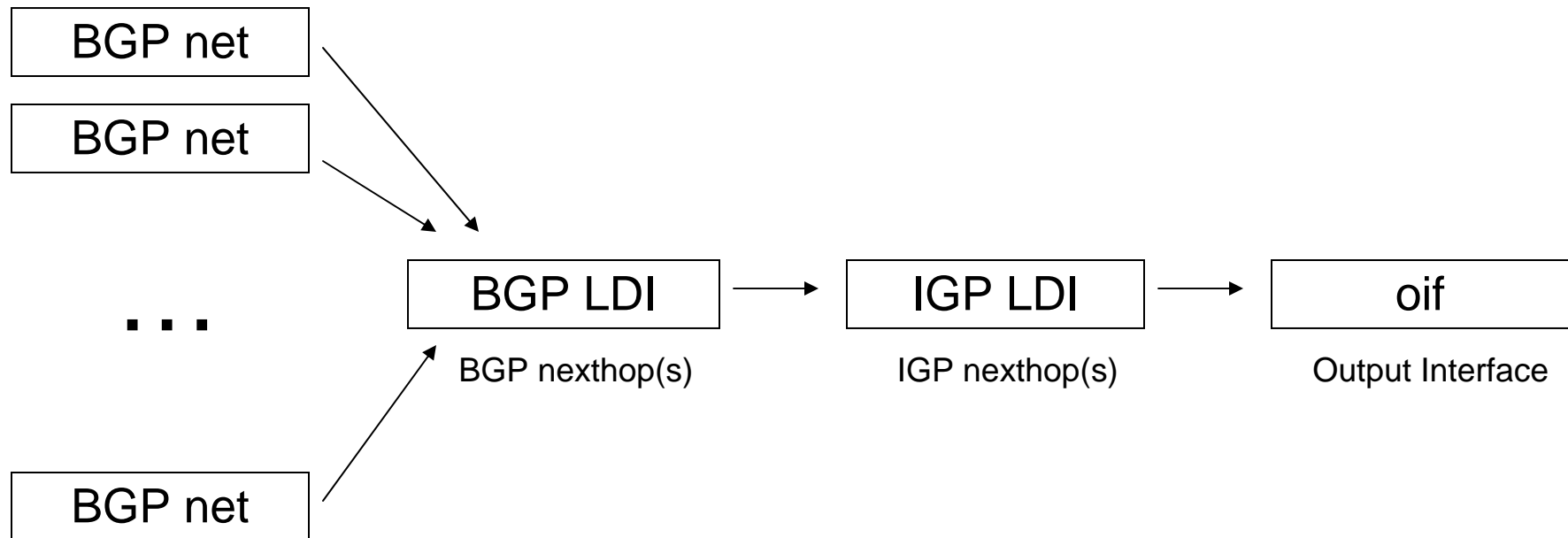
PE1

PoS2

PE2

- Upon a core failure, the IGP finds an alternate path to the BGP next-hop PE2 in a few 100's of msec

- Requirement: the BGP prefixes depending on reachability to PE2 must leverage the new ISIS path as soon as it is updated in the FIB

The right behavior is "BGP PIC/core"

iox-crs1-P-ISIS-BGPipv4-dpi2i-021307_154358-l-nlb

Agilent measurements

The measured convergence to the BGP nhop

The measured convergence to the first BGP dependent to this BGP nhop

Legend:
- 0%
- 50%
- 90%
- 100%
- 0%
- 50%
- 90%
- 100%

The measured convergence to the middle (175k) BGP dependent to this BGP nhop

The measured convergence to the last (350k) BGP dependent to this BGP nhop

Prefix 1 to 5000 are ISIS prefixes. The BGP nhop is an important ISIS prefix and hence is prioritized to the very start of the convergence

Chart: y-axis "traffic loss (msec)" ranging 0 to 400; x-axis "prefix nr" ranging 0 to 7000.

BGP PIC effect

- Testbed: Tier1 ISP topology, CRS1, IOX3.5, 5000 ISIS prefixes, 350k IPv4 BGP dependents to impacted BGP nhop
  - same behavior for VPNv4 as of 3.5
- When ISIS converges, all the BGP dependents immediately leverage the ISIS convergence

The right architecture: hierarchical FIB

```
┌──────────────┐
│   BGP net    │──┐
└──────────────┘  │
┌──────────────┐  │
│   BGP net    │──┤
└──────────────┘  │
                  ▼
   . . .    ┌──────────────┐     ┌──────────────┐     ┌──────────────┐
            │   BGP LDI    │────▶│   IGP LDI    │────▶│     oif      │
          ▲ └──────────────┘     └──────────────┘     └──────────────┘
          │   BGP nexthop(s)       IGP nexthop(s)      Output Interface
┌──────────────┐
│   BGP net    │
└──────────────┘
```

- Pointer Indirection between BGP and IGP entries allow for immediate leveraging of the IGP convergence

The unoptimal way: flattened FIB

BGP net	→	oif
BGP net	→	oif

. . .

| BGP net	→	oif
IGP LDI	→	oif
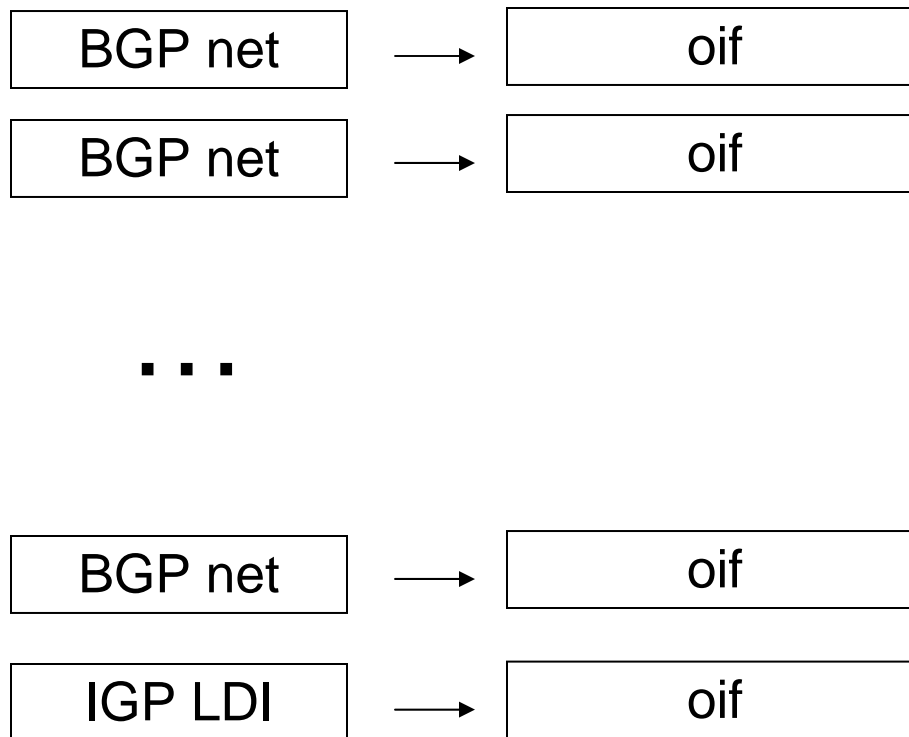
- Control Plane flattens the recursion such that any BGP FIB entry has its own local oif information

Hierarchical FIB - Advantages

- Routing Convergence: BGP PIC Core

  - the BGP dependents converge at IGP convergence of their nhop

- Scaling and Robustness

  - Smaller FIB Memory (thanks to sharing)

  - Much less CPU requirement (no need to reflatten all BGP prefixes upon IGP change)

- Commercially available – proven

BGP PIC Edge

BGP PIC Edge
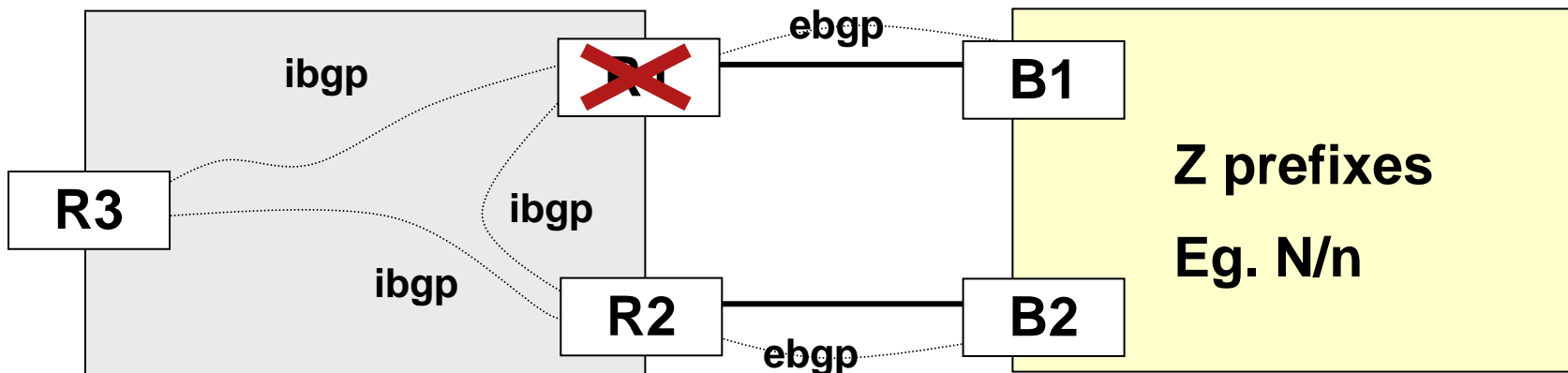
Two ISP's peer at multiple locations and exchange 350k BGP prefixes.

R3, a typical PE within the grey ISP, installs each of these 350k BGP prefixes as multipath pair entries (BGP nhop B1, BGP nhop B2)

We send traffic from R3 to each of the 350k IPv4 BGP prefixes of the yellow ISP. We measure the loss of connectivity upon a peering node failure (R1)

The UUT (R3) is a 12k running IOX 3.3

Results

Agilent measurements

LoC (ms) vs *prefix nr*

Legend: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
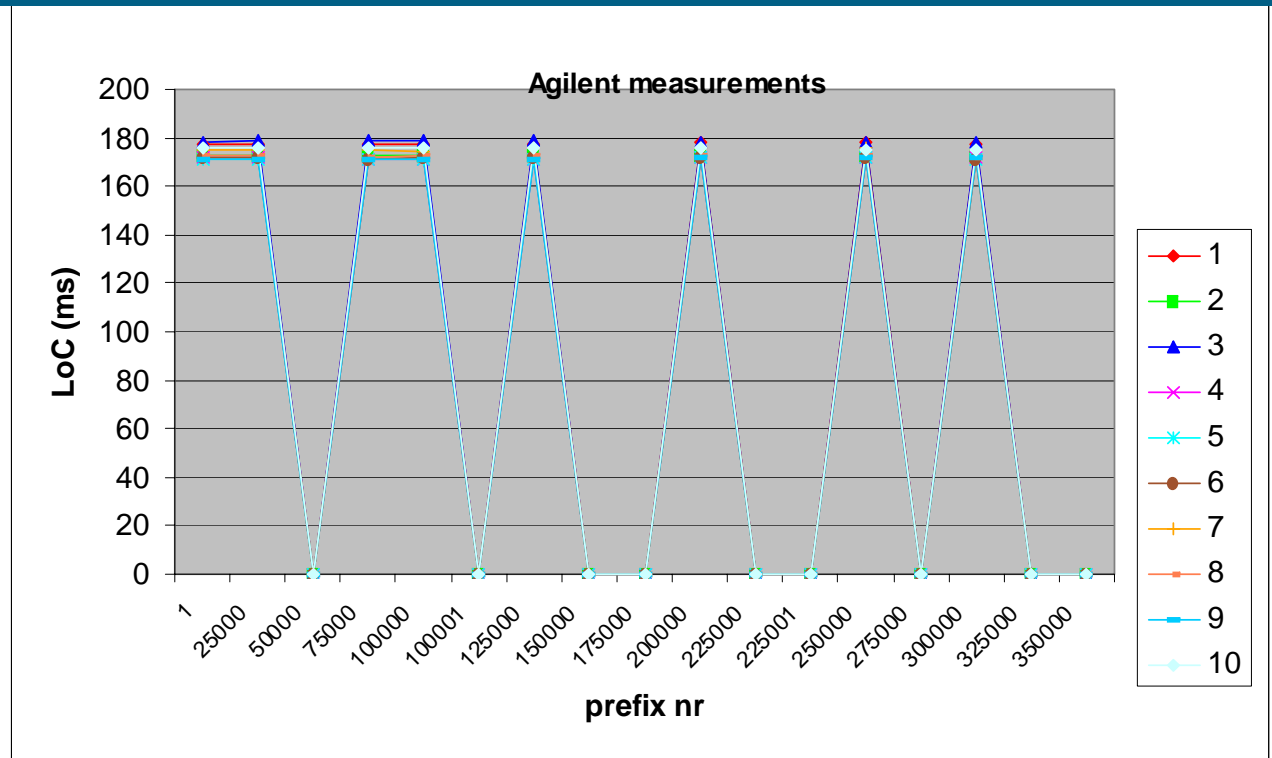
- **180msec !**

  - In 180msec, as a consequence of its IGP convergence, R3 has updated its FIB tree such that all the 350k IPv4 BGP prefixes be forwarded via the alternate path through R2

  - The normal BGP convergence will then take place which involves R2 sending 350k of withdraws and R3 running 350k of bestpath's and RIB modifications. This process will likely take multiple tens of seconds but has no impact on the service thanks to the FIB fixup provided by "BGP PIC" at IGP convergence time.

If Non BGP PIC edge

R3

R2

B1

B2
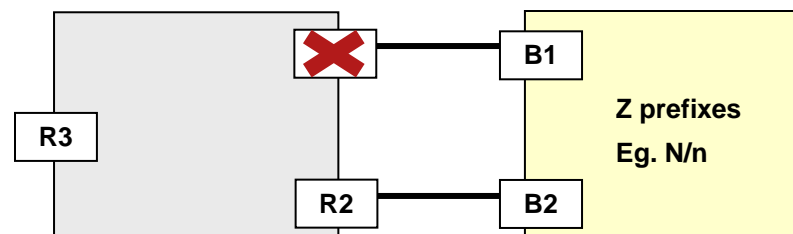
Z prefixes
Eg. N/n

- R3 receives a NHT/down and trigger on a per-prefix basis
  - bestpath (eg. R2 is now best as R1 is gone)
  - RIB/FIB update
  - peer update
- Results: for Z = 100.000 IPv4 BGP prefixes
  - 12k/Eng3/32S: 30 seconds
  - Other vendors: 30 seconds as well
  - Impossible to perform good here due to the prefix-dependency

Another advantage of Hierarchical FIB
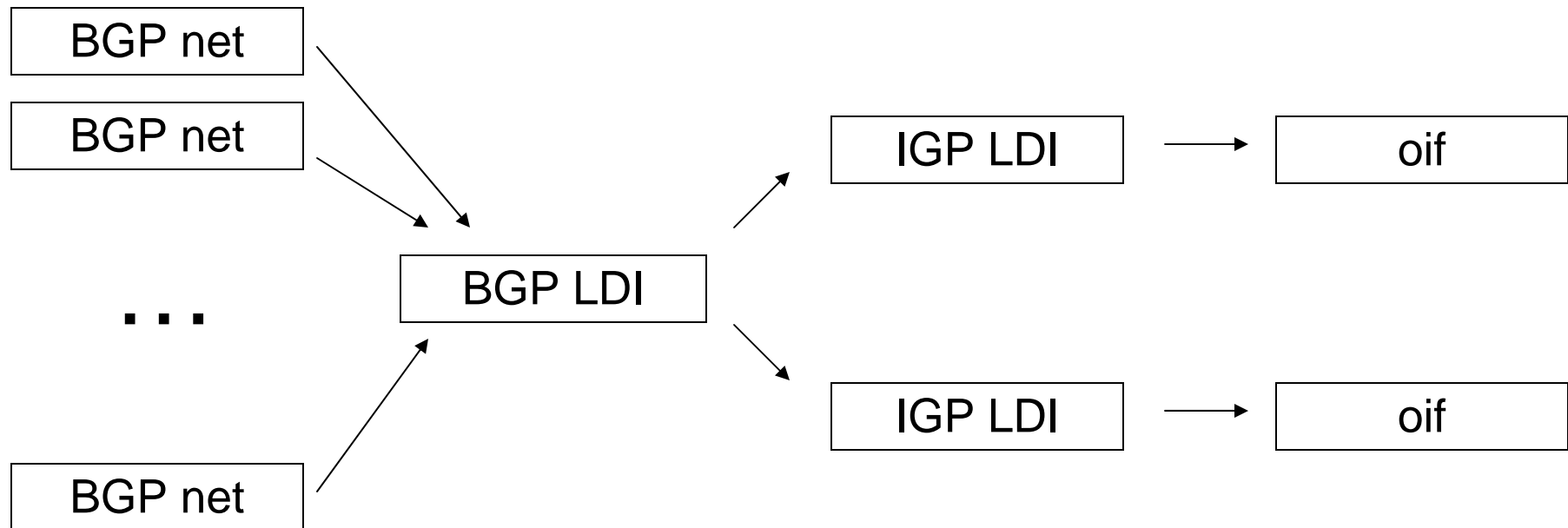
- Pointer Indirection between BGP and IGP entries allow for immediate update of the multipath BGP LDI at IGP convergence

Assumption

- The failure must result into an IGP-based deletion of the BGP nhop

    - PE node failure: its IGP neighbors detect the loss of adjacency and trigger a convergence at all remote PE's which concludes with a deletion of this node from the topology and hence a deletion of any BGP nhop "on that node" (its /32) or on peering links "behind" that node.

    - Peering link failure: next-hop-self should not be used (the BGP nhop is on the peering link) and the IGP should run passive on the peering link. Upon peering link failure, the IGP converges and concludes with a deletion of the BGP nhop

- The BGP prefixes impacted by the failure must have alternate paths installed as multipath BGP entries
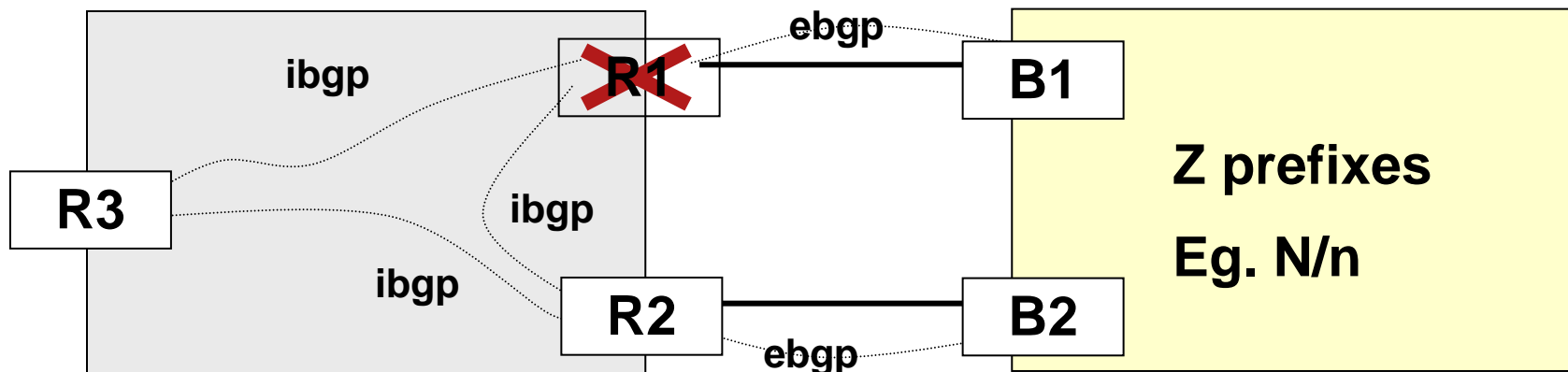
Conclusion

- Hierarchical FIB allows for

  - much lower CPU consumption

  - higher robustness

- And especially

  - BGP PIC Core

  - BGP PIC Edge

Not a futurist talk

- BGP PIC CORE/Edge is available on CRS and 12k, IOX 3.5 for IPv4, IPv6 and VPNv4