To Add-Paths or not to Add-Paths

ftp://ftpeng.cisco.com/raszuk/addpaths/

Robert Raszuk

IOS Routing Development

raszuk@cisco.com

Objective

To present how to achieve fast connectivity restoration and BGP level load balancing for both IP (and 3107) networks without the need for BGP protocol extention and entire network wide upgrade to new operating system.

More important it is to show how the above is possible without need to increasing amount of BGP state in your routers.

Agenda

- A perspective on building IPv4 or IPv6 networks with fast connectivity restoration and load balancing
- Add-paths reality
- Different approach to distribute N paths in BGP with no need for bgp protocol changes

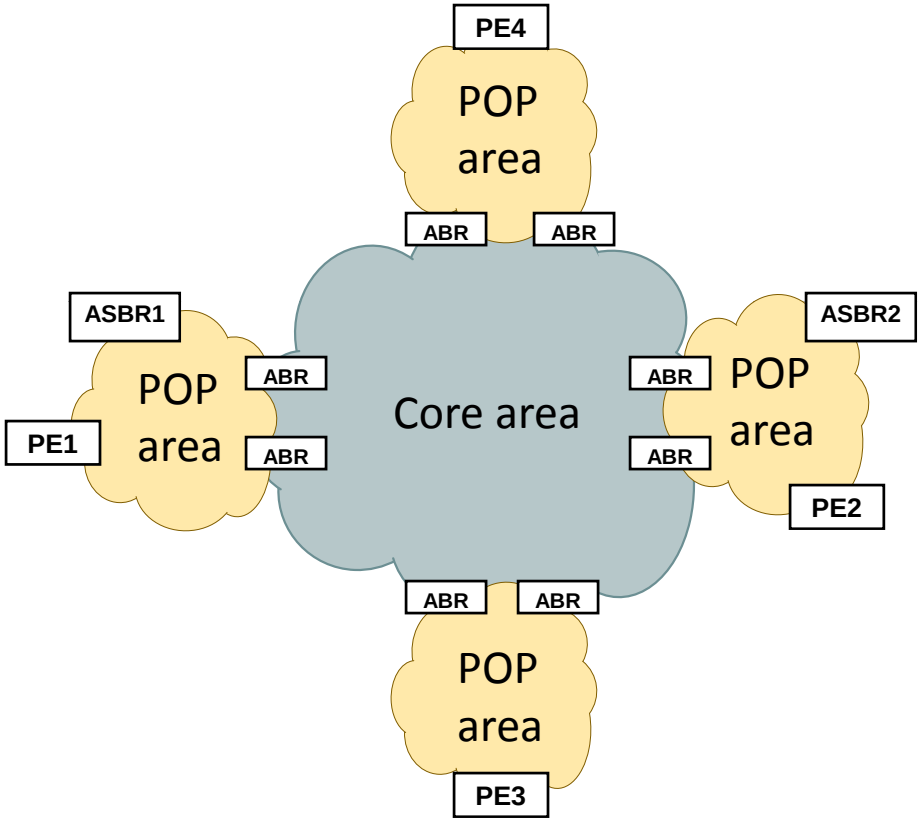
- How to provide path redundancy in RFC3107 networks (if time permits)
- How to operate without new POs when Internet table size exploads on us (if time permits)

Anti-Agenda

- We assume different RD per vrf for VPNv4 & VPNv6 so no issue there. Not a reason to look into add-paths → Out of scope

- Same RD or Inter-as ASBR redundancy issues for L3VPNs can be addressed without add-paths → Out of scope

Hierarchical IGP ISP design



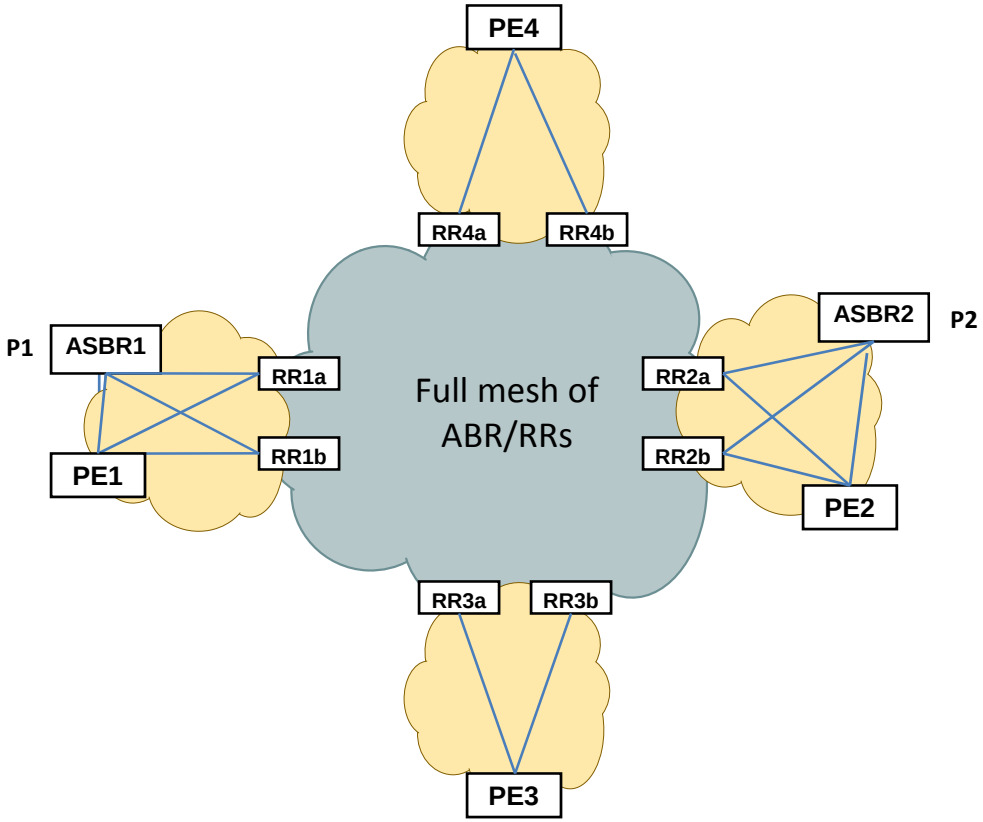
Basic design:

- Hierarchical IGP design
- Each POP in a separate IGP area
- ABRs acting as RRs
- RRs in the data plane on the POP to core boundary
- Fully meshed RRs in the core

So what is missing for Fast Convergence or BGP level load balancing:

- In fact NOT MUCH !
- Enabling best external on edge routers and on RRs is all what is needed in this typical design to provide full BGP paths distribution where functionally needed
- Enable IGP domain wide BGP next hop state propagation

Hierarchical IGP ISP design



Design assumptions:

In the POP:

- ASBRs/PEs advertise their external paths to POP's RRs (even if IBGP path is selected as overall best) – thx to best external feature on ASBRs and PEs
- IBGP full mesh
- Last resort POP's local default route to ABRs/RRs interfacing with the core with admin distance over 200 (IBGP)

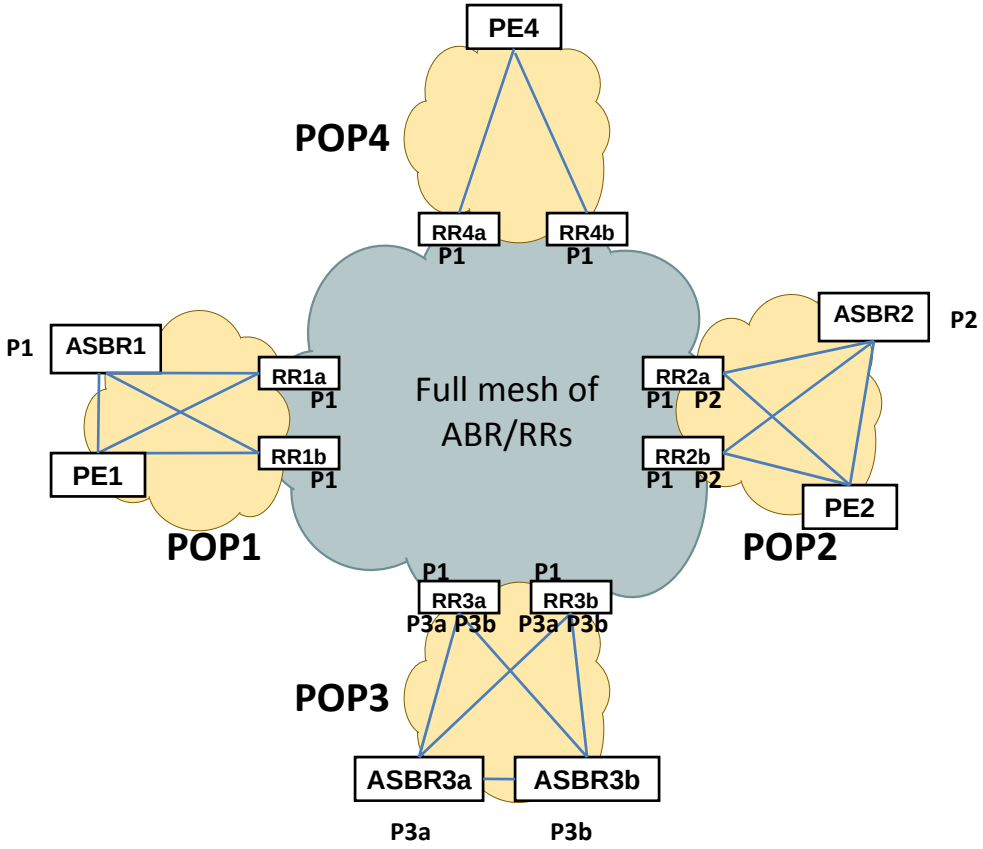
In the core:

- RRs advertise POP's best paths to other RRs
- If best external is enabled on RRs towards the core local/POP's best path will be advertised even if overall path received from other RR's cluster is selected as best.
- IBGP full mesh

Let's walk through main design scenarios for basic IPv4/IPv6 Internet

- **Scenario A – no best external on RRs and no next hop propagation**
- **Scenario B – no best external on RRs and next hops event propagation**
- **Scenario C – enabled best external on RR's and next hops event propagation**
- **Scenario C' – similar case like C, but with flat IGP architecture**
- **Scenario D – DMZ or single POP exit architecture**

Hierarchical IGP ISP design



Intra-domain Path distribution:

Let's observe how paths are being distributed:

Let's assume we have 4 paths for a given prefix P: P1, P2, P3a, P3b.

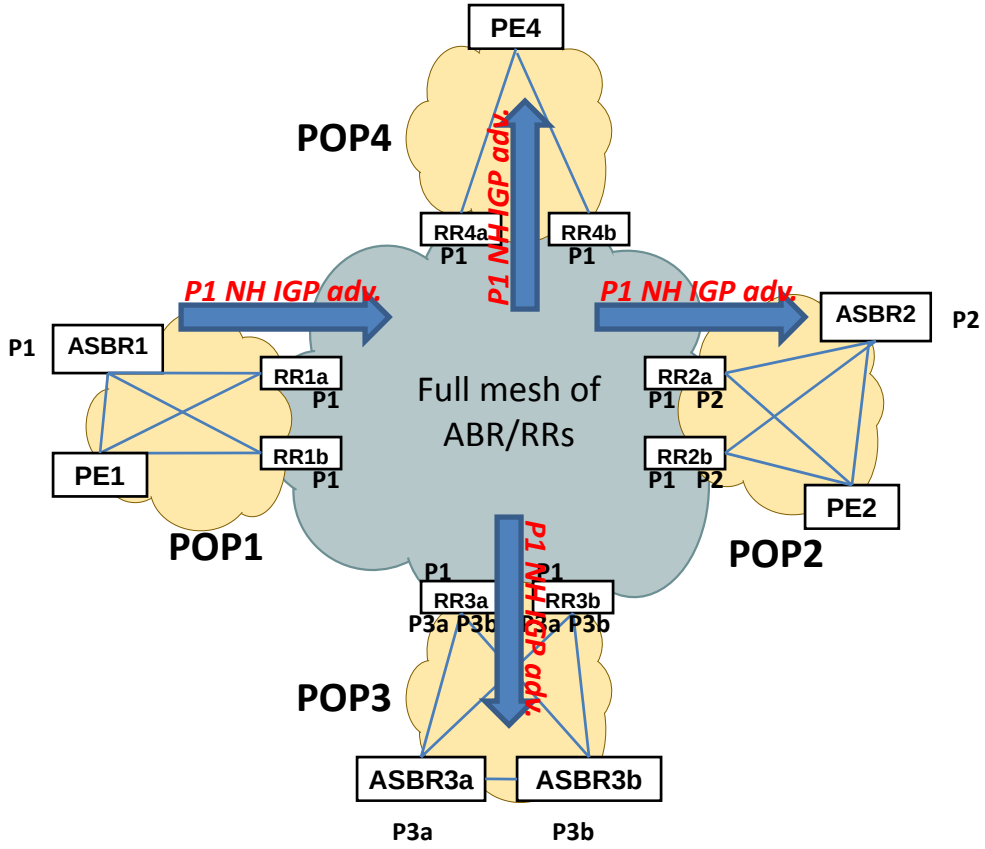
Let's assume P1 is best due to local preference. Best external on edges.

Scenario A – no best external on RRs and no next hop event propagation beyond local (best path's) POP

When P1 goes down RR1s need to withdraw P1 in BGP then subsequent best will be advertised by RR2 or RR3

End-2-end connectivity restoration: SLOW f(bgp) + NO PIC + best POP LB

Hierarchical IGP ISP design



Note ... PIC here takes place primarly on POP to core boundary - ABRs/RRs as well as on ASBRs/PEs within the POP.

Intra-domain Path distribution:

Let's observe how to relax the need for service impacting BGP withdraw

Scenario B – no best external on RRs and next hop event propagation via IGP domain wide

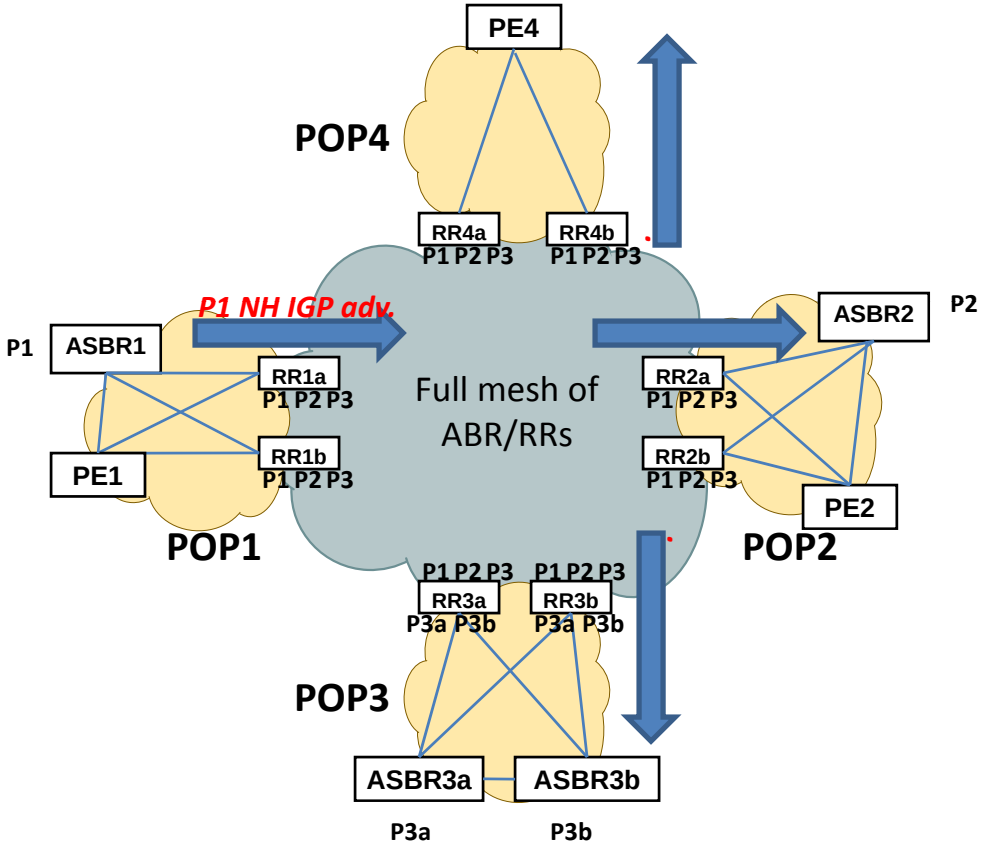
When P1 goes down IGP will flood it's next hop down event network wide. (POP1 and POP4 still needs to wait for BGP !)

This will trigger simultaneus invalidation of P1 on all RRs as well PIC where backup paths are present RRs/ASBRs/PEs

BGP will advertise P2 or P3a/P3b as best

End-2-end connectivity restoration: FASTER + SOME PIC f(igp) + best POP LB

Hierarchical IGP ISP design



Intradomain Path distribution:

Let's observe how to restore connectivity without waiting for BGP:

Scenario C – enabled best external on RR's & advertise next hops events domain wide

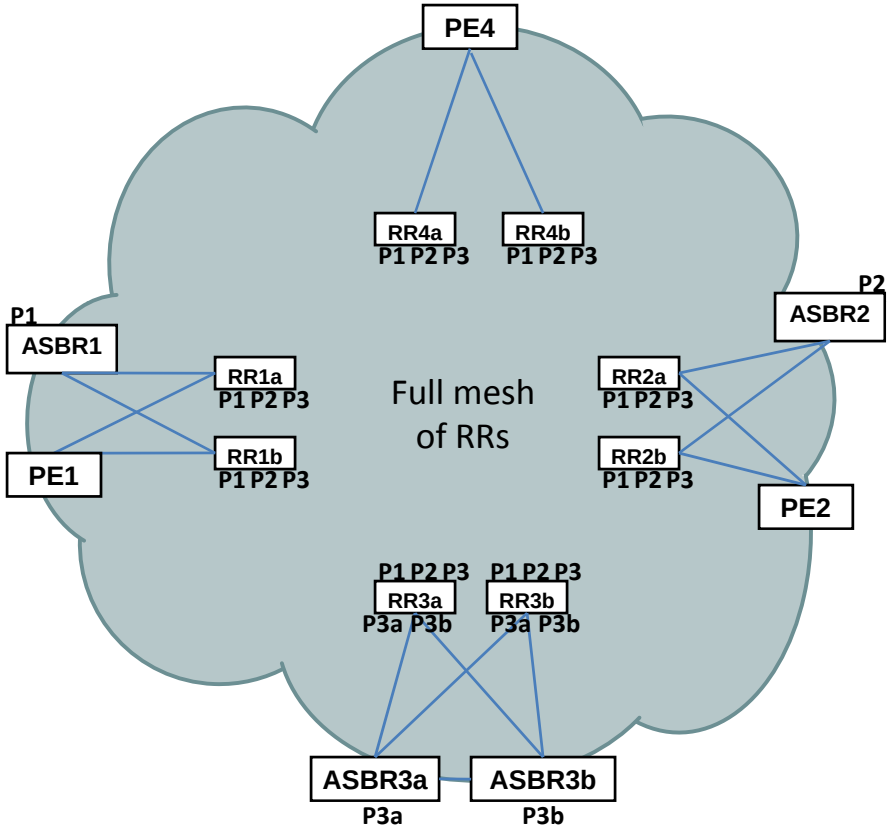
Benefits:

- Fast connectivity restoration
- BGP PIC on RRs, ASBRs and PEs
- IBGP loadbalancing (local and remote)
- No need for upgrade all the ASBRs/PEs
- No need to push all the paths to all ASBRs/PEs

Note ... PIC here takes place primarly on POP to core boundary - ABRs/RRs as well as on ASBRs/PEs within the POP.

End-2-end connectivity restoration:
FASTEST + FULL PIC + FULL LB !

Flat IGP ISP design

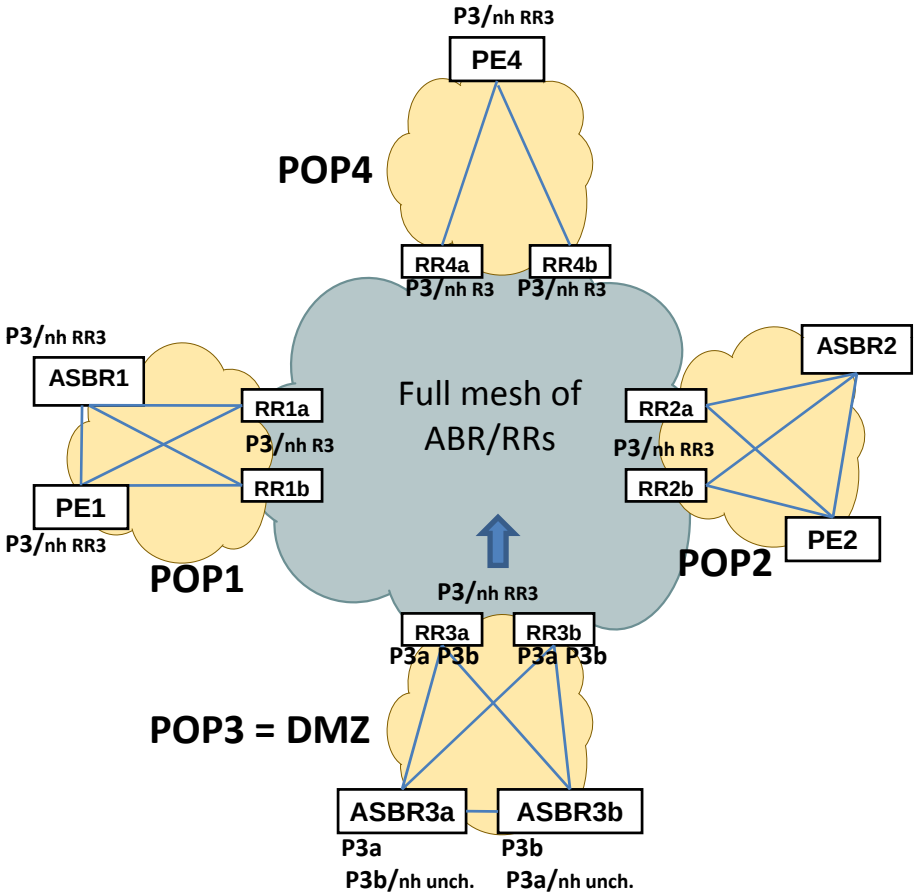


Scenario C' – enabled best external on RR's, advertise next hop state domain wide (flat IGP no need for configured leaking core to POP)

Design guide:

- ASBRs/PEs are peering to two different RRs which are in the data path
- ASBRs/PEs have static default towards RRs with admin distance over 200
- ASBRs/PEs have best external and PIC on
- RRs have best external and PIC on
- Full mesh of RRs directly connected, in a ring or using some encapsulation technology
- Works with next hop self on peering points or with next hop unchanged

Original/typical ISP design



Scenario D – DMZ or single POP exit architecture

- Specific architecture where your entire exit traffic goes via the same POP or same ASBRs

Solution A:

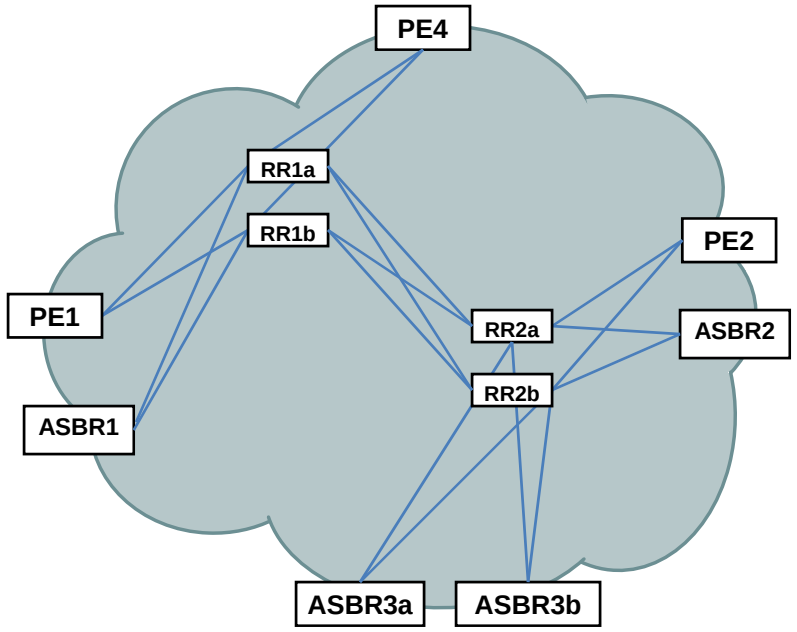
- Split your upstream/peering ASBRs into different POPs → Use scenario C or C'

Solution B:

- Set next hop self on the POP (on RRs or on ASBRs) optionally via anycast address – „ghost loopback".
- Remove fooding and leaking original bgp nh reachability in IGP (not needed)

Why networks got flatten ....

Now pictures of the networks started to looked like this:



And let me just notice that all nice applications like L3VPN (aka 2547) or L2VPNs or multicast VPNs work very well over automatic IP encapsulation in any IGP model. No need to create manually any tunnels, no need to flat your network !

Draw your own conclusions ☺

And there are reasons behind it ....

For any MPLS application end to end LSP needs to be created.

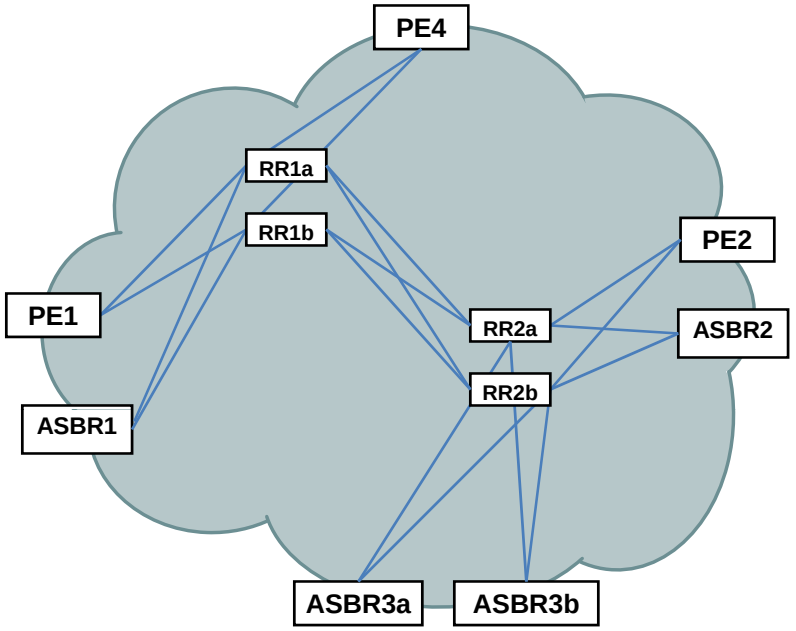
LDP requires the exact match with the RIB (practically /32 host route) – when you have areas you need to leak all /32s across end to end.

TE requires end to end flooding of information about link bandwith reservations to calculate and signal the EROs – Inter-IGP-Area TE is a project on it's own. Another reason for flat IGP.

Is this off-topic ??? Nope ...

Consequences for BGP design

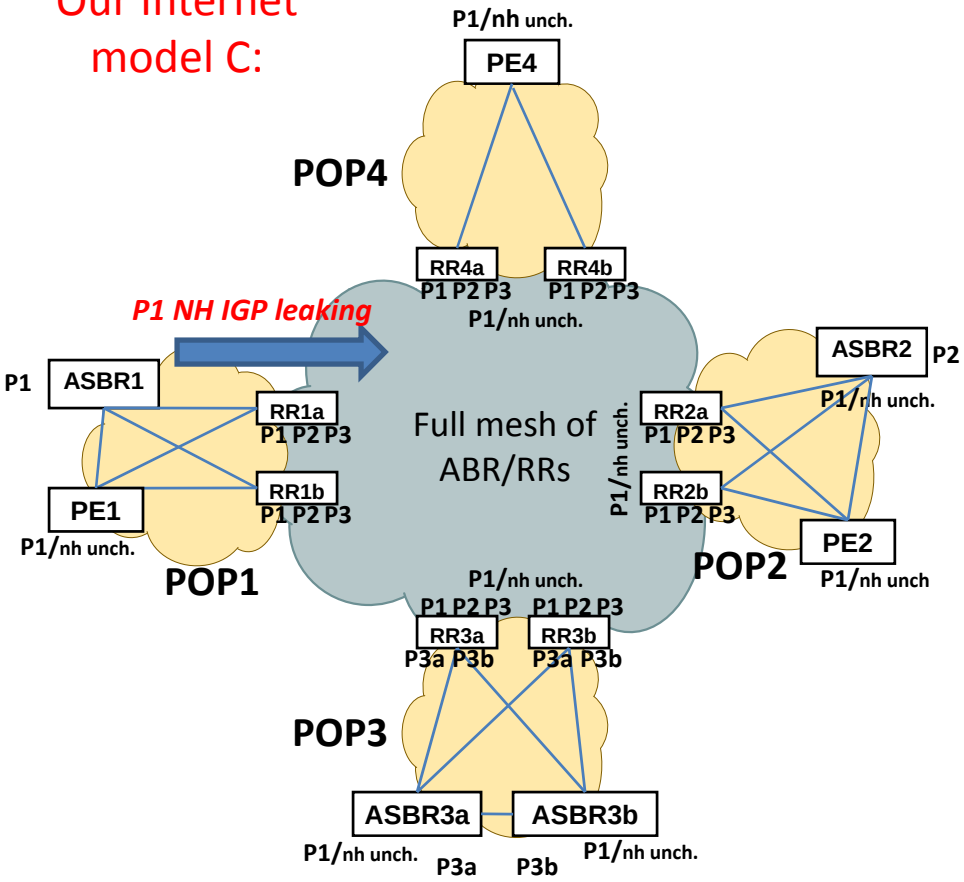
- Concept of flat IGP domains and race between vendors who's IGP code can accomodate bigger areas has started
- As leading application was L3VPN and core could not afford to maintain any VPN state VPN's RRs were naturally control plane devices.
- Some customers had/still have two different networks: flat for VPNs & hierarchical for IPv4/v6, but the price to maintain those is high.
- Some did not flow with the river and started to run L3/L2 VPN services over IP encapsulation - very successfully maintaining their IGP hierarchical model.
- Others following the VPN/MPLS model moved their IPv4/IPv6 RRs into the core control plane devices ... No IP lookup were performed there anyway as LSPs where edge to edge.



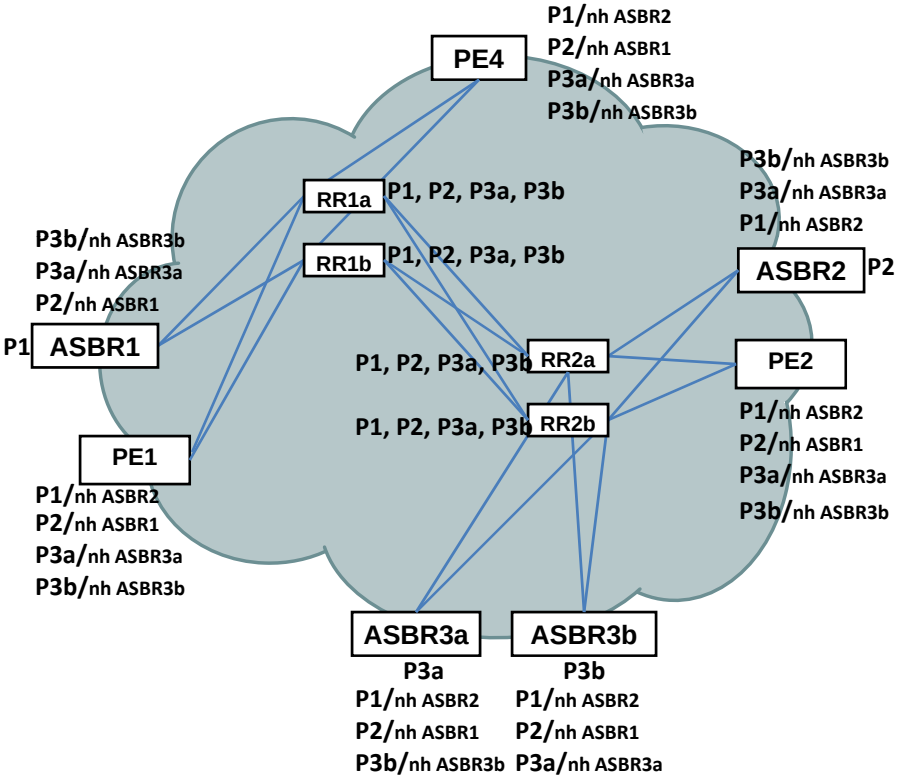
- **Very difficult to provide true hot patato routing without pushing all paths everywhere !**
- **The real issue is that you can not benefit from BGP path virtualization**

So now let's compare our previous designs with the add-paths model:

Our Internet model C:



Let's asume all routers are upgraded to support add-paths:

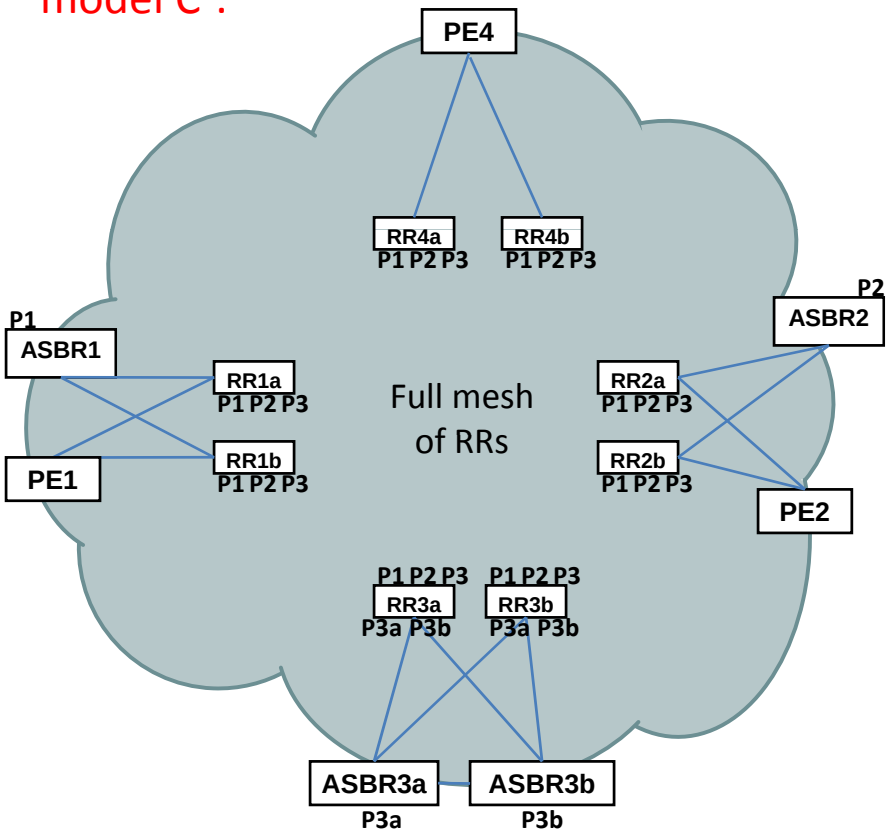


Best external also enabled on all PEs/ASBRs

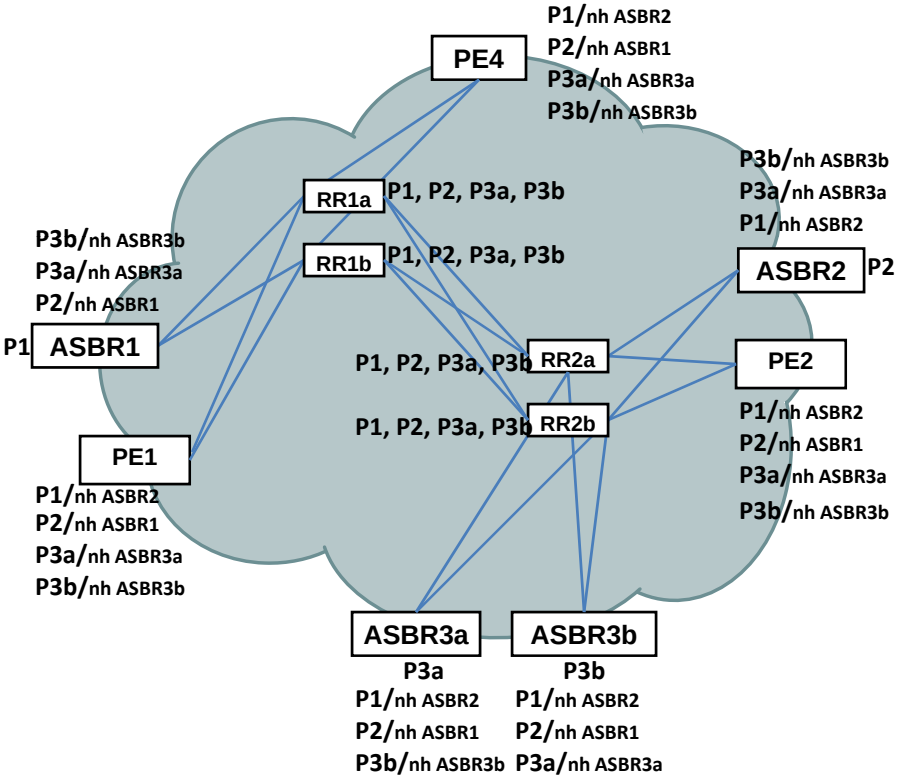
Let's pay attention to difference in the amount of control plane BGP state information on all edge routers (ASBRs & PEs) – and this is only for net with 3 exit points

So now let's compare our previous designs with the add-paths model:

Our Internet model C':



Let's asume all routers are upgraded to support add-paths:

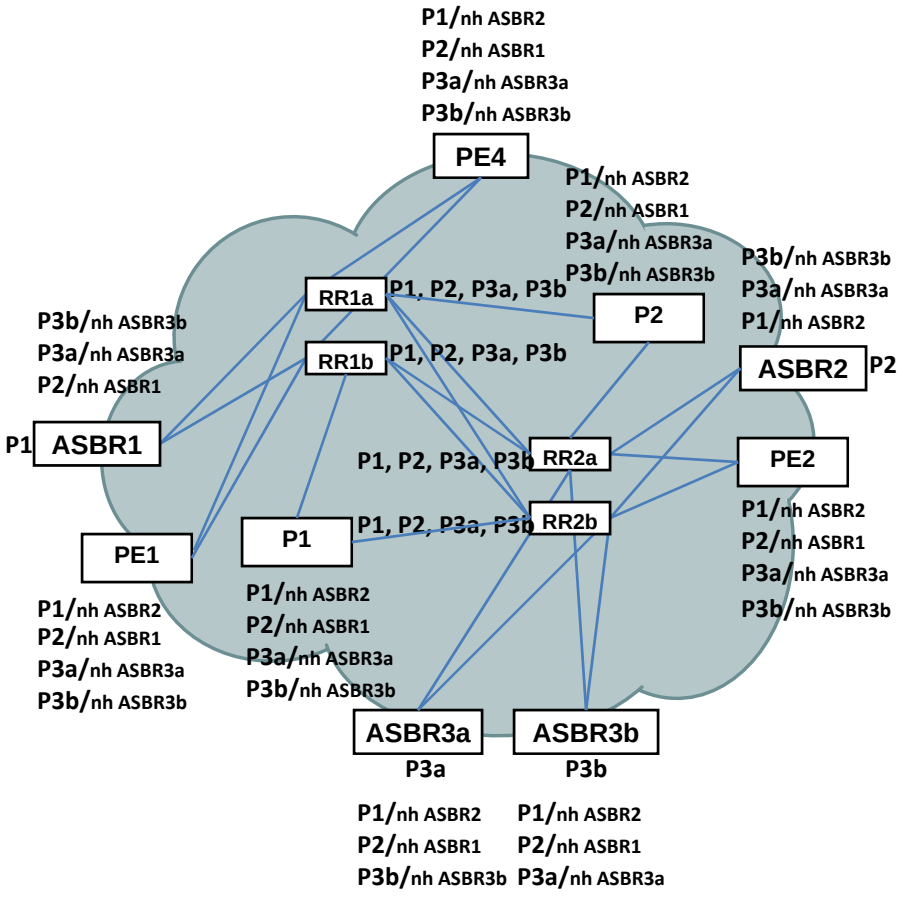


Best external also enabled on all PEs/ASBRs

Let's pay attention to difference in the amount of control plane BGP state information on all edge routers (ASBRs & PEs) – and this is only for net with 3 exit points

Add-paths reality ....

All paths x 2 ! Due to peerings to two RRs



- Notice paths at the core of P1 & P2 ...
- The only other choice is end to end tunneling.

- New BGP protocol encoding, new capability, new network wide upgrade.

New design questions ....

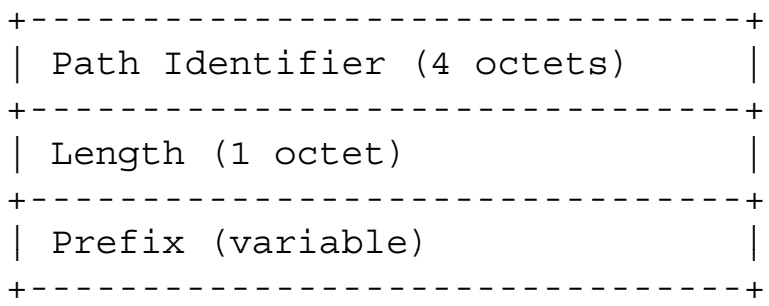
- Which additional paths to distribute ?
  - All paths ?
  - Nth best ?
  - All AS-wide best paths ?
  - Neighbour AS group best paths ?
  - Best local-pref/second local-pref ?
  - Paths at best path penultimate decision

- Would all edge routers will be able to carry the additional control plane load ?

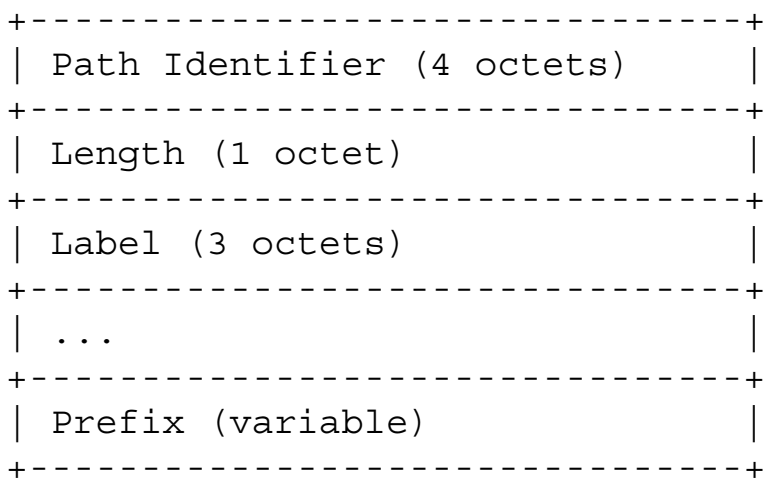
- What is the driver ? Fast connectivity restoration (FC/PIC), load balancing, hot potato routing, oscillations suppresion ? Those can be addressed with the proper network design without add-paths.

Add-paths new NLRI encoding

- NLRI encodings specified in [RFC4271, RFC4760] are extended as the following:



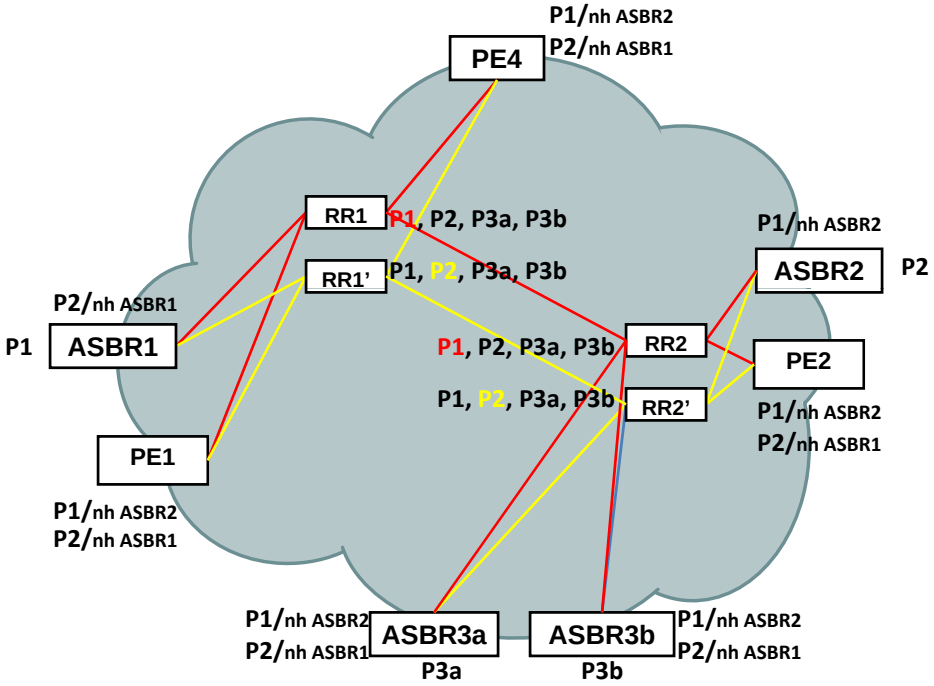
- NLRI encoding specified in [RFC3107] is extended as the following:



Add-paths' easy alternative

In the event of not being able to use design with RRs in the data path ....

Diverse BGP Path Distribution design 1



- RR1' and RR2' are **shadow RRs**
- They are configured to calculate and advertise Nth best path to it's clients
- They can do it on a per AFI/SAFI basis
- Same IGP metric as best RRs or IGP metric disabled on both

- Let's assume that your goal is fast connectivty restoration via FC/PIC and/or IBGP multipath loadbalancing

- P1 overall best, P2 second best

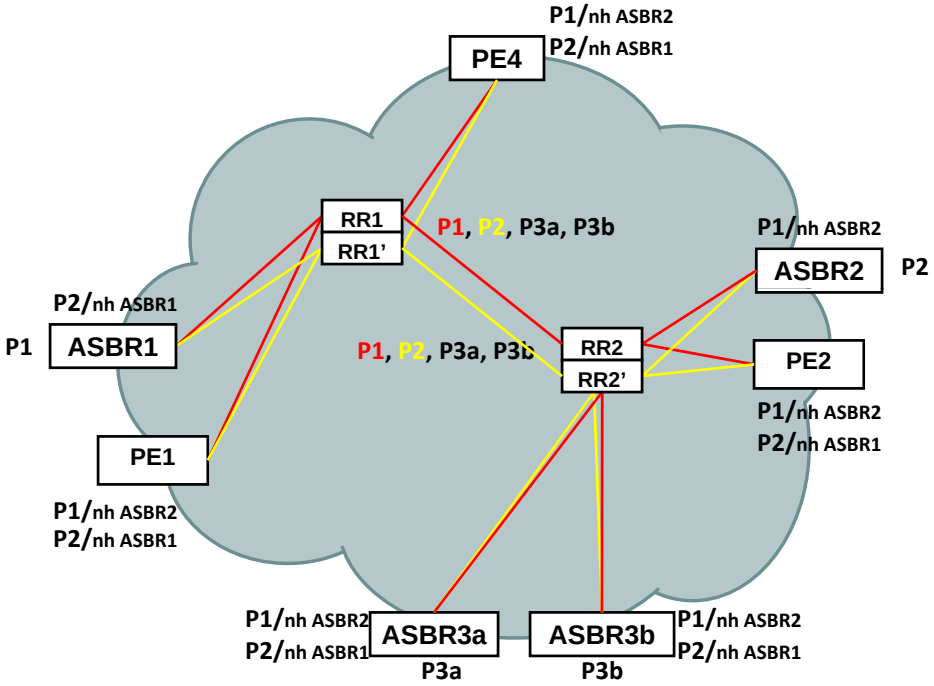
Key benefits:

- Easy deployment – no upgrade of any existing edge router is required, just new IBGP session per each extra path
- One additional „shadow" RR per cluster
- Works within flat domain or within each area of hierarchical network
- No new protocol extension required
- IETF draft: draft-raszuk-diverse-bgp-path-dist-00

Add-paths' easy alternative

In the event of not being able to use design with RRs in the data path ....

Diverse BGP Path Distribution design 2



- RRs and RRs' are **same RRs**
- They are configured to calculate and advertise Nth best path to it's clients on a per neighbor basis
- They can do it on a per AFI/SAFI basis
- Automatic same IGP metric as best RRs

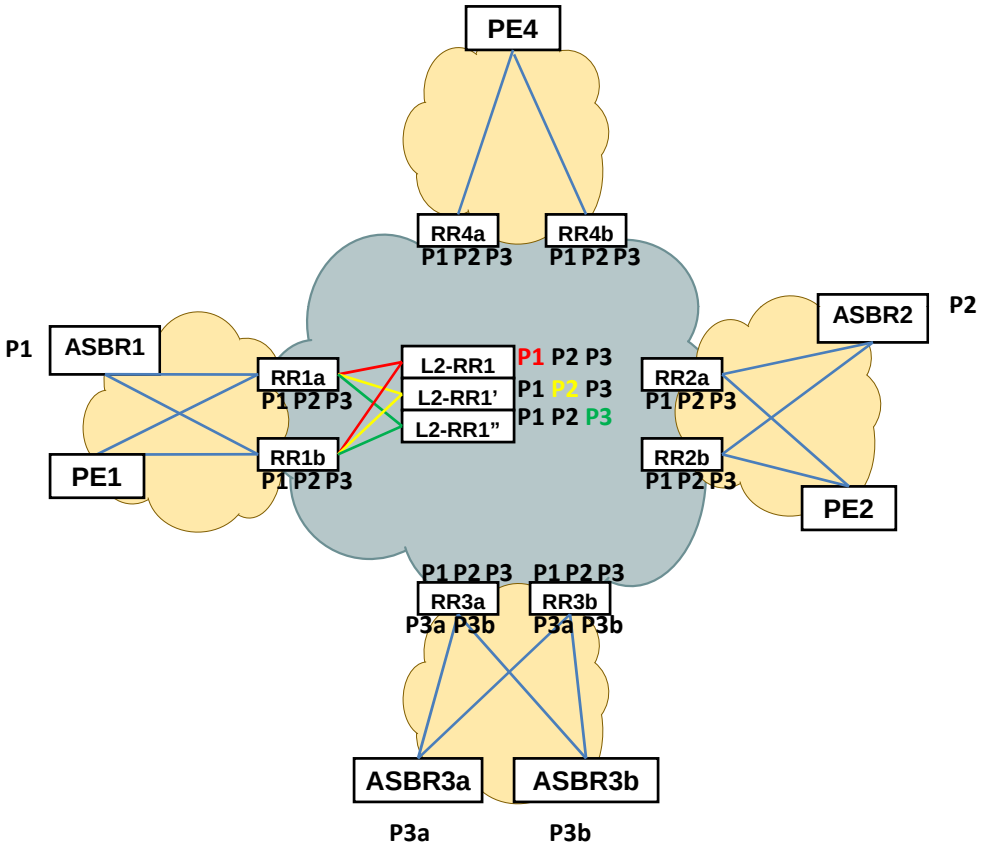
- Let's assume that your goal is fast connectivty restoration via FC/PIC and/or IBGP multipath loadbalancing

- P1 overall best, P2 second best

Key benefits:

- Easy deployment – no upgrade of any existing edge router is required, just new IBGP session per each extra path
- Just few more IBGP sessions and code upgrade on existing RRs
- No new protocol extension required
- IETF draft: draft-raszuk-diverse-bgp-path-dist-00

Instead of full-IBGP mesh in the core/pop …



Design details:

- Let's assume that our goal is to build a redundant network and distribute 2nd and 3rd best paths with different exit points

+

- Removed need to create a lot of IBGP session in the core
- Each shadow RR calculates it's own best path and advertises it to it's clients (POP RRs)
- Any encapsulation can be used within each area IP or MPLS (option).
- Any new application can be build in a scalable manner in such design.

+

- The very same applies to intra POP design instead of assumed full mesh.

Scaling MPLS networks ....

Next 4 slides ...

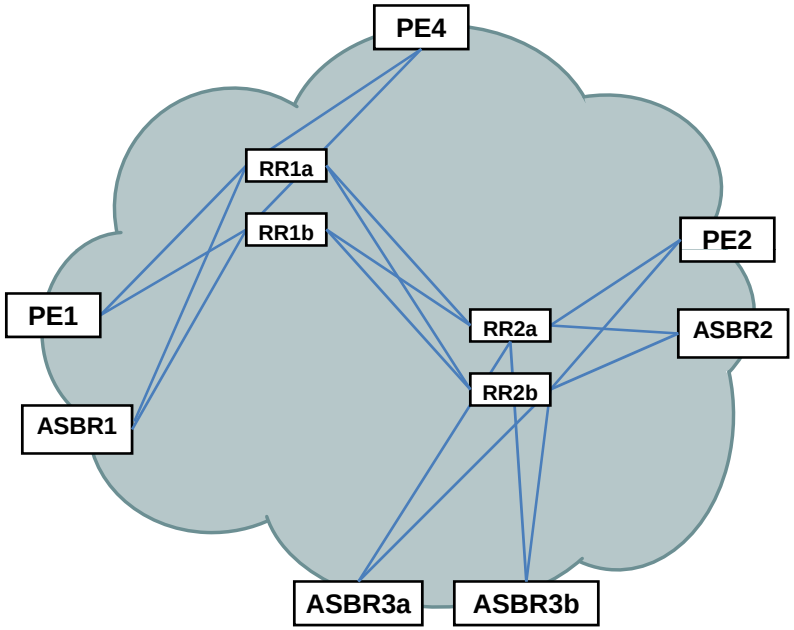
- To present what vendors today tell you on how to scale large MPLS networks
- To address the case of 3107 requirements for add-paths with simple IP like hierarchy

Scaling large MPLS networks

- Remember the basic picture about MPLS end to end LSPs ?
- Perhaps you are considering now how your multiservice network should look like and what should be your choice of encapsulation ?

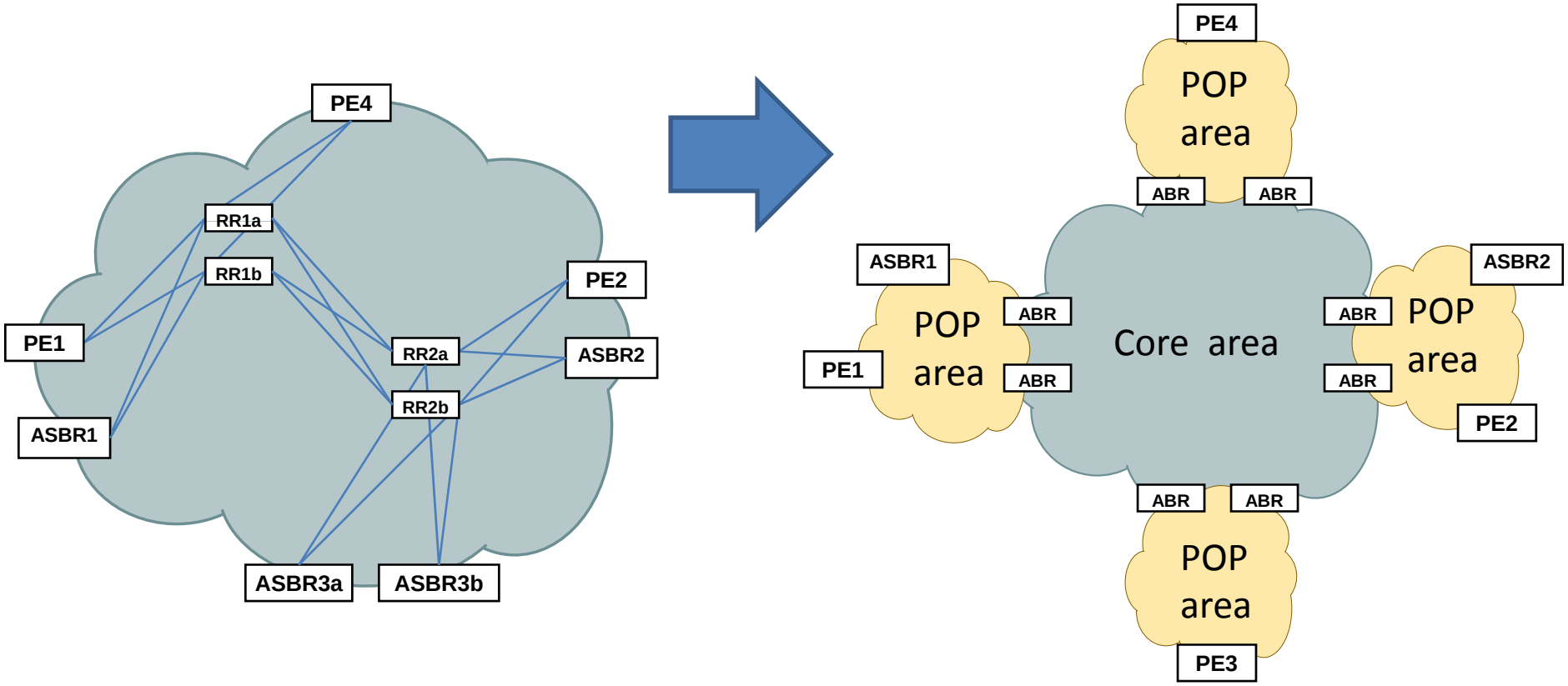
Hint !

- Networks grow and BRAS/DSLAM boxes are becoming PEs ... This results in scale of 10,000 – 30,000 PEs
- I don't think anyone intends to keep it flat.



And please just guess what is the answer to such MPLS scaling challange .....

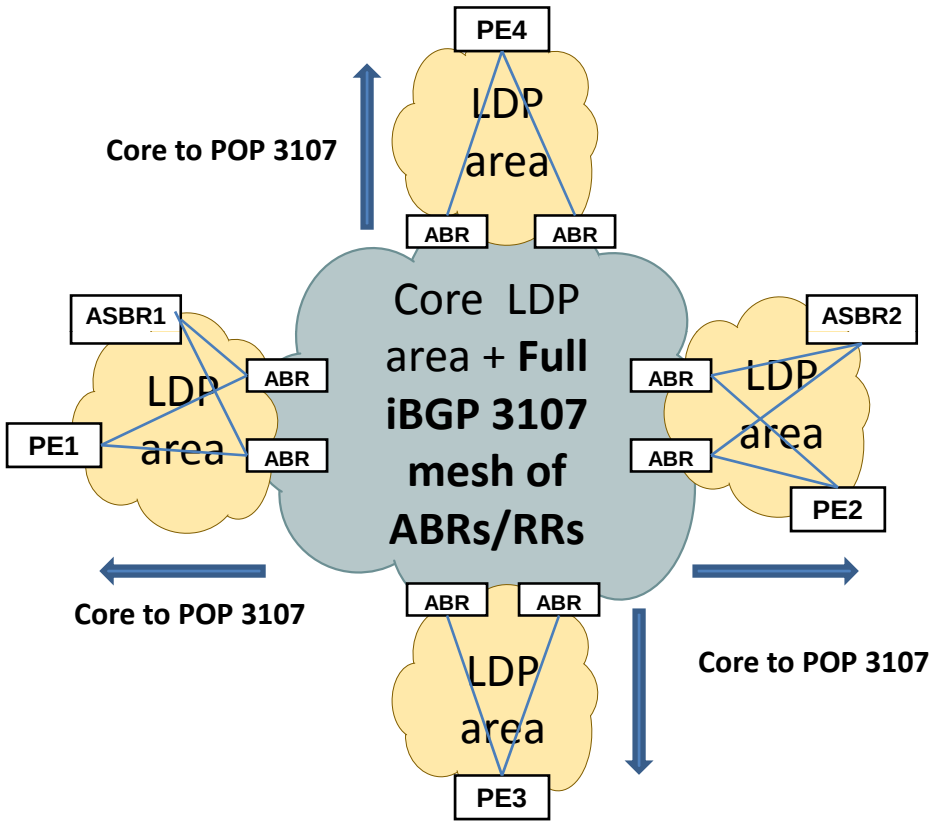
….. to come back to traditional IGP design and introduce hierarchy in MPLS



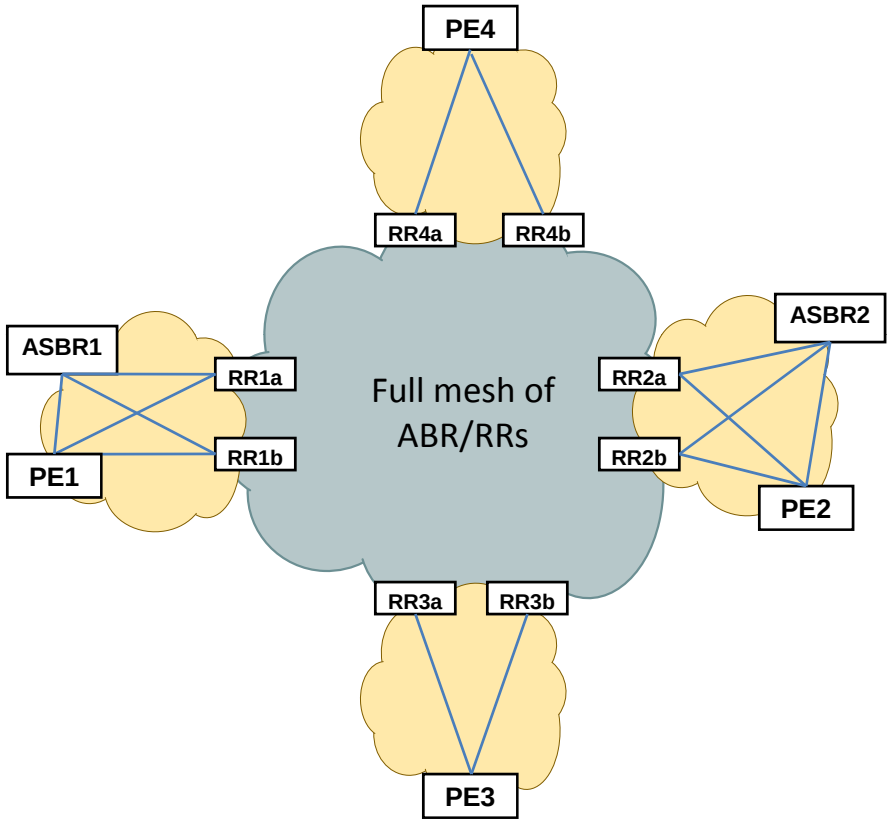
Let's zoom into the MPLS hierarchy .....

MPLS hierarchy vs traditional ISP model

New MPLS hierarchical model for L3VPNs:



IPv4 classic ISP mode from slide 4:

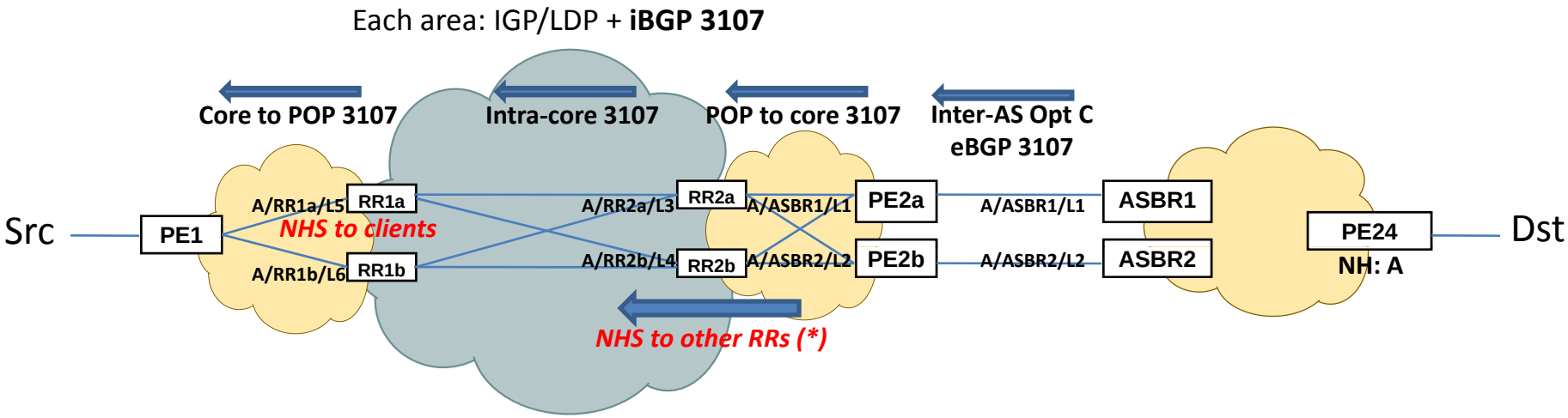


- All PE's loopbacks now carried in BGP SAFI 4 !
- ABRs are RRs doing next hop self for 3107 on /32s
- Origin POP is redistributing IGP+LDP next hops into 3107 BGP
- LDP is local to each area
- Label stack increased +1 due to introduced hierarchy

Q: *Isn't IP encapsulation with it's native summarization cleaner, simpler, nicer ???*

Example of 3107 paths distribution …

Taken scenario of Inter-AS option C to see how it maps to IPv4 ISP scenario C:



RR1's LFIB for prefix A:

L5 -> L3 -> via IGP/LDP to RR2a
L4 -> via IGP/LDP to RR2b

L6 -> L3 -> via IGP/LDP to RR2a
L4 -> via IGP/LDP to RR2b

RR2's LFIB for prefix A:

L3 -> L1 -> via IGP/LDP to ASBR1
L2 -> via IGP/LDP to ASBR2

L4 -> L1 -> via IGP/LDP to ASBR1
L2 -> via IGP/LDP to ASBR2

- 3107 PIC possible at RRs
- 3107 Load balancing possible at PE1 and at each RR

→Benefit of next hop self on 3107 iBGP sessions from RRs

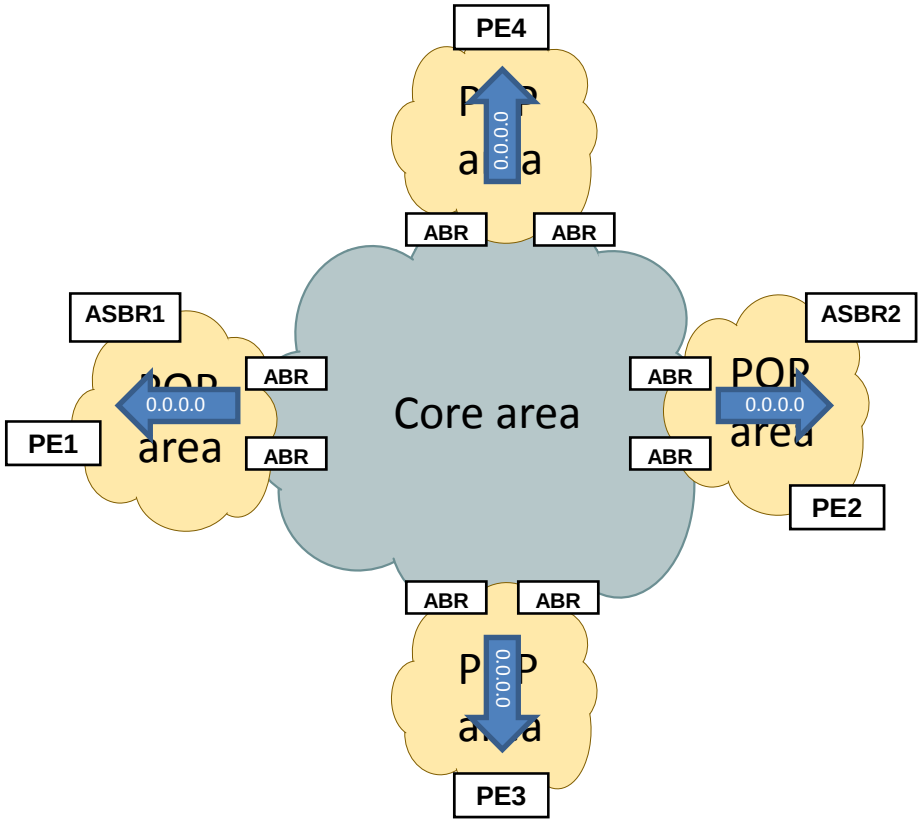
- Inter-AS option C, PE24 next hop: A,
- ABRs are RRs doing next hop self for 3107 on /32s
- LDP is local to each area delivers to RR1s, RR2s & ASBRs
- Label stack increased +1 due to introduced hierarchy
- NHS on the POP to core only needed in single POP exit architecture

Conclusions so far ...

- As we observe there is really close congruency between traditional ISP hierarchical network model and new MPLS „scalable” backbone model
- Your IPv4/v6/3107 RRs being in the data plane path do relax the need for add-paths deployment. It’s all about **path virtualization through hierarchy** rather then brute force network wide path’s flooding end to end.
- „Diverse BGP Path Distribution” can be used to provide more then best path advertisement in the case of hierarchical reflection instead of full meshing within each area. It also works well when your RRs are already control plane only devices.
- *Even for VPNs just consider those RRs to be in VPN data plane executing normal option B – especially with rt-constrain this may be of significant scaling plus*

And when Internet table size explodes ....

..... and some of your edge devices can't fit everything into their FIBs any longer



<u>You have two choices:</u>

1. Enter a new PO with your vendor
2. Configure within each POP a default route to ABRs/RRs for edge FIB install

<u>Details:</u>

- Note that still you have much less paths on the edge then in the flat design
- You still advertise all best nets in the control plane from ABRs/RRs to POPs
- PEs serving stub/domestic customers with smaller FIBs can continue to operate just fine and in FIB they just point to ABRs/RRs
- In flat design comparable solution is very challanging to deploy

That is nothing else then Paul's Francis special case of Virtual Aggregation proposal as described in **FIB Suppression with Virtual Aggregation** - **draft-ietf-grow-va-01.txt**

Final conclusions ....

- Add-Paths proposal or any other proposal is just a choice of technology to flood more BGP paths around in the network.
- The presentation was not about that this particular way of flooding is wrong - add-paths semantics are IMHO just fine – well diverse path idea seems much easier and does not require upgrade of all BGP speakers in your network.
- **The talk is about BGP path flooding to the edges being itself a questionable idea regardless of the encoding used.**
- Any iBGP paths flooding could be used between RRs if needed to establish RR hierarchy – no objections there. Similary any eBGP path flooding could be employed in IX route servers environments when required.
- While the research communities are scratching their heads on how much to flood ie how many paths to distribute to the edges of the network ... while being an excellent in their research findings – the most important point is missed - that this applies only to network architectures which removed RRs from data path.
- **By proper hierarchical architecture of the network - engineering teams are able to aggregate/virtualize BGP paths at the price of an additional line rate lookup.**
- **It also protects networks from exploding their amount of BGP state PEs/ASBRs need to keep in the network.**

Acknowledgement ...

My special thanks should be expressed to the below
Individuals for their inspiration and review of this work:

Prof Lixia Zhang
Randy Bush
Juan Alcaide
Christian Cassar
Cristel Pelsser
Laurent Vanbever
Pedro Marques
Keyur Patel
Rex Fernando

Also many thx to a few of anonymous reviewers who due to political reasons prefered not to expose their current names/associations.

References ...

- **Advertisement of Multiple Paths in BGP**
  draft-walton-bgp-add-paths-06

- **Advertisement of the best external route in BGP**
  draft-ietf-idr-best-external-01

- **Fast Connectivity Restoration Using BGP Add-path**
  draft-pmohapat-idr-fast-conn-restore-00

- **Analysis of paths selection modes for Add-Paths**
  draft-vvds-add-paths-analysis-00

- **Distribution of diverse BGP paths**
  draft-raszuk-diverse-bgp-path-dist-01

- **FIB Suppression with Virtual Aggregation**
  draft-ietf-grow-va-01

Question’s are welcome ...

... both on-line as well as off-line

raszuk@cisco.com

ftp://ftpeng.cisco.com/raszuk/addpaths/