BGP Optimal Route Reflection (BGP-ORR)

draft-raszuk-bgp-optimal-route-reflection-00

Robert Raszuk, Chris Cassar, Eric Aman, Bruno Decraene, Ilya Varlashkin

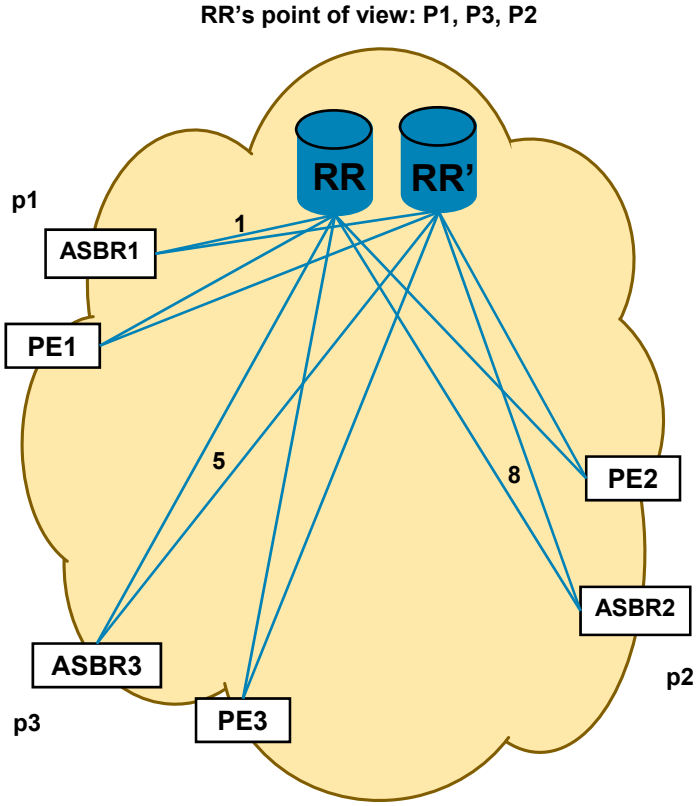
IETF 79, November 2010, Beijing, China

Agenda

- Problem statement
- Solutions
  - Independent IGP metric calculations
  - Next Hop Information Base SAFI (coming in -01)
  - Angular metric approximation
  - Other alternatives
- Flexible logical RR placement/relocation (co-idea)

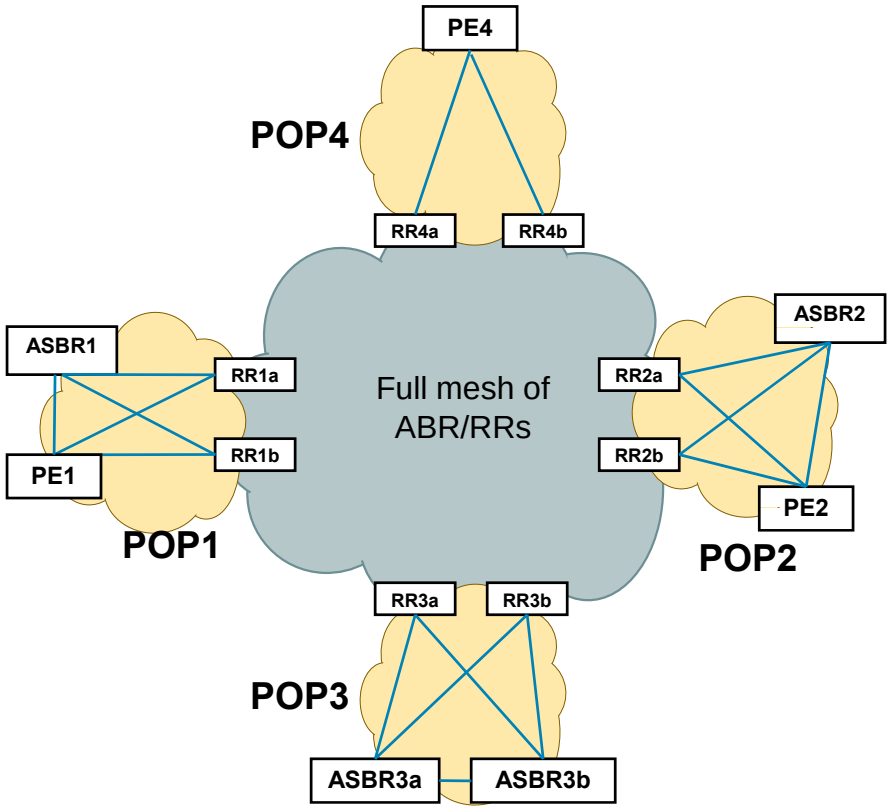
Problem statement

- RRs as control plane only platforms – departure from classic POP to Core location due to end to end encapsulation in networks and emerging Internet free core
- Suboptimal best/2nd best path selection for clients – difficult to ensure hot potato routing
- Position of control plane RRs should not play any role in path selection for clients.



- RRs select p1, p3, p2
- Clients get p1
- PE2 and PE3 exit by ASBR1

Not a goal to modify traditional RR placements - if it ain't broke, don't fix it



- If RRs are on the topological paths between clients and next hops
- If RRs are on the POP to core boundaries in hierarchical IGP model
- And if this design meets your objectives

➔ No need to break it.

Solution

- To calculate customized bgp best path for a given client or group of clients.
- No changes to BGP best path algorithm required .. Instead when we compare IGP metric to next hop for each client or group of clients this „metric” parameter will be different.
- What might such a metric represent ?
  - Could be just an IGP distance between client and next hop
  - Could be client to next hop propagation delay or min link bandwith *
  - Could be per client local exit preference via given next hop *
  - Could be any combination of the above .. Up to operator’s discretion *

  Note ... (*) can be used when edge to edge encapsulation is in place.

Solution

- In all cases network policies like local preference or MED are honored as they are compared before IGP metric to next hop
- This work is applicable to best path propagtion alone, propagation of diverse-path (2nd best) as well as add-paths N option (where N < ALL PATHs).
- The real question stands – How do we find out the right metric value ?
  - Option A ➔ Link state IGPs and SPT remote computation
  - Option B ➔ Next Hop Information Base ()
  - Option C ➔ Angular position approximation

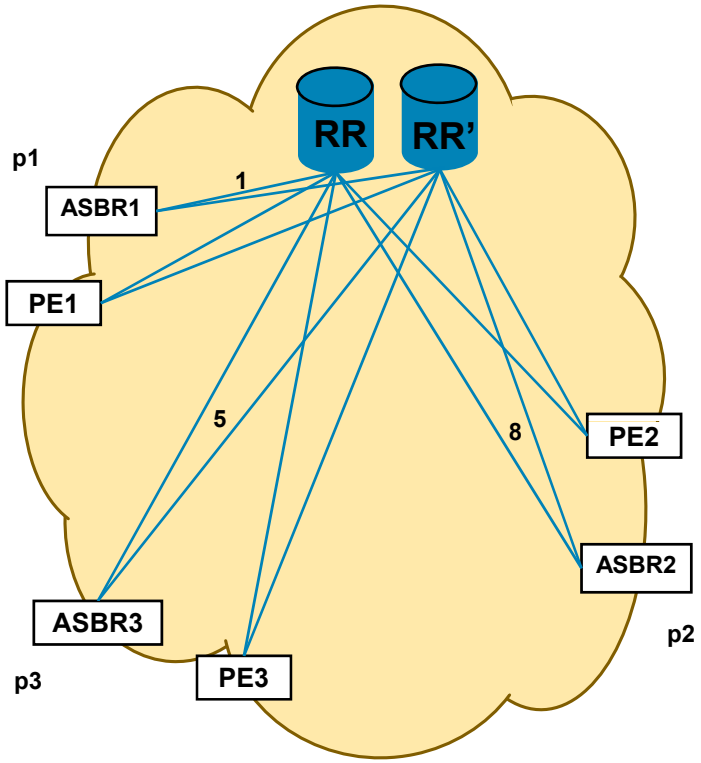
Option A - Link State IGPs remote computation

- In link state IGP, network topology is visible in the given scope of flooding/area by all participating nodes
- RR node can compute its own SPF as well as compute SPF pretending it is some other node
- Technology already used in LFA computation
- RR will position itself as IBGP client and run Dijikstra to get distance of client to every available next hop. Alternatively, as proposed by Aleksi Suhonen, new algorithms can optimize this by calculation of any to any distances in one run.
- Co-located clients (same POP) can be grouped ...
  .... calculation can be reduced to one per update group

Option A - Link State IGPs remote computation

RR's point of view: P1, P3, P2

PE1 point of view: P1, P3, P2
PE2 point of view: P2, P3, P1
PE3 point of view: P3, P2, P1

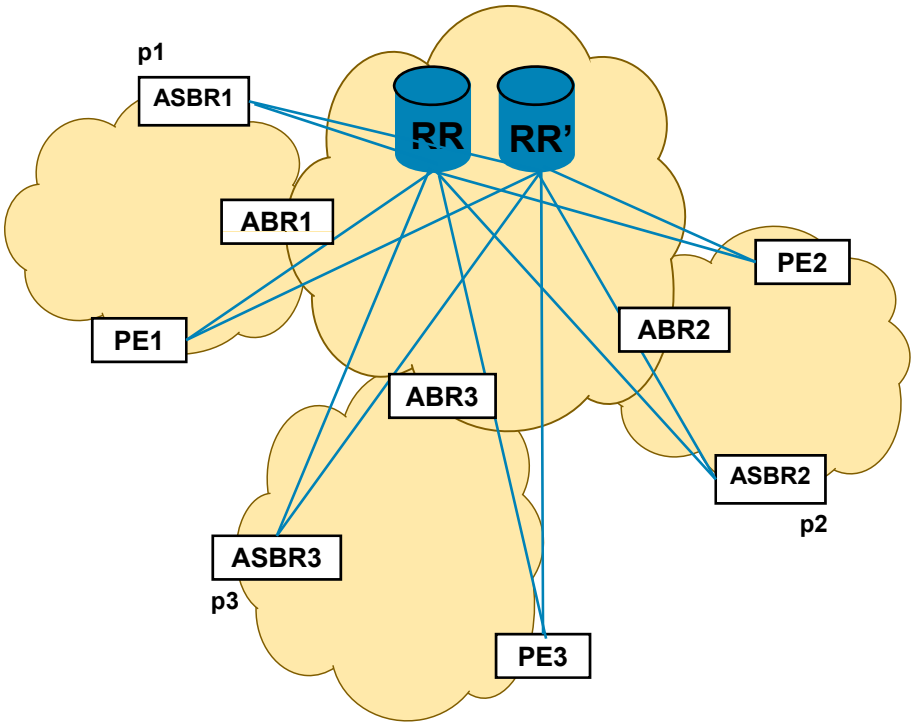


- In flat IGP topology all nodes are visible
- Each node will get an optimal path from it's point of view
- Recalculation during metric changes only when above threshold
- The same applies to areas in hierarchical IGP design where RRs are in each area

Option A - Link State IGPs remote computation

RR's point of view: P1, P3, P2

ABR1 (PE1) point of view: P1, P3, P2
ABR2 (PE2) point of view: P2, P3, P1
ABR3 (PE3) point of view: P3, P2, P1

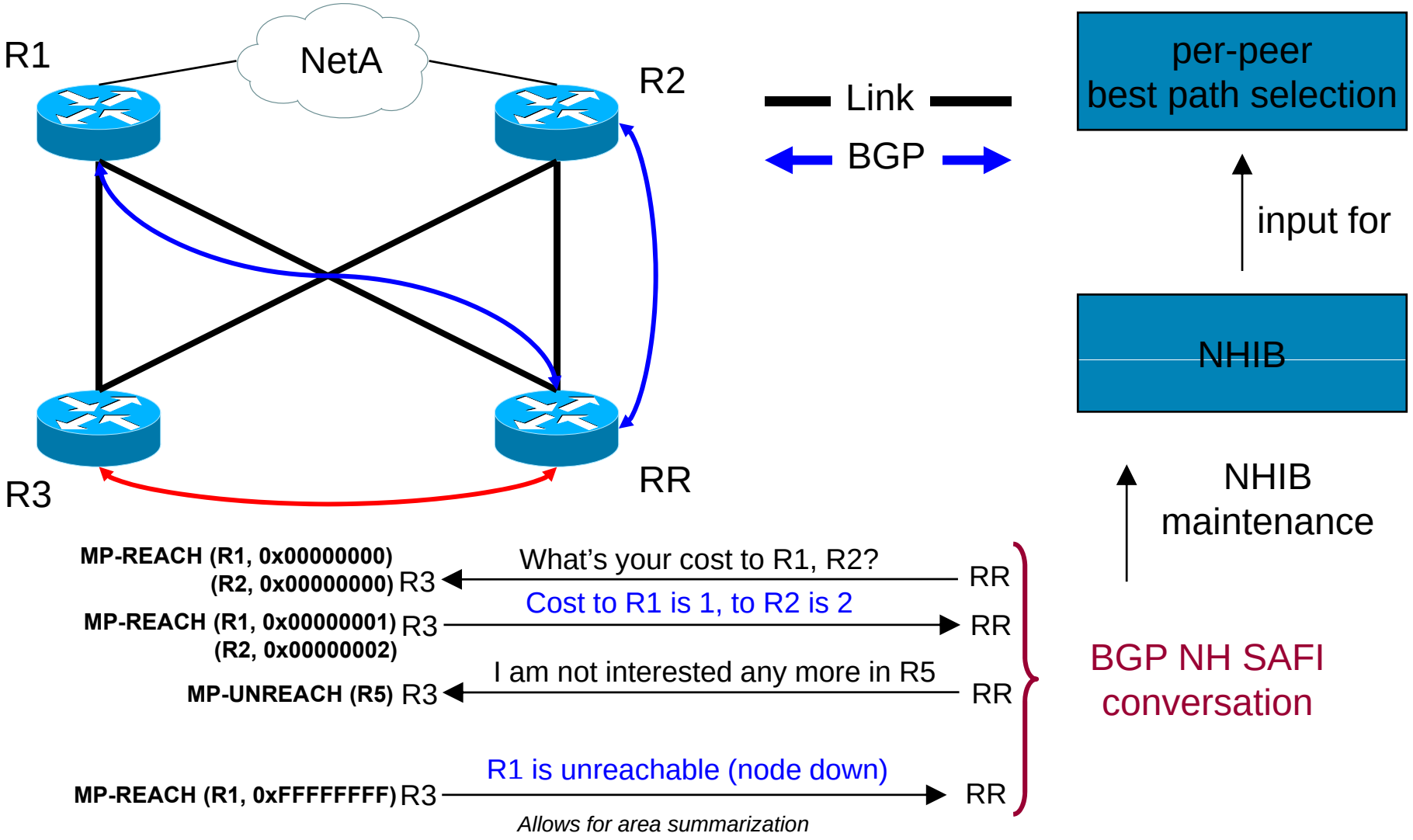


- In IGP hierarchy, centralized RRs can go as far as ABRs
- Metric to next hops will be visible from each ABR point of view either via summary LSAs OSPF or by route leaking (required for end to end LSPs).
- Precision limited to remote area scope (no visibility into intra area topology)

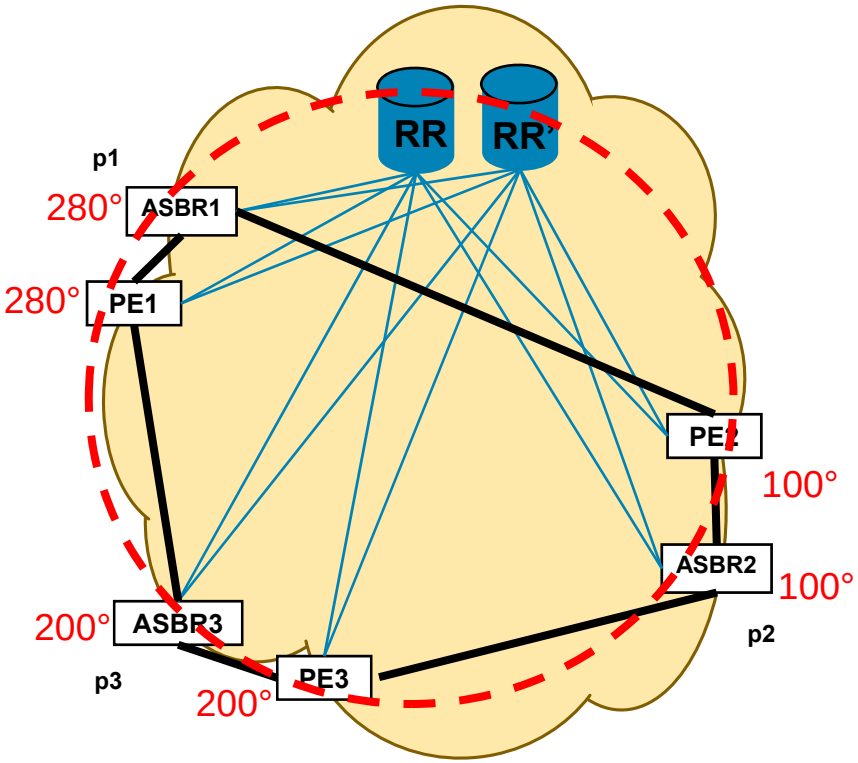
Option B - Next-Hop Information Base + NH SAFI

- Cost to arbitrary Next-Hop from arbitrary router (not necessarily local)
- To be used for BGP best path selection from arbitrary router perspective
- Content can be populated by different methods
- New SAFI facilitates NHIB population via BGP
  - learn Next-Hop cost where IGP has no visibility
- Applicable to both IPv4 and IPv6 (AFI=1 or AFI=2)
- Query/response operations
  - asynchronous (ask when need, inform when have something to say)
  - utilises existing BGP Attribute 14, 15
  - NLRI format (Next Hop, cost to reach next hop)
  - Special cost: 0x00000000 query, 0xFFFFFFFF unreachable

Option B – Using Next Hop SAFI



Option C – Angular position approximation



- Allows to statically map position of RR clients onto a virtual circle/elipse
- Works in specific topologies
- Requires encapsulation edge to edge
- Same metric allocated to co-located nodes (example in a given POP)

P1 – 280, P2 – 100, P3 – 200

PE1 (280) point of view: P1 (Δ=0), P3 (Δ=80), P2 (Δ=180)
PE2 (100) point of view: P2 (Δ=0), P3 (Δ=100), P1 (Δ=180)
PE3 (200) point of view: P3 (Δ=0), P1 (Δ=80), P2 (Δ=100)

Flexible logical RR placement/relocation (co-idea)

- In Option A we observed that within flooding scope of the IGP boundary each node has full visibility of all other nodes in such area.
- That also allows for permanent or temporary logical RR placement at any node of the area without physical connectivity changes
- This can be done globally for an entire RR or per each update group of the RR
- Turns out to be useful in some topologies
- Original feedback also indicates that this is useful for node maintenance without any risk of the BGP best path selection changes in best paths.

Conclusions

- This proposal attempts to ease introduction of control plane/centralized Route Reflectors
- It enables operators to manage their best paths selection policy within the AS beyond the traditional rules
- It opens new possibility for logical route reflector placement in an arbitrary network location without need to physically extend the connectivty to particular point
- Enables easier migration towards Internet routes free core without loosing ability to provide strict hot potato routing.