draft-vroonen-idr-bgp-bestpath-nh-selection-00

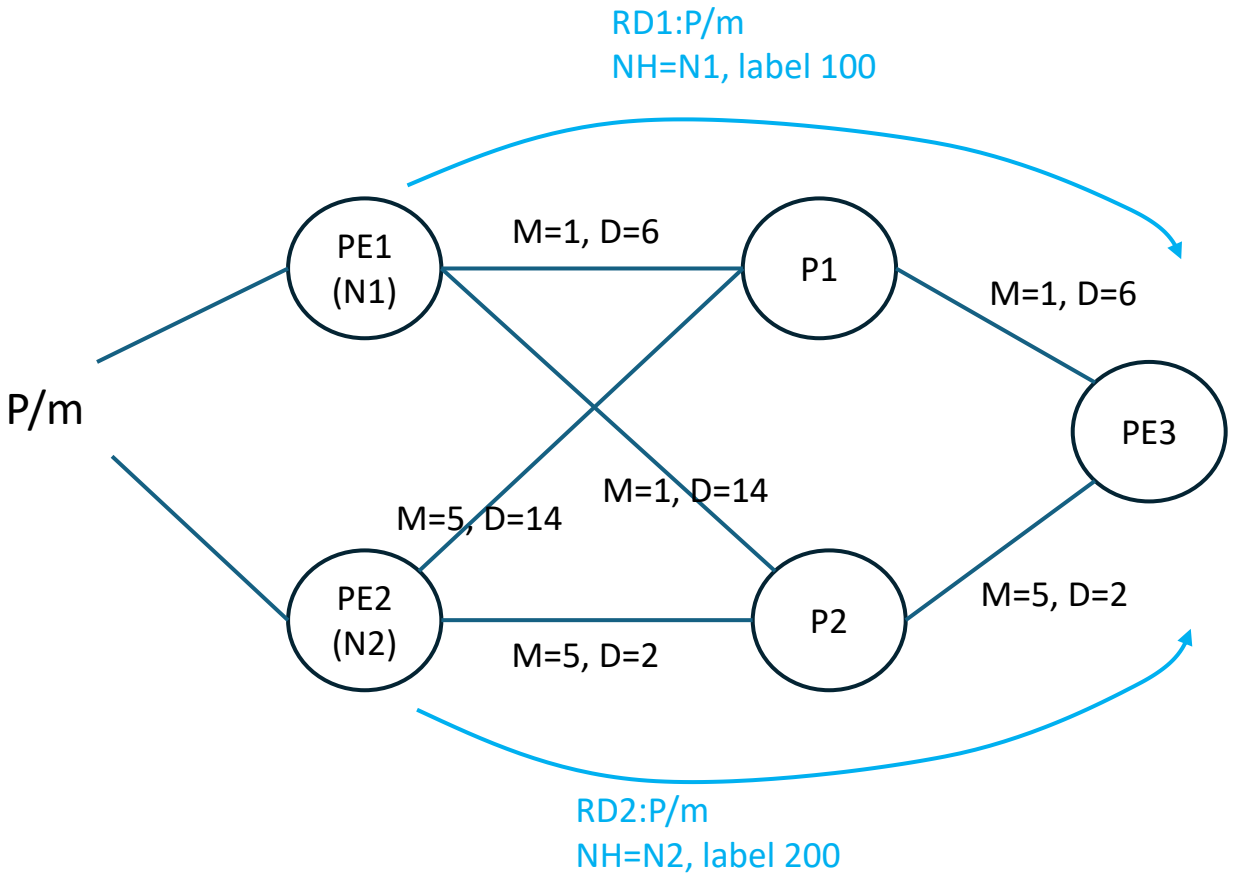
O. Vroonen, Cisco

S. Litkowski, Cisco

Problem statement

- RFC4271 is considering that NEXT_HOP is actually used for forwarding
  - Resolvability condition
  - iBGP paths tiebreaker based on metric (taken from RIB) to find closest exit
- With new dataplane coming in (tunnels…), this assumption may not be true anymore with side effects:
  - Traffic drop
  - Suboptimal routing
- Most of the implementations already address these problems, the goal of the draft is to standardize these behaviors

Example of issues: traffic drop



PE3 IPv4 RIB:

N1 via P1, metric=2
N2 via P1, metric=6

PE3 "MPLS tunnel RIB":

N2 via P1, metric=6, label {2100}
No LSP to N1

PE3 BGP VRF table:

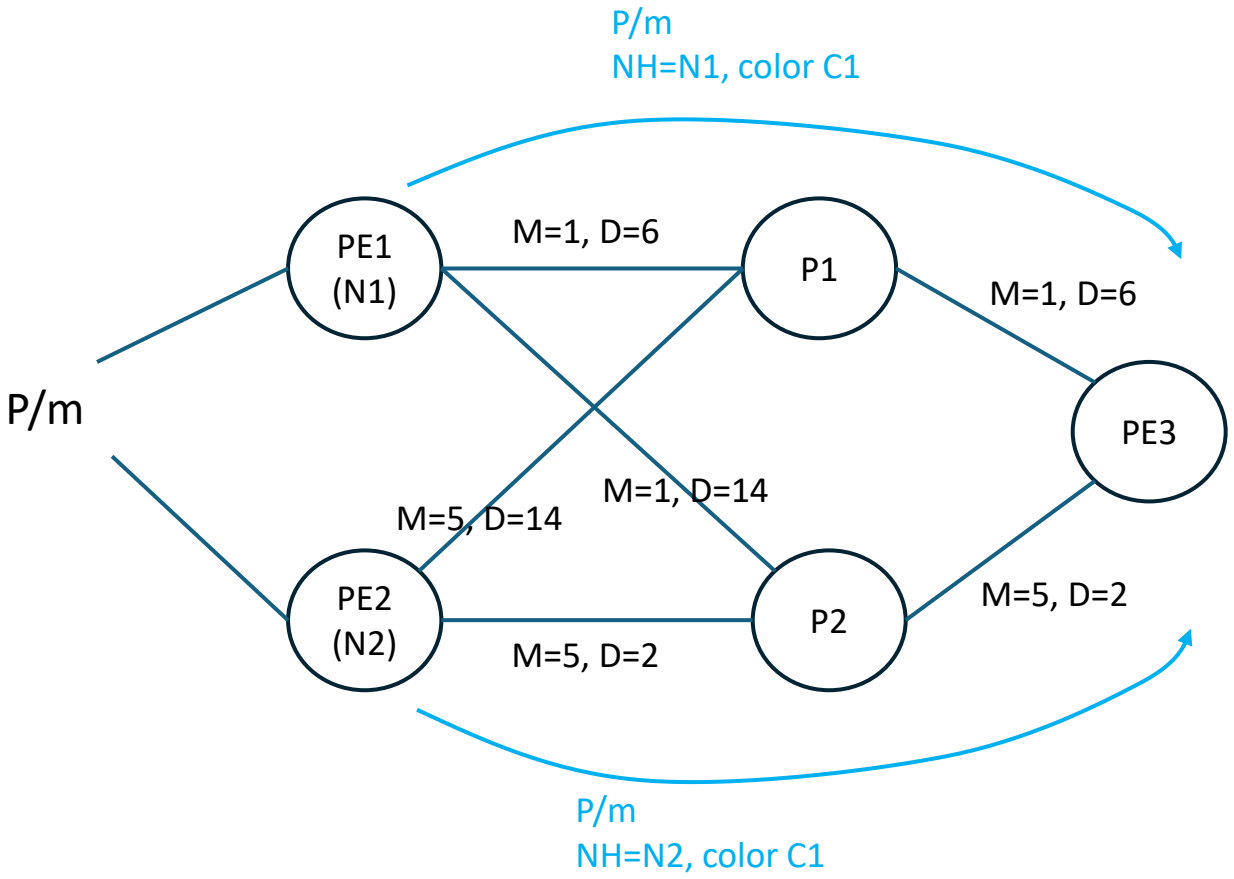
P/m:
via N1 (cost 2), label 100 => BEST

via N2 (cost 6), label 200

PE3 will drop traffic to P/m as there is no transport tunnel to reach N1

Example of issues: suboptimal path

P/m prefix uses color C1 that is associated with low-delay routing



Prefix	Next hop	Cost
N1	P1	2
N2	P1	6

Table 1: IPv4 routing table of PE3

SR Policy	Next hop	Cost	State
(C1, N1)	P1	12	Up
(C1, N2)	P2	4	Up

Table 2: SR Policies of PE3

PE3 BGP table:

```
P/m:
  via N1 (cost 2), color C1 => BEST

  via N2 (cost 6), color C1
```

PE3 picks up the path from PE1 as best while it has the highest delay

This is a generic issue

- Not only MPLS related
  - SRv6 enhanced route resolvability condition in RFC9252
  - RFC9012 enhanced routes resolvability condition when tunnel encaps is used
  - MPLS resolvability condition tried to be address in draft-ietf-idr-bgp-best-path-selection-criteria
- Resolvability is partially addressed
- No draft talks about path suboptimality

Forwarding address and context

- In the BGP path, we need to identify the forwarding address:
  - This is the address that is actually used to forward packets (it may not be the NEXT_HOP)
  - For tunnel encaps, this is the tunnel endpoint
- A forwarding address has a forwarding context defining the characteristics of the forwarding path to be used:
  - For tunnel encaps endpoint, forwarding context is tunnel type and its characteristics
- The forwarding context defines a set of resolution constraints:
  - E.g.: table to be used for lookup…
- Resolution constraints may be tuned by configuration

Modify route resolvability condition

- Route resolvability check for the NEXT_HOP must always be done

- If the forwarding address is not the NEXT_HOP, route resolvability check should be done for it by applying the resolution constraints:
  - Lookup the forwarding address in a specific table
  - Verify that there is an entry that matches the required path characteristics

Modify internal cost computation

1. Retrieve the cost associated with the route to the forwarding address (lookup is done a specific table as per resolution constraints)

2. Costs may not be comparable depending on the forwarding address and its context:

```
P/m
    Path1: NH=10.0.0.1 Color: 200
    Path2: NH=2001::2 SRv6-SID: cafe:0:2:e002::
    Path3: NH=10.0.0.3 Tunnel-encap(L2TPv3, endpoint: 10.0.0.3)
    Path4: NH=2001::4 SRv6-SID: cafe:0:4:e002::
```

Delay based cost

BW based cost

No cost

IGP metric based cost

3. As RIB do, we need to introduce a notion of preference to make paths comparable

```
P/m
    Path1: NH=10.0.0.1 Color: 200,
           preference 100 (from table), cost 1001
    Path2: NH=2001::2 SRv6-SID: cafe:0:2:e002::,
           preference 10 (from BGP), cost 12
    Path3: NH=10.0.0.3
           Tunnel-encap(L2TPv3, endpoint: 10.0.0.3, sessID: 1),
           preference 1000 (from BGP), cost max
    Path4: NH=2001::4 SRv6-SID: cafe:0:4:e002::,
           preference 10 (from BGP), cost 14
```

Lowest preference wins

Modify internal cost computation

- When comparing internal paths:
  - Remove from consideration any routes with a highest preference value:
    - Preference is retrieved from table lookup or set by BGP (default or user config)
  - For remaining paths, we pick the route with lowest cost
- This should be turned on by configuration

NH and Forwarding address tracking

- NH / Forwarding address resolvability condition and internal cost computation must not be done only when the BGP route is learned

- Any time the cost is changing or the address becomes unreachable/reachable, BGP should reevaluate the best paths.

Next steps

- The draft clarifies behaviors that are already implemented by multiple vendors to solve traffic drop and suboptimality

- We welcome the feedbacks to make this work progressing