Transitive BGP Graceful Restart (draft-zhang-idr-transitive-gr-01)

Haifeng Zhang, Alvaro Retana

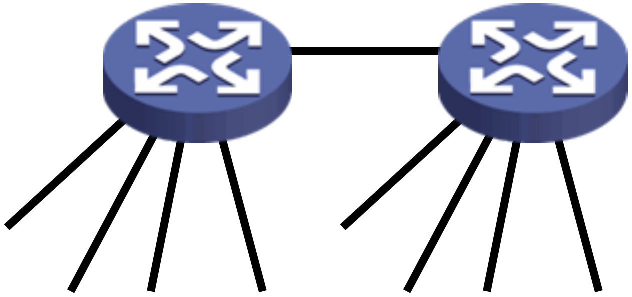
zhanghf@h3c.com,
aretana@cisco.com

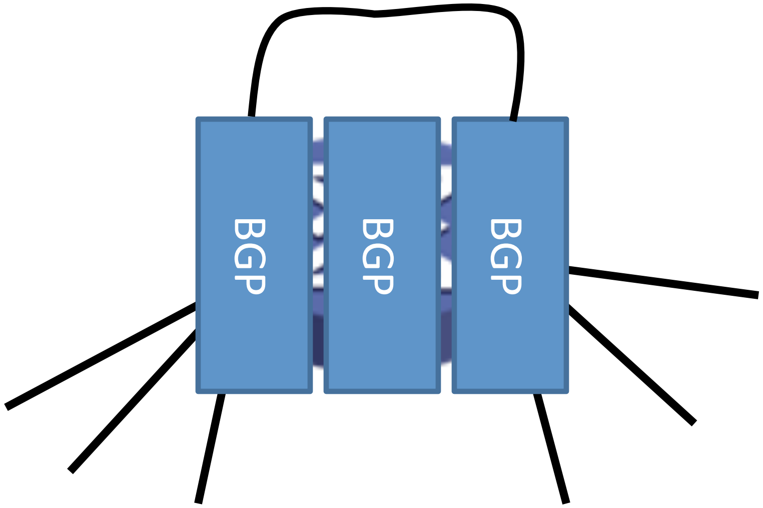
Background

- Forwarding in a Network is Transitive
  - routing protocols independently calculate paths based on consistent information distributed throughout the network
  - adjacencies confirm the transitive relationship between neighbors
  - graceful restart verifies transitivity by signaling that the forwarding state has been preserved through a restart…even if routing is not active

What problem are we trying to solve?

- Two (or more) restarting routers next to each other.





Transitive BGP Graceful Restart

- Objective:
  - **Reduce the impact of multiple restarting routers on the data plane.**
- Problem:
  - **BGP Graceful Restart (rfc4724) defines a mechanism that clears the forwarding table for any restarting peer, even if the forwarding state has been maintained by both routers. The effect is route instability, traffic loss, etc.**

```
          NR1                R1               R2
            |   RIB+EOR      |                |
            |--------------->|                |

                 Best Path Selection

                  Clean RIB                   -- t1

            |   RIB+EOR      |                |
            |<---------------|      RIB       |
            | End GR         |<-------------->|

                    Update RIB               -- t2

NR - Non-Restarting Router;    R - Restarting Router
```

"...the forwarding state of the speaker MUST be updated and any previously marked stale information MUST be removed ." [rfc4724]

Transitive BGP Graceful Restart

- Objective:
  - **Reduce the impact of multiple restarting routers on the data plane.**
- Solution:
  - **Extend the operation of BGP Graceful Restart to allow a router to fully synchronize with other restarting peers before cleaning the forwarding table.**

```
        NR1             R1              R2
          |   RIB+EOR    |               |
          |------------->|               |

                 Best Path Selection
                                          -- t1
          |              |    RIB+EOR    |
          |              |<------------->|

                 Best Path Selection

                 Clean RIB                -- t3

          |   RIB+EOR    |               |
          |<-------------|               |-- t4
          | End GR       |               |

NR - Non-Restarting Router;    R - Restarting Router
```

“…the BGP speaker MAY advertise the Adj-RIB-Out to the remaining peers …and MAY wait for the corresponding End-of-RIB marker from the restarting ones.” [Proposed Addition.]

Not used Routing Information

- During the recovery period of multiple restarting routers, a BGP speaker may advertise routing information that is not being used at the time.
  - the forwarding state of the speakers remains unchanged (from that at the restart)

- There are 3 types of routes being advertised by the restarting router (all through its non-restarting peers):
  1. Routes that correspond to routes in the RIB. Represent the majority!
  2. Routes not previously in the RIB. (Installation will be delayed until t3.)
  3. Routes that in the RIB point to the other restarting router.

"…minimize the effect of routing flaps, it is noted that when a BGP Graceful Restart-capable router restarts…there is a potential for transient routing loops or blackholes in the network if routing information changes before the involved routers complete routing updates and convergence." [rfc4724]

Transitive GR across more than 2 Nodes

- To maintain the transitive property when more than two BGP speakers peering with each other restart...:

  “If a restarting BGP speaker has multiple restarting peers, sending the End-of-RIB marker SHOULD be delayed until all the markers from restating peers have been received. The BGP speaker with the lowest BGP Identifier on a given connection SHOULD send its End-of-RIB marker if the pair hasn't sent or received UPDATES for a locally configured time period (which should be significantly less than the Selection_Deferral_Timer).” [Proposed Addition]

Routers in a Line

```
NR1 ----- R1 ----- R2 ----- R3 ----- NR2
 |        |        |        |         |
 |RIB+EOR |        |        | RIB+EOR |
 |-------->|        |        |<---------|
 |        |RIB+EOR | RIB+EOR|         |
 |        |-------->|<--------|         |
 |        |        |        |         |
 |        |RIB+EOR | RIB+EOR|         |
 |        |<--------|-------->|         |
 |RIB+EOR |        |        | RIB+EOR |
 |<--------|        |        |-------->|
 |        |        |        |         |
```

More Routers in a Line

```
NR1 ----- R1 ----- R2 ----- R3 ----- R4 ----- NR2
 |        |        |        |         |         |
 |RIB+EOR |        |        |         | RIB+EOR |
 |------->|        |        |         |<--------|
 |        |RIB+EOR |        | RIB+EOR |         |
 |        |------->|        |<------- |         |
 |        |        |        |         |         |
 |        |        |   RIB  |         |         |
 |        |        |<------>|         |         |
 |        |        |   EOR  |         |         |
 |        |        |------->|         |         |
 |        |        |  EOR   |         |         |
 |        |        |<-------|         |         |
 |        | RIB+EOR|        |         |         |
 |        |<-------|        |         |         |
 |RIB+EOR |        |        | RIB+EOR |         |
 |<-------|        |        |-------->|         |
 |        |        |        |         | RIB+EOR |
 |        |        |        |         |-------->|
```

Routers in a Triangle

```
                   R2
               /          \
              /            \
             /              \
NR1 ----- R1 -------------- R3 ----- NR2
  |        |        |        |         |
  |RIB+EOR |        |        | RIB+EOR |
  |------->|        |        |<--------|
                   RIB
  |        |------->|        |         |
  |        |---------------->|         |
  |        |        |<-------|         |
  |        |<----------------|         |
  |        |        |        |         |
                  RIB
  |        |<-------|------->|         |

                  EOR
  |        |------->|        |         |
  |        |---------------->|         |
  |        |        |        |         |
                   EOR
  |        |<-------|------->|         |
  |        |        |<-------|         |
  |        |<----------------|         |
  |        |        |        |         |
  |RIB+EOR |        |        | RIB+EOR |
  |<-------|        |        |-------->|
  |        |        |        |         |
```

Next Steps

- Comments/Questions/Feedback
- Adopt as a WG Item
  - Updates rfc 4724
  - Standards Track