BGP Persistence

draft-uttaro-idr-bgp-persistence-00

James Uttaro	AT&T
Adam Simpson	Alcatel-Lucent
Rob Shakir	C&W
Clarence Filsfils	Cisco Systems
Pradosh Mohapatra	Cisco Systems
Bruno Decraene	France Telecom
John Scudder	Juniper Networks
Yakov Rekhter	Juniper Networks

Why?

- BGP built on the premises that forwarding plane share fate with BGP control plane / sessions.
- But for some (new) BGP applications, this coupling is less valid
  - e.g. L2 VPN auto-discovery, dedicated route reflectors in L3 VPN, BGP signaled multicast…
- People / business relies on network
  - less and less likely to accept failures for a long duration (hours).
  - The bigger, the more converged the network, the less acceptable.
- “Persistence” targets catastrophic BGP failures when both nominal & backup BGP sessions are down,
  - if it is felt that some (degraded) network is better than nothing.
- “Persistence” as a last resort safety net for BGP sessions.

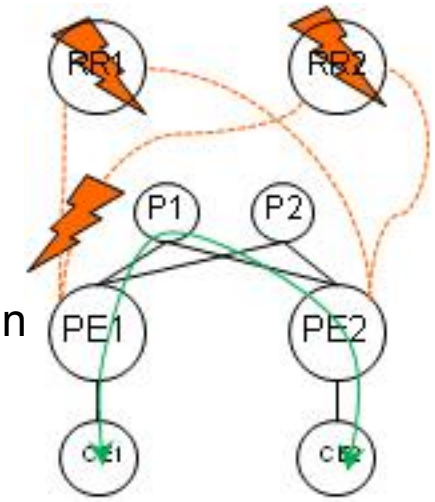
What? (1): BGP session failure

- Routes are kept for the duration of persist-timer.
  - could be hours or more.
- Routes are de-preferenced → prefer non-stale routes
  - And tagged with a “STALE” community to inform downstream BGP routers
  - → leads to BGP re-advertisement.
- BGP Next-Hop reachability is (still) checked.
- Some routes may be defined as non eligible for BGP persistence
  - Tagged by upstream BGP peers with community “DO_NOT_PERSIST”.

What? (2): BGP session re-establishment

- Stale routes are replaced by newly received routes.
- When EoR is received, remaining STALE routes are removed, best path computation performed and routes re-advertised.
  - additional local timer if EoR not received.
- If session fails again before EoR is received:
  - routes still marked as STALE are kept
  - all routes are marked as STALE (again)

Example of use cases



- Types of failures:
  - Double failure of both dedicated Route Reflector.
  - Failure of both iBGP client sessions between a PE an its 2 RRs.
- Type of networks:
  - VPLS/VPWS (L2 VPN)
    - BGP routes are not exchanged with customers and fairly static (provisioned).
  - L3 VPN
    - Note: dual attached customers would switch to backup path.
  - First IP node of residential customers
    - e.g. BRAS, BNG, IP MSAN
    - Customers known to be single attached and not moving → very static routes

Caveats

- If all routers experiencing the iBGP sessions failures are not persistent capable, different routers have different routing states.
  - Resulting effects are AFI/SAFI specific.
    - L2 VPN & L3 VPN cases are discussed in the draft
  - routes using tunnels to reach BGP Next-Hop are less affected (vs hop by hop routing).
  - Not new / specific to BGP persistence: idem for GR, gr-notification, optional-transitive.
- During the double iBGP failures, routing states are not updated anymore, especially dynamic states learn from others AS. Hence quality and consistency of routing is expected to degrade over time.
  - Tradeoff to consider when setting max value of the persistence timer. Application and AS specific.
  - Particularly important for L3 VPN as VPN isolation are based consistent VPN labels across all PEs. Discussed in the security consideration.

Why not Graceful Restart (GR)?

- Different assumption on the nominal path
  - GR: assumes nominal path is still usable (disregard the failure) → strong assumption
  - Persistence: assumes nominal path is less trusted and should only be used as last resort → lighter assumption
- Different routing result
  - GR: keep nominal path → no churn, short duration only (strong assumption may be wrong over time)
  - Persistence: switch to backup path → churn, longer duration (lighter assumption easier to assume over time)
- GR as some limitations with regards to the persistence requirements:
  - limited duration (68 min.), does not address consecutive session restart or BGP notifications
    - draft-keyur-idr-enhanced-gr-00 would address the latter

Next Steps

- Still work in progress
  - Feedback and comments welcomed.
- Next version (-01) will address comments received:
  - Interaction between Graceful Restart & Persistence
  - Incremental deployment
    - Intra AS & between ASes
  - Discuss, in the security consideration, the use of “GR mechanism for BGP with MPLS” (RFC 4781).

thank you