Prefix Unreachable Announcement

draft-wang-lsr-prefix-unreachable-annoucement

A. Wang (China Telecom)
G. Mishra (Verizon)
Z. Hu (Huawei Technologies)
Y. Xio (Huawei Technologies

IETF-110, March 2021

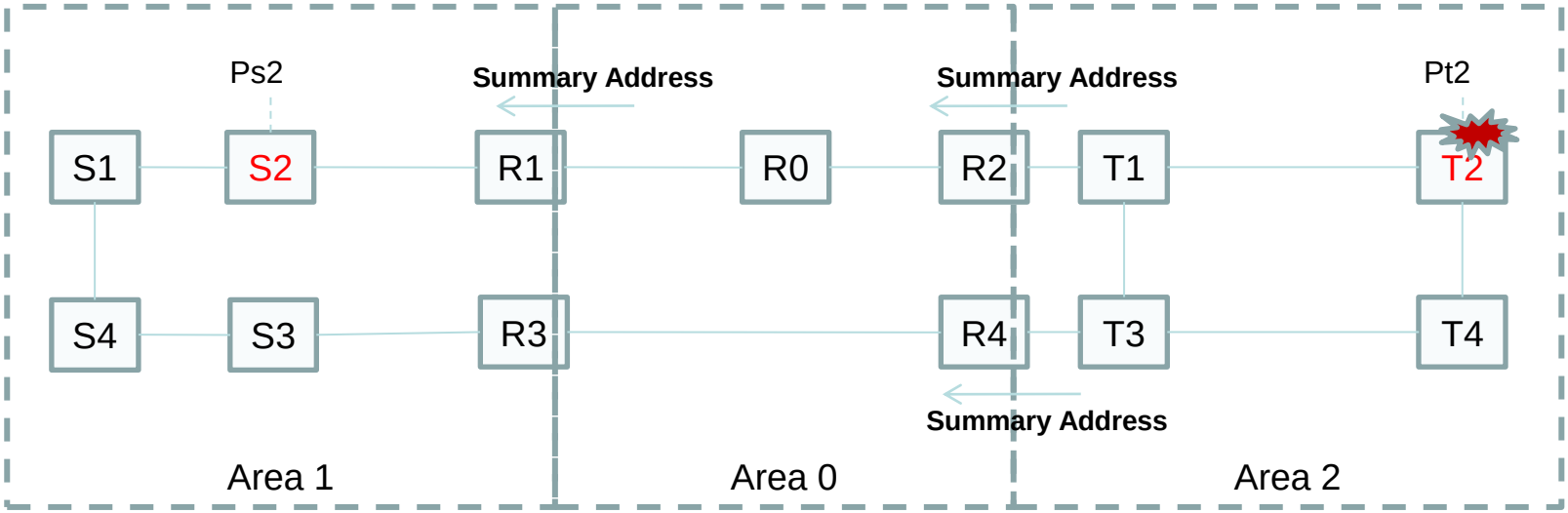
Motivation & Problem Statement

- The motivation behind this draft is based on either MPLS “Exact Match” host route FEC binding, or SRv6 BGP Service overlay using traditional unicast routing (uRIB) Longest Prefix Match (LPM) forwarding plane where the IGP domain has been carved up into OSPF or ISIS areas & summarization is utilized..
- Summarization of Inter-Area types routes propagated into the backbone area for flood reduction are made up of component prefixes. It is these component prefixes that the “Prefix Unreachability Announcement” tracks to ensure traffic is not “black hole” sink routed due to a PE or ABR failure.
- This draft provides a control plane signaling mechanism to detect the component prefix failures that are part of a summary prefix to force immediate control plane convergence to an alternate path.

Updated Contents

- Updated Scenarios
- Updated Action based PUA message
- Further Action

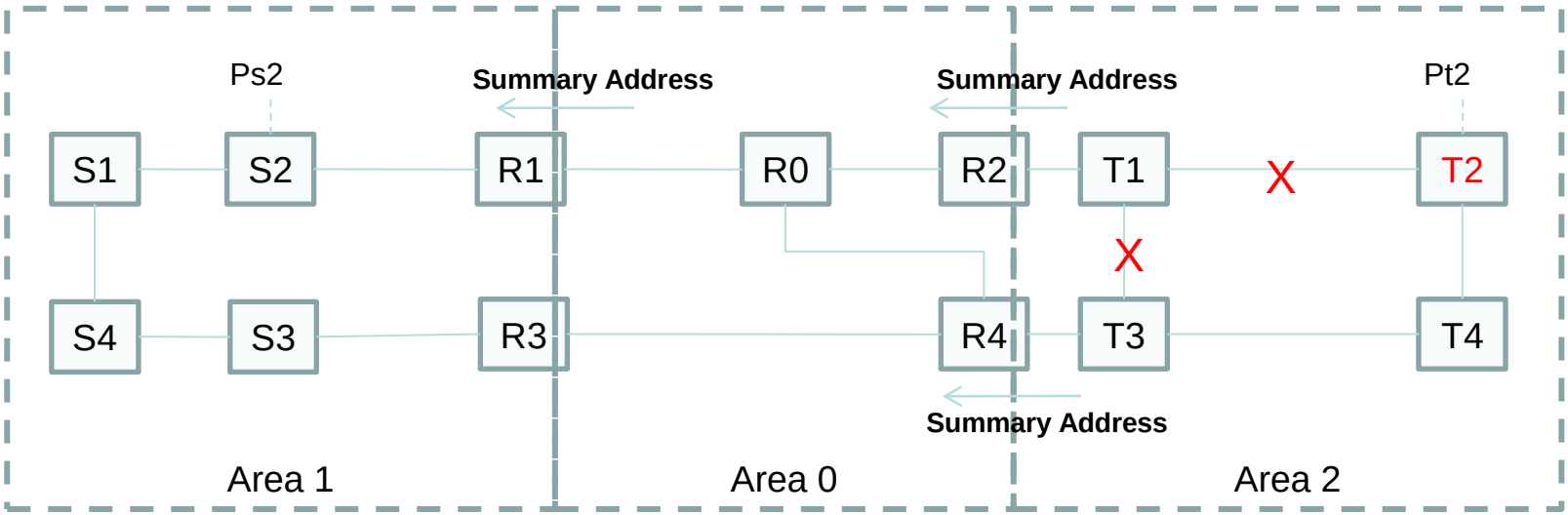
Updated Scenarios(1/2)



OSPF Prefix Unreachable Scenario (Node Failure)

- ✓ ABR R2/R4 do the summary action, send only the summary address to Area 0, Area 1.
- ✓ S2 has BGP session with T2, which provides the control connection for VPN services between them.
- ✓ When node T2 is failure, the summary address is still advertised and so the LSP is still built to R2. Black Hole LSP dead end on R2.
- ✓ S2 doesn't know that T2 down.
- ✓ Service Traffic will be breakout during this duration of T2 down.

Updated Scenarios(2/2)



OSPF Prefix Unreachable Scenario (Link Failure)

- ✓ ABR R2/R4 do the summary action, send only the summary address to Area 0, Area 1.
- ✓ S2 has BGP session with T2, which provides the control connection for VPN services between them.
- ✓ When link between T1/T2 and T1/T3 are broken, R2 can't reach T2, but it still announces the summary address. R0 still takes R2 as the next hop to T2 and LSP is still built to R2. Black hole LSP is dead end on R2.
- ✓ Traffic to T2 will be broken at ABR R2 until T2 is restored.

PUA Mechanism

- Upon receiving the node/link failure information, which prefix is within the range of advertised summary address, the ABR or L1/L2 border router will:
  - Generate one new summary address, with the failure prefix associated, but set its originator information to NULL.
  - For ISIS, we use “IPv4/IPv6 Source Router ID” sub-TLV, which is defined in RFC 7794
  - For OSPF, we use “Prefix Originator Sub-TLV”, which is defined in draft-ietf-lsr-ospf-prefix-originator
  - Such summary message will be flooded across the boundary as normal OSPF/IS-IS procedures.

Updated Action based on PUA message

- For scenario 1(node failure)
  - When node within one area receives the PUA message from All of its ABRs, it will trigger the switchover of the control plane, which is run on top of it.
  - For scenario 1, the BGP session between S2/T2 will be notified, S2 can then begin the BGP session switchover immediately.
- For scenario 2(link failure/network partition)
  - When only some of the ABRs can't reach the failure prefix, the ABRs that can reach this prefix should advertise one specific route to this PUA prefix.
  - Same procedures as RIFT.

BGP next Hop MPLS / SR-MPLS / SRv6 Use Case

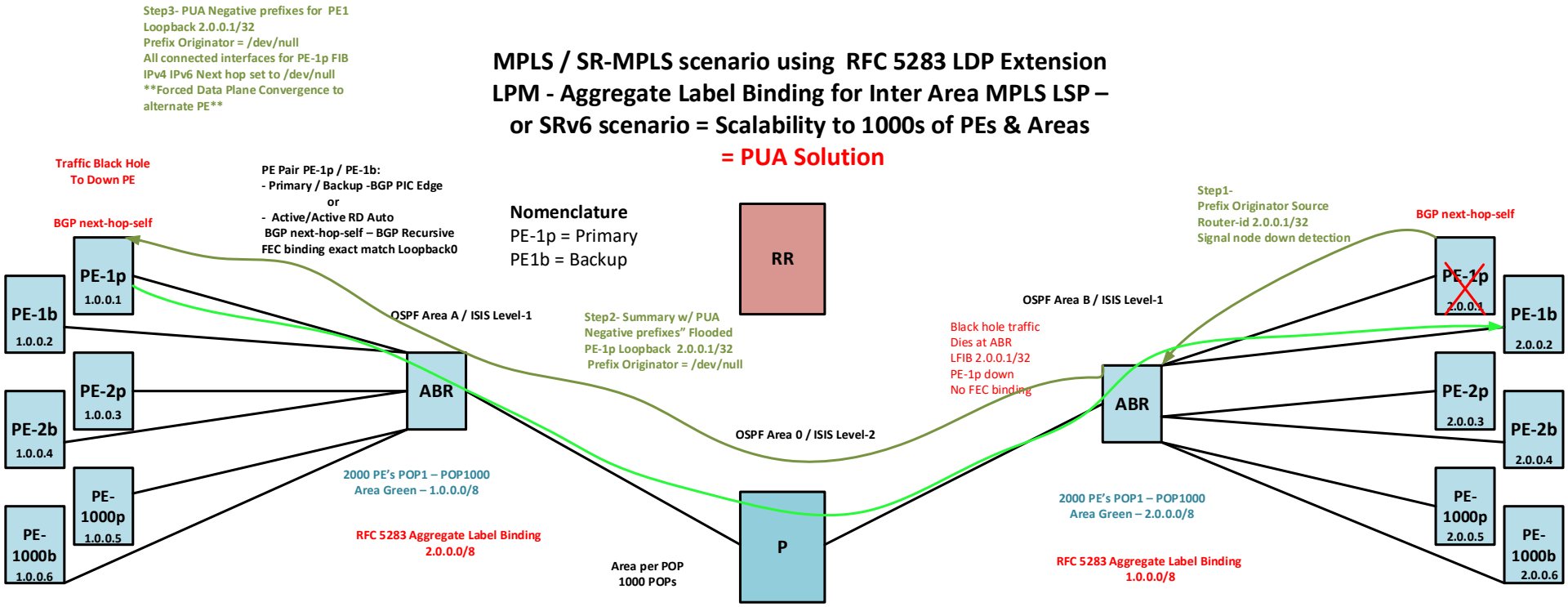
Use case BGP Next Hop Data Plane Convergence

- In an MPLS or SR-MPLS service provider core, scalability has been a concern for operators which have split up the IGP domain into multiple areas to avoid flooding of BGP next hop reachability throughout the domain. RFC 5283 defined LDP extension for inter-area LSP aggregation. MPLS FEC binding for LSP instantiation is based on egress PE "exact match" of /32 host route Looback0. RFC 5283 LDP inter-area extension provides the ability to LPM(Longest Prefix Match), so now the RIB match can now be a summary match and not an "exact match" of /32 host route of the egress PE for an inter-area LSP to be instantiated. The caveat related to this feature that has prevented operators from using the RFC 5283 LDP inter-area extension concept is that when the component prefixes are now "hidden" in the summary prefix, and thus the visibility of the BGP next-hop attribute is now lost. Thus in a case where a PE is down, and the RFC 5283 LDP inter-area extension LPM summary is used to build the LSP inter-area, now the LSP remains partially established black hole on the ABR performing the summarization. This MAJOR gap with RFC 5283 inter-area extension forces operators into a workaround of having to flood the BGP next-hop domain wide. In a small network this is fine, however if you have 1000s PEs and many areas, the domain wide flooding can be painful for operators as far as resource usage memory consumption and computational requirements for RIB / FIB / LFIB label binding control plane state. The ramifications of domain wide flooding of host routes is described in detail in RFC 5302 "Domain wide prefix distribution with 2 level ISIS" section 1.2 Scalability. As SRv6 utilizes LPM (Longest Prefix Match), this problem exists as well with SRv6 when IGP domain is broken up into areas and summarization is utilized.

Solution to BGP Next Hop Control Plane Convergence

- PUA is now able to provide the "Negative prefix" PUA component now flooded across the backbone to the other areas along with the summary prefix with Next hop set to Null0, which is now immediately programmed into the FIB control plane used by the forwarding plane. MPLS LSP "Exact Match" or SRv6 LPM match over failover path can now be establish to the alternate egress PE. No disruption in traffic or loss of connectivity results from PUA. Further optimizations such as LFA & BFD can be done to make the convergence hitless. The PUA solution applies to MPLS or SR-MPLS where LDP inter-area extension is utilized for LPM aggregate FEC, as well a SRv6 IPv6 control plane LPM match summarization of BGP next hop.

Applied Scenarios



Implementation Consideration

- Considering the balance of reachable information and unreachable information announcement capabilities, the implementation of this mechanism should set one MAX_Address_Announcement (MAA) threshold to control the advertisement of PUA and summary address.
  - If the number of unreachable prefixes is less than MAA, the ABR should advertise the summary address and the PUA.
  - If the number of reachable address is less than MAA, the ABR should advertise the detail reachable address only.
  - If the number of reachable prefixes and unreachable prefixes exceeds MAA, then advertises the summary address with MAX metric.

Further Action

- Comments?
- Adopt as WG document?

wangaj3@chinatelecom.cn
gyan.s.mishra@verizon.com
huzhibo@huawei.com
xiaoyaqun@Huawei.com

IETF110@Online(Virtual)