Scaling Seamless MPLS networks using “MPLS Namespaces”


IETF 110

Kaliraj Vairavakkalai

Juniper Networks
March 10, 2021
Agenda

• Problem statement, and scope
  • Scaling problem.
  • Convergence problem.

• Describe solution
  • MPLS namespaces, signaled by BGP.
  • A new inter-AS option. (option BC+)

• Discuss Advantages
  • Scaling advantages
  • Convergence advantages
  • Brownfield deployment advantages.
Problem statement, and scope

• Problem Scenario
  • Seamless MPLS network (option-C inter-AS) may have huge number of nodes. Each node lo0 could be a Protocol Nexthop (PNH) in BGP routes.
  • MPLS labeled-routes and nexthops resources consumed on BNs is proportional to total number of nodes in the network
  • MPLS nexthops consumed on ingress-SNs is proportional to number of PNHs received, and ECMP across them.

• Goal
  • Reduce the number of PNHs visible to the BNs in the network.
  • Improve scaling properties of BNs and SNs in the network by doing so.
  • Improve convergence properties too of the network, by detecting and repairing traffic much closer to failure.

• Following two slides describe with example topology, the scale problem and convergence problem.
Problem: Scaling on all BNs is a Function(num PEs) times num-ECMP.
Problem: PE1 failure needs to propagate until PE2 for PIC convergence.
Solution – MPLS-namespaces signaled by BGP

• As number of PNHs grows in the network, some abstraction is needed across regions.
• “MPLS namespaces” can abstract PE lo0s (PNHs) in a region from rest of the regions. Thus providing this abstraction. All PEs in a region are represented by a “Context PNH” as the nexthop in BGP updates.
• “BGP MPLS namespaces” is a new BGP family (RFC4364 style) to populate upstream allocated labels in MPLS Context tables.
• This family needs to be negotiated only between Service-RR and BNs in the same region. It will mirror “Private labels” to the BN’s context-tables.
• Other regions will have route, nexthop state for one Context-PNH, instead of “N” PE-lo0 PNHs. Thus saving forwarding resources on all BNs and ingress-SNs
• We will see how forwarding to the N PEs works, while being abstracted by this one Context-PNH.
• PE advertised service-labels will be re-written with mpls-namespace private labels, and advertised with the context-PNH. Similar to L3VPN option-B. But without having service-routes at the BN.
• So in essence, this is a new inter-as option. Which has the benefits of option-B, option-C, and some more.
• Let us see how it works.
Solution: Scaling on all BNs becomes Function(num-Regions)
Solution: PE1 failure is absorbed by BNs for BGP PIC.

L3VPN: RD:Pfx1, PNH=PE1, VL5
RD:PfxN, PNH=PEn, VLn

"MPLS-Namespace" Routes
RD:PL-CE1
PNH=PE1, VL5
PNH=PE2, VL6
Region-1

BL4
BL3
BN2
BN3
BN4
P3
PE2

LU: CPNH1, PNH=BN1, BL3
LU: CPNH1, PNH=BN3, BL1
LU: CPNH1, PNH=BN4, BL2

PE100

VL5
VL6
CE1
PE1
PE2

L-PE1
VL5
Payload

L-BN1
BL3
PL-CE1
Payload

Juniper Internal Presentation
Solution: PE1 failure is absorbed by BN1,2 for BGP PIC.
Solution: BN1 failure is absorbed by BN3,4 for BGP PIC.
Advantages

• Scaling needs on BNs and SNs get reduced, they don’t need all PE lo0 routes. Just one CPNH per region.

• Less turbulence in the network during failures and M/W. Churn absorbed within region.

• The MPLS-namespace routes absorb the shocks when the topology shakes, sans affecting service routes.

• PIC/EP like protection features provided at MPLS-namespace layer, is actually BGP Service family agnostic. Can provide sub-second traffic restoration against BN failure, and PE-failure also.

• Works with legacy PEs also. Only upgrade required is on RR, BNs. Localized feature between RR-BN provides network wide scale and convergence.

• Described procedure works for BGP-LU networks as-well as BGP-CT networks. The LU/CT route will advertise the CPNH address with Context-Label.
References

- [RFC-4364] BGP/MPLS IP Virtual Private Networks (VPNs)
- [RFC-5331] MPLS Upstream Label Assignment and Context-Specific Label Space
Thank you.