BGP Classful Transport Planes

https://datatracker.ietf.org/doc/draft-kaliraj-idr-bgp-classful-transport-planes/13

IETF IDR Interim Meeting

Jan 24, 2022

Presenter: Kaliraj Vairavakkalai

On behalf of coauthors: Natarajan Venkataraman, Balaji Rajagopalan, Gyan Mishra, Mazen Khaddam, Xiaohu Xu, Rafal Szarecki, Deepak Gowda

Agenda

- Recap – Problem statement.
- Recap – Solution, BGP-CT and advantages.
- BGP-CT: Current status, executive summary.
- CT vs CAR comparison.

BGP-CT recap: Problem statement.

- A domain has intra-AS tunnels with varying TE characteristics (gold, silver, bronze).

- There could be multiple tunnels to the same destination. And different tunneling protocols creating those tunnels.

- These tunnels may need to be extended inter-domain, while preserving their TE characteristics end-to-end.

- Different Service routes want to resolve (put traffic) over intra/inter-domain tunnels of a certain TE characteristic, with an option to fallback on tunnels belonging to a different TE characteristic, including best-effort tunnels. ***So, doing 'Intent driven Service-mapping' is the problem.***

- Solution should be agnostic of transport (RSVP, SRTE, Flex, IP-tunnels, etc..) and service layer (L3VPN, IPv6, Flowspec, Static, L2VPN, EVPN, etc..). i.e. works with any of these protocols in service and transport-layer.

- ***How to extend BGP to signal these pieces of information, and get the job done.***

BGP-CT recap: Solution constructs.

- Transport Class (TC): collects tunnels with same TE characteristics (gold, silver, etc). Transport-Class Identifier: 32-bit Color.

- BGP-CT is a new BGP transport layer address-family (SAFI: 76, "Classful Transport") that follows RFC-4364 procedures and RFC-8277 encodings.

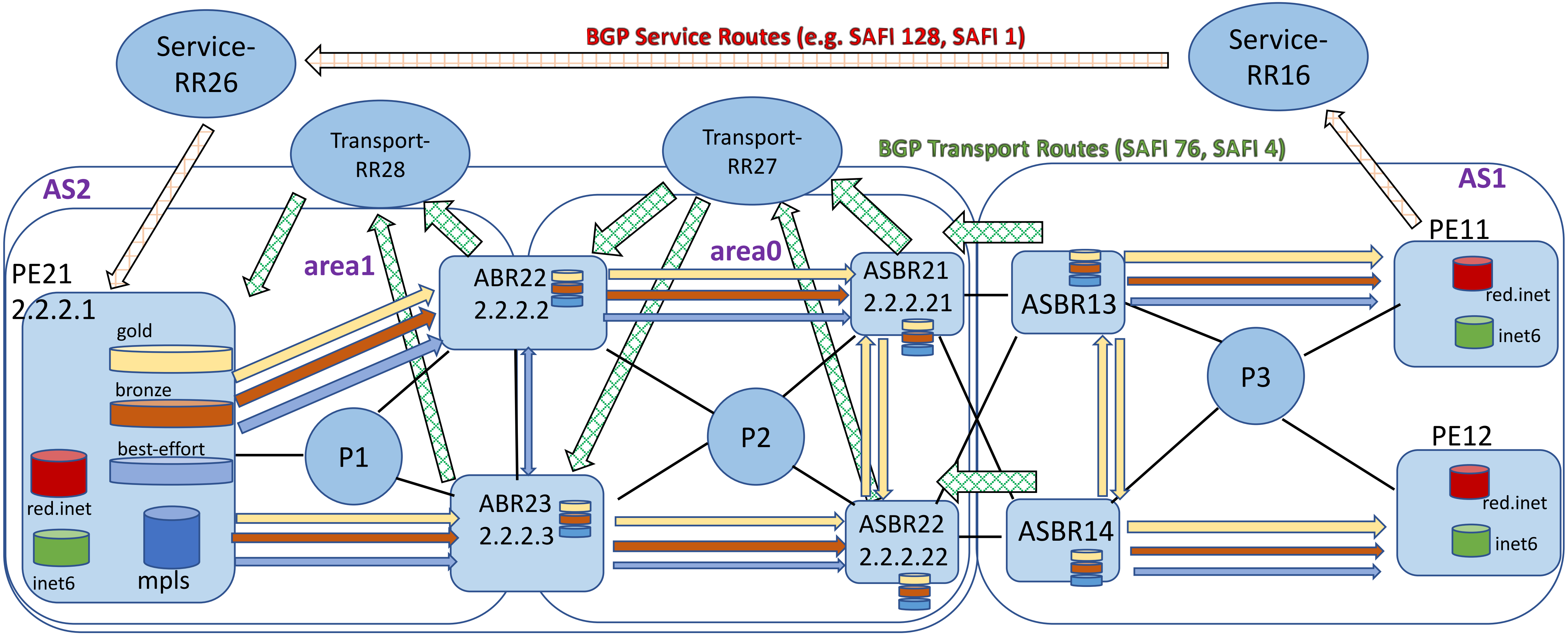
- Egress-PE originates BGP-CT family routes for it's own lo0, the tunnel-endpoint.
  - With "Route Distinguisher:TunnelEndpoint" as the NLRI.
  - And "Transport Class Route Target" that identifies the TC it belongs to. aka *Transport-Target.*
- Intra-domain tunneling protocols (RSVP, SRTE, FlexAlgo) install ingress routes to tunnel-endpoints in TC Transport-RIBs.

- BGP-CT extends the tunnel across inter-domain boundaries, while preserving the same Transport class end-to-end.
  - Resolve BGP-CT route's NH using tunnels belonging to the same Transport class, as specified by Transport-Target on the route.
  - Follow RFC-4364 option-C style procedures, to create swap-routes on domain boundaries.
  - Works in conjunction with option-A, option-B scenarios as-well.

- Service routes want to resolve using a Resolution scheme asper user intent *(e.g.. use tunnels of a certain Transport class, with an option to fallback on Best-effort or another Transport class)*.

- Desired Resolution scheme is signaled via "Mapping community" on BGP route. E.g:
  - Color:0:<n> on the service-route. Resolves over Color "n" tunnels, with fallback on 'best-effort' tunnels.
  - Transport-Target on BGP-CT route. Resolves strictly over Color "n" tunnels.

BGP CT : Transport Class based Network Slicing



- Transport Class (e.g. gold, bronze, best-effort) provides the "**Topology Slice**" in Network Slicing
- Intra-domain Transport routes are populated in Transport class RIBs by tunneling protocols (e.g. RSVP, Flex, SRTE).
- Inter-domain Transport routes are populated in Transport class RIBs by BGP-CT family (SAFI 76).
- Service-routes (e.g. L3VPN, Internet) map to a "Toplogy Slice" by using appropriate Mapping community (e.g. Color extended community).

BGP-CT: BGP protocol observations

❑ Desired intent (Resolution scheme) is signaled via “Mapping community” on BGP route. E.g:

- Transport-Target on BGP-CT route. Intent: Resolve strictly over Color “n” tunnels.
- Color:0:<n> on the service-route. Intent: Resolve over Color “n” tunnels, with fallback on ‘best-effort’ tunnels.

❑ Community/Route-Target rewrite. If the domains have different route-target values to represent an intent, then receiving domain-BN rewrites the received Transport-target to the appropriate Transport-target value for the local domain. This is similar to how L3VPN domains do route-target rewrite on AS boundaries today. Or, how Color:0:<n> communities are rewritten today in such scenarios.

❑ . “Route Distinguisher” is used to distinguish between different Transport-class routes for the same TunnelEndpoint, when propagating across route-selection pinch-points.

- Besides Route-Distinguisher, Add-path ID is also used, when crossing redundant domain-BNs and subsequent RR.
- RD allows to uniquely identify the originating PE, across a multiple domains, which is helpful in troubleshooting.

❑ The BGP CT spec does not specify any changes to how *forwarding semantics (*label or SIDs) are carried. It follows already existing mechanisms. E.g.

- RFC 8277 specified encoding is used to carry label.
- SIDs are carried Prefix-SID attr, as defined in mechanisms specified by SR specs.
- Any changes in these mechanisms will also work for BGP CT, as long as those changes are backward-compatible to these existing standards. E.g. multinexthop attribute.

❑ Utilizes RT-constraint family RFC-4684 to prune BGP-CT route distribution to required PEs only.

BGP-CT: advantages of reusing 4364 machinery

- Using RFC-4364 style “Route Distinguisher”.
  - Avoids using multiple loopbacks on Egress-PE, Avoids path-hiding when transiting RR/ASBRs,
  - Allows unambiguously identifying the originating PE, for debugging.
  - Supports TunnelEndpoint being an Anycast-address participating in multiple domains.
  - RD is not used when doing per-prefix-label allocation, thus confining ripple of link/node failures local to the region where failure happened.

  Basically, RD is an identifier of convenience. Use it when needed, Strip it when not needed. Preserved end-to-end.

- Using RFC-4364 style “Route Target” to propagate Transport-Class allows:
  - Forming Venn diagrams of color domains as desired.
  - E.g. Core network having more fine-grained colors than Access networks.

- Treating “Color” as an attribute (adjective), rather than part of NLRI (noun)
  - Helps in cases where domains have different numbering of color values. Attribute rewrites is easier than rewriting NLRI.

BGP-CT: advantages (contd.)

- BGP-CT organizes transport prefixes in RIB scoped per Color. So nexthop resolution is done by LPM (longest prefix match) on the transport-class RIB in an efficient manner.

- Union of transport RIBs with staggered preference allows achieving fallback to other color/intents.

- ODN using Route Target Constrain RFC-4684 procedures.
  - Service-routes can have a clean API with Transport-layer, to request for only the BGP-CT routes required by service-routes.

- Re-using the time tested, well deployed, RFC-4364 machinery:
  - Cuts down implementation, testing time. Improves reliability of the solution, and time to deploy.
  - **Protects the investment operators have made in operational training, tooling, and procedures. Inventing new things just for fun, creates new OpEx**

- BGP-CT preserves ROI of existing deployments, by supporting all transport-tunneling protocols including RSVP, SRTE, FlexAlgo.

BGP-CT: Current status, executive summary

- Draft submitted March 2020. Five IETFs ago.

- Juniper Implementation available since Junos21.1R1. Uses IANA allotted code-points. New features getting added every release since then.

- Customers interested to deploy, undergoing lab-trials. But request other vendors interoperability.

- Request IDR WG to help by adopting BGP-CT as a WG document.

CT vs CAR - proposal observations

- CAR “Where is my Color?” ambiguity. Prefix or LCM (more on next slide)
  - routes with different effective colors are keyed against the same NLRI/Prefix.
  - May result in inefficient lookup to find a matching nexthop with desired color

- RTC cannot be used for CAR routes. If LCM is attached to all routes, RTC mechanism may be extended to be applied to CAR routes. Whereas RTC readily works with CT routes.

- The CAR proposal tries to leverage ORF like mechanism by requesting interested NLRI. ORF has hop-by-hop behavior and RTC was invented to provide better scalability solution.

- CAR NLRI combines key and non-key fields. Lot of non-key fields. Which is always a bad idea. Need special care in dealing with ambiguity in how withdrawals are sent and processed. Increased implementation complexity.

- CAR SAFI seems to be used as transport-family between PEs, and service-family between CE-PE in VPN-CAR use-case. This overloading will also cause confusion in deployment. Seems not well thought thru.

- Also, like VPN-CAR will there be VPLS-CAR, EVPN-CAR, etc for each service-family? Seems like an over-kill. CT solves the problem with new attributes, that can work with all existing service-families.

CT vs CAR proposal observations (2)

To give a concrete example of why basic resolution over routes with color will not work efficiently with CAR encoding, consider the following prefixes:

- Rt1: Prefix: EP:1.1.1.1 Color:100
  - Attribute: LocalPref:100
- Rt2: Prefix: EP:1.1.1.1 Color:200
  - Attribute: LCM: 100, LocalPref:100
- Rt3: Prefix: EP:1.1.1.1 Color:300
  - Attribute: LCM: 400, LocalPref:100
- Rt4: Prefix: EP:1.1.1.1 Color:400
  - Attribute: LCM: 100, LocalPref:200
- Now, to find the ‘best path’ Color 100 tunnel for 1.1.1.1, an implementation needs to walk the full table, all entries, and find Rt4 in above example. Just a LPM lookup on the Prefix does not yield any meaningful result. Because the color in NLRI “may not” be the real color. So it doesn’t scale well.
- This is just basic resolution over routes with color. Imagine how fallback works!. May be not thought thru?
- In essence, what CAR provides is an unorganized set of transport-reachability data, with an ambiguous hint on what the color is. What CT provides is organized transport-reachability information nicely organized in scope of the transport-class(color).

CT vs CAR - proposal observations (3)

- CAR requires addpath enabled EBGP-peering, which is also a contentious feature. And advertising non-bgp paths in Addpath, which most implementations don’t support.
- Addpath-PathID does not help identify originator of the route. Like RD. Troubleshooting will be hard.
- Scaling method proposed in CAR draft (Hierarchical transport) increases the recursive resolution and ECMP-nexthop load on the ingress-PEs. Which may be low-end legacy devices. So it is not practical.
- Where-as scaling method proposed in CT architecture (MPLS namespaces) works with no changes on ingress-PEs. Only BNs and SHs need to be upgraded.   This reduces NH resources usage by reducing NH-addresses visible thru out the network.
- Ironing out CAR Interop issues will take its own sweet time. Again higher-cost. Overall operational complexity and training cost is higher.
- So the question needs to be asked: what is the advantage for the higher cost. When existing VPN machinery (that CT re-uses) solves all these problems in an elegant and “more efficient” way.

References

- BGP-CT: https://datatracker.ietf.org/doc/draft-kaliraj-idr-bgp-classful-transport-planes/
- PCEP RSVP Color
  draft-rajagopalan-pcep-rsvp-color-00
- Seamless SR – use cases.

https://datatracker.ietf.org/doc/draft-hegde-spring-mpls-seamless-sr/

- SRv6 and MPLS interop.

https://datatracker.ietf.org/doc/html/draft-salih-spring-srv6-inter-domain-sids/

- MPLS namespaces: signaled via BGP

https://datatracker.ietf.org/doc/draft-kaliraj-bess-bgp-sig-private-mpls-labels/

- Generic RTC

https://datatracker.ietf.org/doc/draft-zzhang-idr-bgp-rt-constrains-extension/

- BGP CAR: https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-03

Thank you.