CAR-CT adoption Poll – Part 3 – Operational differences

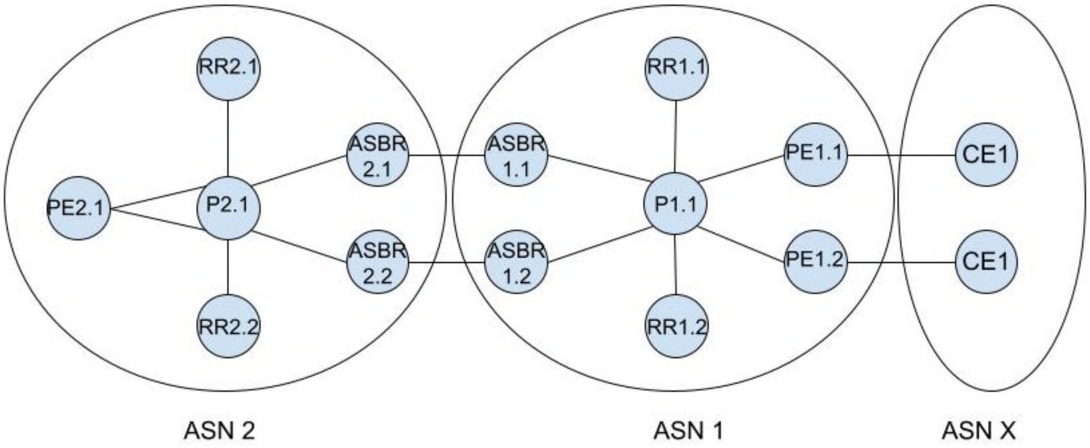
Contents

1: Questions on CT draft

F3-CT-Issue-1: SAFI 76 only in Option C

1. [Robert Raszuk]: For SAFI 76, is [it] only used in Option-C?

Topology:



Kaliraj Diagrams:

I had implied the topology diagram in section 18.1 https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-18.1 ,

when referring to section 18.3. Sorry if that was not clear.

Explanation:

The topology is classic Inter-AS, I prepared the illustration (limiting the drawing to only relevant nodes). The path hiding problem is well-known for SAFI 128 and its derivatives.

Let's assume that IGP metric of all links equals 1.

In a nutshell ASBRs flatten two dimensions of original ASN advertisements. Here ASN1.

In this ASN1 paths from PE1.1 and PE 1.2 are unique AS different RDs are combined with different next hops. That works like a charm in ASN1 in a full mesh or via RRs. However the moment you cross ASN boundary ASBR to ASBR will set next hop self - flattening the diversity yet keeping the RDs untouched.

Then when ASN2 RRs get the updates they get 4 paths ... two paths with next hop corresponding to ASBR2.1 and two paths with next hop corresponding to ASBR2.2.

So here comes the issue - when RRs reflect the paths they may choose the same next hop hence you get on the PE2.1 still two paths but with the same next hop of single ASBR2.1 or 2.2 therefore you lose redundancy.

[Robert Raszuk's Conclusion:] CT by default would be subject to this issue.

ADD-PATHs or DIVERSE-PATH proposals mitigate the issue by forcing to either advertise all paths from all RRs (former) or making sure that you create sessions on RRs as primary and backup so you get diversity and ingress get's two different (by next hop) paths. But the point here is that clearly using ADD-PATHs on top of RD based distribution models is simply an overkill.

[Robert]: Hope this helps to visualize the scenario we have been discussing. Good news is that there is no dispute on the issue and CT folks admitted it.

[Kaliraj: See Section 10.6 of draft-ietf-idr-bgp-ct-00.txt. [This discussion] was to helping Robert understand [this section]. There is no open issue on this point of "Addpath usage with BGP-CT".

Text:

CT draft says in abstract:

> "Though BGP CT family is used only in the option-C inter-AS [networks],*

CT draft also says in many sections that next hop is rewritten (example [in] Section 9]

> *whereas Classful Transport routes are readvertised over EBGP single hop sessions with "nexthop self" rewrite over inter-AS links.*

Topology:

Question [Robert] So for SAFI 76 how can the draft say that it is used only in option-C where it is evident that CT really uses option-B model with next hop rewrite ?

[https://mailarchive.ietf.org/arch/msg/idr/W8wofBQ3wCwrQBtyHg8NgROvqXY/]

Response-1: [Reshma Das]: This is True SAFI 76 is only used in inter-AS option C. It is known SAFI 4 (BGP-LU) in Option C at BN does Next hop self.

Stating an example from Cisco learning network:

https://learningnetwork.cisco.com/s/question/0D53i00000Ksqy9CAB/interas-option-c

Similarly for SAFI 76 at BN, if RD1:EP1 (transport-target:0:100 - gold) has a gold path to reach EP1 then RD1:EP1 is advertised upstream with Next hop self.

Counter-Reply-1: [Robert Raszuk]: Options for Inter-as route distribution are *ONLY* defined in RFC4364 section 10. Neither RFC3107 nor RFC8277 define Inter-as options. So you should not confuse anyone using terminology reserved for SAFI 128.

But nomenclature aside, … you realize that this was just to set the stage [for question part 2].

Question Part 2 [Robert Raszuk]: The real question I would like to ask is: how do deal in BGP-CT with day one L3VPN specification bug in the Inter-as scenario?

- As everyone deploying SAFI 128 across Inter-AS ASBRs is aware when you set next hop self (do SAFI 128 option B) you are losing redundancy. This is due to the fact that for any set of RDs from original ASN ASBR is overwriting their different next hops. Then ASBRs are sending those routes over IBGP to RRs which now chooses closest (or random) ASBR as exit for routes received from a pair (or more) of ASBRs. So effectively ingress PE is receiving multiple NLRIs but they all are going to the single ASBR !
- The only defined solution today is to use ADD-PATHs on top of SAFI 128 to force propagation of all paths (all ASBRs next hops). But CT does not like ADD-PATHs - so how are you planning to solve this original L3VPN distribution issue?

Response-2: [Kaliraj Vairavakkalai]:

Text: see https://datatracker.ietf.org/doc/html/rfc4364#section-10

> “c) Multi-hop EBGP redistribution of labeled VPN-IPv4 routes between source and destination ASes, with EBGP redistribution of labeled IPv4 routes from AS to neighboring AS. In this procedure, VPN-IPv4 routes are neither maintained nor distributed by the ASBRs.”

As described above, main difference between option (c) and (b) is that there is no service-route (SAFI 128) state at ASBRs. The ASBRs are “Transport ASBRs” that carry BGP-LU (SAFI 4) or BGP-CT (SAFI 76) routes. So, there is no ambiguity that SAFI 76 is a Transport-layer family, that enables Inter-AS option (c) solution.

Text: Further, Ref: https://datatracker.ietf.org/doc/html/rfc8277#section-3.2.2

RFC 8277 does define how nexthop-self for SAFI-4 and SAFI-128 routes achieve option-C or option-B label-swap forwarding at the BGP speakers re-advertising these routes with nexthop-self. Irrespective of whether it is SAFI-4, SAFI-128 or SAFI-76, the same procedures apply. Viz.

- Nexthop-self creates new label route for MPLS transit-forwarding, this label is advertised to BGP peers in the 8277 family.
- Those labels cross-connect the tunnels that the advertised BGP routes resolved over.

A BGP family is classified as service-family or transport-family based on what prefixes it carries. Transport-family (used in option-c) carries prefixes that are used as nexthops in service-families. SAFI 4, SAFI 76 are examples of Transport families. Service-families carry prefixes that appear in destination-field in the PDU on wire.

This is described in the following section:

https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-9

> “A new family also facilitates having a different readvertisement path of the transport family routes in a network than the service route readvertisement path. Service routes (Inet-VPN) are exchanged over an EBGP multihop session between Autonomous systems with nexthop unchanged; whereas Classful Transport routes are readvertised over EBGP single hop sessions with "nexthop self" rewrite over inter-AS links.

> The Classful Transport family is similar in vein to BGP LU, in that it carries transport prefixes. The only difference is that it also carries in Route Target, an indication of which Transport Class the transport prefix belongs to and uses RD to disambiguate multiple instances of the same transport prefix in a BGP Update."

[Kaliraj] *Responding to:* But CT does not like ADD-PATHs - so how are you planning to solve this original L3VPN distribution issue ?

Further, about Addpath, CT doesn't dislike Addpath 😊. BGP-CT does use Addpath to avoid the redundant ASBR path-hiding problem you describe, as stated in the following sections:

- https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-10.6
- https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-18.3

> "Addpath is enabled for BGP CT family on the sessions between RR27 and ASBRs, ABRs such that routes for 1.1.1.1:10:1.1.1.1 with the nexthops ASBR21 and ASBR22 are reflected to ABR23, ABR24 without any path hiding. Thus giving ABR23 visibiity of both available nexthops for Gold SLA."

We just don't think Add-path-ID is a good end-to-end distinguisher. Because it is per-session scoped. RD plays a better role of being an end-to-end distinguisher.

[Robert Raszuk- reply to Kaliraj]: The fundamental difference between SAFI 128 Option C vs Option B is that in the former next hop is not changed between ASNs while in the latter it is. I do not see any reason why not to say that CT uses Option-B especially based on the fact that ASBR do maintain state and do rewrite next hop.

- You seem to be stuck with Option C which simply is not what CT is doing. Why ?
- And also you keep bringing SAFI 4 over and over to the discussion which is completely irrelevant to the subject.

[skipping down to Robert's portion of the text]:

- So you confirm that your solution when deployed inter-as requires ADD-PATHs. Cool. We are making progress.
- You also need to educate all of your customers that without ADD-PATHs use the CT solution becomes single point of failure. And that is easier said then done. ADD-PATHs has been supported for SAFI 1 for a while but not necessarily for SAFI 128. So new code is needed which I am sure will be part of OS which includes SAFI 76.
- Note that path hiding on the RRs can also be addressed using BGP Diverse Path RFC 6774 without the need to use ADD-PATHs.
- I mention this here as those customers mobilized globally to support CT must be aware that analogy to SAFI 128 is not identical when it comes to SAFI 76.
- As far as your last comment of ADD-PATHs being session scoped and using this as counterargument let's observe that for transport layer where you set next hop on transport anchors actually per session scope makes much more sense as opposed to carry luggage from other ASNs into your domain.

[Kaliraj]: BGP-CT does use Addpath to avoid the redundant ASBR path-hiding problem you describe, as stated in the following sections: [10.6 and 18.3] [topology is in section 18.1] – when referring to 18.3

https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-10.6

https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-18.3

> "Addpath is enabled for BGP CT family on the sessions between RR27 and ASBRs, ABRs such that routes for 1.1.1.1:10:1.1.1.1 with the nexthops ASBR21 and ASBR22 are reflected to ABR23, ABR24 without any path hiding. Thus giving ABR23 visibiity of both available nexthops for Gold SLA."

[Kaliraj] To answer your question whether [an] To answer your question on whether implementation for add-path exists?

- Addpath for SAFI-128 has existed for long time in Junos production code, and yes shipping code exists for SAFI-76 as-well.
- Such usage of 'Addpath with RD:EP at RR' in SAFI-76 is similar to usage of 'Addpath with EP:Color at RR' in CAR. As stated in CAR draft sec 6.1:
    - "When a transport RR is used within the domain or across domains, ADD-PATH is enabled to advertise paths from both egress BRs to it's clients."
- So the overhead want to emphasize that BGP-CT is not against Addpath. BGP-CT uses existing Mechanisms like Addpath for solving path-hiding problems where applicable.
- Just that, from the troubleshooting POV, RD helps to identify who originated the route, but Addpath-ID doesn't.

[Robert]: Yes I think we got very much in sync on the path hiding issue.

- As far as adding ADD-PATHs to RD based route distribution I think this is an overkill. Personally I would rather algorithmically modify RD on the ASBRs. Just that, from the troubleshooting POV, RD helps to identify who originated the route, but Addpath-ID doesn't.
- Well if this is the only reason I think we can do much better then using RDs to identify the originator of the path. Besides after import at the dst node RD information is gone.

F3-CT-Issue-1: Issues to Load in Issues Tracker in github

- Add text to clarify the operational of CT under topologies for option A, B, or C
- Expand current Add Paths Example for service and transport AFI/SAFI families
- Add text on why RFC8277 (Appendix)
- Explanation of how RD can be used as an end-to-end distinguisher. In this explanation, please explain what happens at border routers (BRs).

F3-CT-Issue-2: Sizing Discussion – beginning in NLRI format discussion

Author: Ketan Talaulikar (ketant.ietf@gmail.com)

Text: This is regarding the NLRI encoding in the new BGP CT proposal where there exists an embedded MPLS label field in the NLRI per RFC8277 encoding.

Kaliraj's response on the other thread is here:
https://mailarchive.ietf.org/arch/msg/idr/uv-bQATEgL-cLtmJFbw3DYK37K4/

I don't think we can let this slip claiming the support for RFC8277 – a reminder this is a new SAFI design and there is no compulsion to use the label encoding from RFC8277.

So I agree with what Robert said on that thread:
https://mailarchive.ietf.org/arch/msg/idr/mg3k6GRT6WLsLpQclV8cO57NHSY/

Further in Section 17 where SRv6 support is covered, there is the proposal to adopt an approach similar to RFC9252 (BGP Services over SRv6) that overloads the MPLS label field.

There have been other proposals in the past (e.g., RFC8365 EVPN) where this has been done, but in all of those cases, it has always been about extending to other newer encapsulations something designed originally only for the MPLS data plane.

This is the first proposal that seeks to perpetuate that mistake into a brand new SAFI designed with full awareness of the existence of multiple encapsulations in the transport layers in today's networks.

Note that during the review of RFC9252, there were concerns raised (one of the examples in [1]) on this very topic and the WG should not be adopting new proposals that make the same protocol design mistakes.

Thanks,
Ketan

[1] https://mailarchive.ietf.org/arch/msg/bess/JyzFH7Z9SjbS4Ni82_Knv9Ou-iM/

Responses:
[Kaliraj-1]: Hi Ketan - Thanks for raising this issue, with pointer to [1] in your email thread. Please find the clarifications below.

- BGP-CT follows existing SRv6 procedures described in RFC9252. I agree that the 'Transposition mechanism' overloading the MPLS-label field of RFC-8277 families is not a good idea. As it leads to security and robustness issues as pointed to by [1], and also causes interop issues during migration.
- **We plan to** disallow SRv6 Transposition for SAFI 76 in BGP-CT draft. We understand this may not be desirable for update packing purposes. But given we have to choose between SRv6 procedures that suite either update-packing or 'security and robustness', we choose the latter.
  - **[Ketan-2]:** Good to know this. I hope we'll see this update in the next version of your draft.

- Things like SiteOfOrigin community already affect update packing. So I don't believe we need to micro-optimize for update-packing, especially with such 'security and robustness' tradeoffs.
    - **[Ketan-2]:** It is interesting that you have taken the example of SiteOfOrigin for a transport SAFI like CT. Could you clarify the use case that you foresee for it with BGP CT?
    - **[Ketan-2]:** I would not characterize update-packing as micro-optimization. For the "intent-aware" transport solution, I can foresee a significant scaling challenge coming up with the increase in transport routers that will get multiplied with the types of "intents" to be realized. Update-packing mechanisms help optimize the BGP messaging load. This is especially critical in times of reconvergence and network churn.
        - **[Jeff-**1]: A detail I was pondering while I was doing some of the work for the chairs' slides covering update packing was intended scale. The customers I'm personally dealing with at this time are somewhat atypical of regular scale.
            - What do you envision as being typical scale for number of CAR routes of any color in a larger provider?
        - **[Jeff-1]** The motivation for the question was roughly packing impacts vs. scale. For Internet mix with the usual packing, thousands of updates per second are regularly happening.[1] Convergence time for 10k routes with color won't be terribly significant, even with zero packing. Some of the more insane network slicing scaled scenarios may be better points of comparison. Clearly update packing will reduce that even more so. But what I think we're starting to hit as a point of comparison is "for expected loads for the mechanisms, what's the general time for each?"
        - **[Ketan-3-reply]:** The scalability requirements are captured here: https://datatracker.ietf.org/doc/html/draft-hr-spring-intentaware-routing-using-color-00#section-6.3.2
          This is the merged document that, I believe, captures the consensus that both the CAR and CT solutions aim to address.
            - **[Jeff-2]:** Thanks, Ketan. Roughly 1.5 million routes. Presuming an example 10k update per second handling, roughly 2.5 minutes of convergence time without packing optimizations.
            - **[Robert-reply-to-Jeff-2]:** So I have a few different questions here. (Editor's note – this is move to a Discussion point on CT
              [see F3-CT-Q2 in the text above]

Robert discussion starts a new topic (Editor)
[Robert-reply-to-Jeff-2]: So I have a few different questions here. Assume in CAR/CT enabled domain one color has transport problems ... say low latency is becoming not so low due to interface queuing is transiently congesting for whatever reason between P1 and P2 nodes (not even running any BGP).

- Q1 - How (by what exact protocol) and how fast such issue with forwarding a given color via this domain will be visible at the CAR/CT layer?
- Q2 - Assume Q1 is done - do we now need to withdraw 300K routes based on one color brownout?
- Q3 - According to your math such CAR/CT reaction will take 30 sec. What if transport problem is transient and occurs for say 5-10 sec every 40 sec ?
- Q4 - Is there in any document an analysis on dynamics of CAR/CT signalling needed to make this at all practical in real deployments vs ppts?

[Robert]: We keep burning energy on encoding, but apologies if I missed it but I am not seeing the full picture here. Why not advertise just 5 colors between those domains in 5 NLRIs and define a new attribute to carry all the interdomain color mappings in it ? 5 being an example from the section 6.3.2 ... but realistically we could perhaps vastly simplify this if we define day one set of well-known colors instead of each domain inventing their own definition :) Maybe I am just too practical here - but your math inspired those questions :)

- **[Jeff-3]:** A partial comment from my mobile device. Withdraw encoding will pack much denser. On a total withdraw you likely could pack 200 or more prefixes per update. Implicit withdraw via replacement is clearly same speed as initial advertisement. The stability dynamics and impact of service route re-resolution are largely the same as BGP labeled unicast. Thus, beware churning your transport routes.
    - **[Robert-2]:** I would quite not agree with the above. Reason being that labeled unicast is about reachability. Here we are talking about real promises of data plane "performance" hence we are dealing with completely different set of triggers for various data plane issues.
        - **[Jeff-**4]: I make no comment on how it is intended to be deployed. Only that the consequences protocol-wise are the same.
            - **[Robert-**3]: My point is that this is the first time we are facing the introduction of BGP invalidation (as you stated no resolution) by performance (or under-performance) of data plane metric. I think this has new consequences to the protocol which are nowhere near SAFI 4. Perhaps it could work just fine for reasonable scale. But the numbers being quoted of 1.5M color routes seems way too excessive and rather suggest different protocol encoding or one more layer of hierarchy/indirection needed.

F3-CT-Issue-2: Issues to Load in Issues Tracker in github

- Add text on Why CT uses RFC8277 (Appendix) that includes scaling discussion.
    - Sizing discussion should reference the following (at least): draft-hr-spring-intentaware-routing-using-color-00#section-6.3.2, and
    - packing for SRv6 SIDs

F4-CT-Issue-3: BGP-CT and RTC

[Originator: Jeffrey Zhang]

Issue: An incorrect claim in
https://mailarchive.ietf.org/arch/msg/idr/AB9pSxxfG0fwOTOgBuiej9W8mpQ/

Quote: "BGP CT proposes RTC for constraint distribution. RTC filters on RT and thus need to carry PE address based RT in CT route. This means same PE address is carried in 2 places i.e. NLRI (RD:PE address) and Route Target: PE:TC. This breaks BGP packing completely. Each update can carry only single CT route"

BGP-CT text: (missing)
Topology: (missing)

[Jeffrey Zhang] Discussion: For #2, I want to point out that unless you want to prevent the propagation of a certain BGP-CT NLRI to where it does not have to go (e.g. a low-end edge only wants to receive routes for certain destinations that it cares about) based on the destination itself, you don't need an IP address specific RT. By default, the BGP-CT updates will only have a RT that both encodes the TC and controls the propagation and importation of the routes of that TC to where it is needed.

Responses:

[Robert Raszuk]: On #2, I agree with what you said.

[Swadesh]: I don't know how you interpreted my statement, but the claim is not incorrect. Section 14.2 of draft-kaliraj-idr-bgp-classful-transport-planes-17 clearly states the following for constraint distribution in scale network.

- "An egress SN MAY advertise BGP CT route for RD:eSN with two Route Targets: transport-target:0:<TC> and a RT carrying <eSN>:<TC>. Where TC is the Transport Class identifier, and eSN is the IP-address used by SN as BGP nexthop in its service route advertisements.

[Swadesh]: It is MAY, of course, but for scaling that is the described solution. And when <eSN:TC> RT is added, it is unique per RD:eSN route. Hence each bgp update can carry only single CT route.

F3-CT-Issue-3: Issues to Load in Issue Tracker in github

- Expand Section 18.3 in draft-ietf-idr-bgp-ct-00 to indicate the lack of need for "an IP address specific RT" for scenario given by Jeffrey Zhang.
- Clarifying section 14.2 to address Swadesh comments regarding RTs with "<eSN>:<TC> Clarifying should be alignment with F3-CT-Issue-2

F3-CT-Issue-4: Q3-CT-ANSWER-2 Clarification to Bruno on BGP-CT NLRI format.

[Kaliraj] to this Part3 email-thread, as suggested by Susan. To answer your question:

- Yes BGP-CT follows RFC-8277, and advertises a single label, when "Multiple Labels Capability" is not negotiated. We don't foresee the need to use Multiple Labels Capability with BGP-CT.

[Bruno] It's interesting that you mention a BGP capability to advertise multiple labels.

- RFC3107 did not require one. RFC8277 was obliged to define one because the same AFI/SAFI than RFC3107 is used and there is a need to distinguish between implementation compliant with 3107 (i.e. accepting a label stack) from implementation pretending to be compliant (not accepting a stack). The situation is different for BGP-CT as a new AFI/SAFI is defined hence there is a priori no need to define two NLRI encodings (stack vs single label) and a capability. Why do we need this extra complexity?

- So, Section 7 of the BGP-CT draft illustrates the BGP-CT NLRI with single-label.
  - "For better readability, the following figure illustrates a BGP Classful Transport family NLRI when single Label is advertised:"
- Procedures with respect to whether single-label or label-stack is encoded is same as SAFI-4 or SAFI-128.

[Bruno] "BGP-LU" allows advertising multiple labels from day one (both RFC3107 and 8277). So I understand that BGP-CT allows the advertisement of multi labels in the NLRI. Please correct me if I'm wrong.

F3-CT-Issue-4: Issues to Load in Issues Tracker in github

- Revise CT text to indicate whether CT allows for multiple labels in NLRI

F3-CT-Issue-5: Perpetuating known issues with embedded MPLS label field in a new SAFI

[original]: https://mailarchive.ietf.org/arch/msg/idr/R98YR27K31ZlbdDvsbMX1f-liTY/

[Ketan Talaulika]: This is regarding the NLRI encoding in the new BGP CT proposal where there exists an embedded MPLS label field in the NLRI per RFC8277 encoding.

- Kaliraj's response on the other thread is here:
  https://mailarchive.ietf.org/arch/msg/idr/uv-bQATEgL-cLtmJFbw3DYK37K4/
  (editor: check reference)

[Ketan initial Discussion]: I don't think we can let this slip claiming the support for RFC8277 **-** a reminder this is a new SAFI design and there is no compulsion to use the label encoding from RFC8277.

- So I agree with what Robert said on that thread: https://mailarchive.ietf.org/arch/msg/idr/mg3k6GRT6WLsLpQclV8cO57NHSY/

Text:

- Further in Section 17 where SRv6 support is covered, there is the proposal to adopt an approach similar to RFC9252 (BGP Services over SRv6) that overloads the MPLS label field.

[Ketan's discussion continues]:

- There have been other proposals in the past (e.g., RFC8365 EVPN) where this has been done, but in all of those cases, it has always been about extending to other newer encapsulations something designed originally only for the MPLS data plane.
- This is the first proposal that seeks to perpetuate that mistake into a brand new SAFI designed with full awareness of the existence of multiple encapsulations in the transport layers in today's networks.
- **Note**: that during the review of RFC9252, there were concerns raised (one of the examples in [1]) on this very topic and the WG should not be adopting new proposals that make the same protocol design mistakes.

[Responses]:
[Kaliraj]: (https://mailarchive.ietf.org/arch/msg/idr/5_zqGzLuazwcWRFUc_Ktjd80DFY/)
Thanks for raising this issue, with pointer to [1] in your email thread. Please find the clarifications below.

- **[Kaliraj]:** BGP-CT follows existing SRv6 procedures described in RFC9252. I agree that the 'Transposition mechanism' overloading the MPLS-label field of RFC-8277 families is not a good idea. As it leads to security and robustness issues as pointed to by [1], and also causes interop issues during migration.

- **[Kaliraj]:** We plan to disallow SRv6 Transposition for SAFI 76 in BGP-CT draft. We understand this may not be desirable for update packing purposes. But given we have to choose between SRv6 procedures that suite either update-packing or 'security and robustness', we choose the latter.
  - **[Ketan-reply]:** Good to know this. I hope we'll see this update in the next version of your draft.

- **[Kaliraj]:** Things like SiteOfOrigin community already affect update packing.
  - **[Ketan-reply]:** It is interesting that you have taken the example of SiteOfOrigin for a transport SAFI like CT. Could you clarify the use case that you foresee for it with BGP CT?
    - **[Kaliraj-2**]: I meant, "community identifying origin site", which could be SoO, or any other BGP-community. Or attributes like AIGP as-well affect packing – these are used in BGP-LU deployments today. Bottomline is, looking at how much packing we get in

current BGP-LU deployments will give an idea. We cannot get more packing than that, by stuffing things in the NLRI.

- **[Kaliraj]:** So I don’t believe we need to micro-optimize for update-packing, especially with such ‘security and robustness’ tradeoffs.
  - **[Ketan-**reply]: I would not characterize update-packing as micro-optimization. For the "intent-aware" transport solution, I can foresee a significant scaling challenge coming up with the increase in transport routers that will get multiplied with the types of "intents" to be realized. Update-packing mechanisms help optimize the BGP messaging load. This is especially critical in times of reconvergence and network churn.
    - **[Jeff Haas]:** A detail I was pondering while I was doing some of the work for the chairs' slides covering update packing was intended scale.
      - The customers I'm personally dealing with at this time are somewhat atypical of regular scale.
      - **What do you envision as being typical scale for number of CAR routes of any color in a larger provider?**
      - The motivation for the question was roughly packing impacts vs. scale.
      - For Internet mix with the usual packing, thousands of updates per second are regularly happening.[1] Convergence time for 10k routes with color won't be terribly significant, even with zero packing.
      - Some of the more insane network slicing scaled scenarios may be better points of comparison. Clearly update packing will reduce that even more so.
      - But what I think we're starting to hit as a point of comparison is "for expected loads for the mechanisms, what's the general time for each?"
    - **[Kaliraj (KV2)]:** I agree packing is desirable, but not at expense of ‘correctness’. So Transposition must be deprecated and disallowed for RFC-8277 NLRIs.
    - **[Ketan]:** The scalability requirements are captured here:
      - https://datatracker.ietf.org/doc/html/draft-hr-spring-intentaware-routing-using-color-00#section-6.3.2
      - This is the merged document that, I believe, captures the consensus that both the CAR and CT solutions aim to address.
    - **[Jeff]: Thanks, Ketan. [Scaling is]**
      - Roughly 1.5 million routes.
      - Presuming an example 10k update per second handling, roughly 2.5 minutes of convergence time without packing optimizations.
    - **[Robert]:** Sure 300k times 5 colors makes it 1.5M. So I have a few different questions here. Assume in CAR/CT enabled domain one color has transport problems
      - Say low latency is becoming not so low due to interface queuing is transiently congesting for whatever reason between P1 and P2 nodes (not even running any BGP).
      - Q1 - How (by what exact protocol) and how fast such issue with forwarding a given color via this domain will be visible at the CAR/CT layer ?
      - Q2 - Assume Q1 is done - do we now need to withdraw 300K routes based on one color brownout?
      - Q3 - According to your math such CAR/CT reaction will take 30 sec. What if transport problem is transient and occurs for say 5-10 sec every 40 sec ?
      - Q4 - Is there in any document an analysis on dynamics of CAR/CT [signaling] needed to make this at all practical in real deployments vs ppts?
        - **[Jeff-3]:** A partial comment from my mobile device. Withdraw encoding will pack much denser. On a total withdraw you likely could pack 200 or more prefixes per update. Implicit withdraw via replacement is clearly same speed as initial advertisement. The stability dynamics and impact of service

route re-resolution are largely the same as BGP labeled unicast. Thus, beware churning your transport routes.

- **[Robert]** We keep burning energy on encoding, but apologies if I missed it but I am not seeing the full picture here.
    - Why not advertise just 5 colors between those domains in 5 NLRIs and define a new attribute to carry all the interdomain color mappings in it?
    - Maybe I am just too practical here - but your math inspired those questions :)

- **[Kaliraj]:** Further, to answer the important question of 'why carry MPLS-label along with SRv6 encap', answer is: it is required for Interop during migration.
    - **[Robert]:** So what happens post migration? Your protocol proposal makes it unnecessary luggage for years to come. While migration is important it should have a transient character when designing a new protocol extension.
- **[Kaliraj]:** E.g. consider the following scenario we encountered in a real customer deployment. They were rolling out IPIP in their network, and wanted compatibility with existing "MPLS-only" L3VPN PEs in their network.

```
                     +--------R2  [MPLS only]
                     |
          R1---------P
     [MPLS + IPIP]   |
       (Egress)      +--------R3  [IPIP + MPLS]
```

[Kaliraj]: Here, BGP-speaker R1 can interop with both R2 (which supports only MPLS) and R3 (which supports IPIP also), by advertising both:

· MPLS-label in 8277 NLRI, and
· IPIP encap in TEA attribute.

- **[Ketan]:** You have picked up a good example to support your point :-). I note that all the endpoints are MPLS capable here. Your next example is more interesting.
- **[Kaliraj]:** I had snipped it for brevity. Let me state it for completeness: Consider R4 with IPIP-only support. R4 sends the BGP route with TEA, with semantics to 'ignore-embedded-label', such that MPLS label in 8277 NLRI is ignored at R1.
    - https://www.rfc-editor.org/rfc/rfc9012.html#name-embedded-label-handling-sub
- **[Kaliraj(KV2)]:** In fact this is what happens at R3 as-well, when it choses to use IPIP-encap towards R1.Similarly, for SRv6 one can imagine:

[SRv6 example]

```
                      +-------R2  [MPLS only]
                      |
          R1----------P-------R3  [SRv6 + MPLS]
     [MPLS + SRv6]    |
       (Egress)       +-------R4  [SRv6 only]
```

- **[Kaliraj]:** Node R1 can interop with R2, which supports only MPLS, R3 which supports both MPLS and SRv6, and R4 which supports only SRv6 – if the MPLS label field is left alone without getting polluted by Transposition.

  - o **[Ketan]:** What is your proposal that R4 set in the MPLS label field in this case? We also need to consider an RR in the middle. I look forward to seeing more details on these aspects in the upcoming update(s) of your document.
  - o **[Kaliraj (KV2)]:** Sure**. Will specify in the next update. Pls stay tuned.**

- **[Kaliraj]:** We don't need to stop using RFC-8277 because SRv6 Transposition causes the problem. We can stop using SRv6 Transposition.
- **[Kaliraj]:** Irrespective of where the forwarding info is carried (NLRI or Attribute), a BGP Route should be able to advertise forwarding information pertaining to multiple mechanisms. So that the receiver of the route can choose which ever mechanism it supports.
- **[Kaliraj]:** Looking at the comments on 'no need to carry MPLS-label in SRv6 deployments', it appears only "greenfield deployments" are being considered. As against both "brownfield and greenfield deployments" being the focus in BGP-CT.

  - o **[Ketan]: KT]** The WG is designing a new SAFI now in 2022 which needs to stay relevant for many more years in operator networks. Looking at BGP-LU, it has been around for over 2 decades now. However, we all agree that the pace of change is much faster these days than in the times before. No doubt the focus needs to be on greenfield and brownfield both - also the migration in between. However, what you are referring to as greenfield technologies now will be tomorrow's brownfield - and there would be a "greener-field" coming up ;-). Should we not learn from our experiences in the current transition of encapsulations and design with a more forward-looking extensible approach?

Summary comments:

[Kaliraj (KV2)]: It seems there is unanimous consensus that use of Transposition with RFC-8277 NLRI is a bad idea. Because it overloads MPLS-label field which may cause mis-routing at MPLS PEs; may even conflict with well-known special purpose MPLS labels.

  - o With this problem, it seems SRv6 cannot be deployed in today's MPLS networks.
  - o If we were aware of this, why was 9252 allowed to become an RFC?
  - o Is backward compatibility and migration considerations 'not a thing'?
  - o Should the WGs work on a 9252-bis, disallowing SRv6 Transposition for all RFC-8277 NLRIs?
  - o This will preserve sanity of the MPLS-label field in RFC-8277 NLRIs.

- **[Ketan (KT2)]:** Unanimous? :-) ... The context of my thread and our following conversation was about the NLRI design for a new SAFI. This was not about discussing the best options for extending existing BGP SAFIs for newer encapsulations.
  - o **[Kaliraj**-3] We are continuing this constructive discussion for existing families as-well. Because they are affected the same way by the mis-routing caused by SRv6 transposition schemes. And you seemed to be in agreement with John in the below thread about this problem:
    - ▪ https://mailarchive.ietf.org/arch/msg/bess/JyzFH7Z9SjbS4Ni82_Knv9Ou-iM/

  - o **[Kaliraj-3]:** That's why I said unanimous consensus.
    - ▪ Why do you want to stop the constructive discussion at the line of new SAFIs?
    - ▪ It is more important in context of existing SAFIs IMO, because they are already deployed in the field.
- **[Robert Raszuk]:** Do you recognize the difference between transport layer and service layer? From your rant below it looks like you are completely mixing documents discussing the service layer with this thread discussing the transport layer.
  - o **[Kaliraj-3]:** Sure I do Robert! I also recognize both service and transport layer need correctness and are affected by misrouting problems being discussed in referred thread:
    - ▪ https://mailarchive.ietf.org/arch/msg/bess/JyzFH7Z9SjbS4Ni82_Knv9Ou-iM/

- **[Kaliraj-3]** What is your argument for ignoring this problem for service layer?
    - **[Robert-2]:** I do not think anyone ignored it. In fact, if you read RFC9252 you will notice that there is a detailed description of how VPN label can be part of SID.
        - See SRv6 came way after RFC2547/4364 so it is expected to handle legacy services.
        - Contrary to this entire CAR/CT debate which is a clean slate in the way to look for a new transport model. There is no legacy (other than some specific vendor's dogmas) which we need to care about.
    - **[Robert-2]:** With that maybe we could start a separate IDR or BESS thread where you can enumerate what is not correct in RFC9252 ?

[Kaliraj (KV2)]: In (the discussion) thread [above] there is agreement we should do this for BGP-CT. I think we should solve the same problem for other existing RFC-8277 NLRIs also. Same proposal of disallowing transposition will help those cases as-well.

- Frankly, SRv6 should not get to do this mistake in the first place and then also demand that RFC-8277 should not be followed anymore because of this mistake. That doesn’t seem fair.
    - [Robert Raszuk]:
- **[Ketan (KT2)]:** I thought we were having a good and constructive discussion thus far on this thread until this misdirection. We have two different scenarios:
    - there is (1) where we have had to extend support for existing SAFIs to newer encapsulation like SRv6 and VXLAN, and then there is
    - (2) where we are introducing a new SAFI design.
- **[Ketan (KT2)]:** The prior existence of RFC3107/8277 label encoding in the NLRI does not make it right to introduce the same design bugs in newer WG products.
    - And just to be clear, the "bug" that I refer to is not the transposition itself but the base design of the carrying of MPLS labels as a fixed/mandatory field in the NLRI.
    - MPLS labels are optional and not mandatory in today's transport networks.

Overall-response

- **[Gyan]:** I was involved in the WGLC discussions related to the SRv6 BGP Services BGP Prefix SID attribute RFC 8669 encoding of SRv6 Service TLVs into 2 new TLVs SRv6 L3 Service TLV and SRv6 L3 Service TLV and transposition schemes providing the corresponding equivalent functionality provided by MPLS labels with L2 or L3 service route transposed into the variable function/argument portion of the SRv6 SID.
    - The issue with mis routing was related security issue with route leaking to the internet outside of the closed provider domain. The SRv6 SID is an IPv6 128 bit address that if it were leaked that would expose the traditional VPN overlay BGP NLRI encoding, which now flattened into a single layer with SRv6 encoded into the IID host portion of the IPv6 address SRv6 function/argument which could be a security issue. That was the problem at hand being discussed.
    - As noted in RFC 8402 S6 was also discussed on ML to secure the limited or closed domain perimeter with infrastructure ACLs inbound to block “any to B:N::/48” as well as only advertise the summary B:N::/48 to the internet.
    - What I had brought up and Ketan on the WGLC agreed that as next hop self is deployed on all the Gao-Redford model typical peering that really the B:N::/48 summary does not need to be advertised at all.
        - https://datatracker.ietf.org/doc/html/rfc8402#section-8.2
    - During WGLC was mentioned as well to why the flattening with SRv6 of the BGP NLRI into a single data plane layer versus traditional MPLS underlay/overlay clean layering and that this was the 2nd time a design was completed and published with the first being with EVPN RFC 7432 providing the same trickery setting the precedent with flattening of the NLRI to support both MPLS and non MPLS NVO RFC 8365.

        - With that question as the train had left the station for both there were no comments as I recall as to that flattening precedent being sent and the possible issues surrounding.
        - In that discussion update packing and issues and optimizations as well as overhead related to update packing was not discussed.

    - With regards to new SAFI and green and brown field discussions as well as the best optimized method of NLRI encoding I do see how the SRV6 thread discussed above came into play here, as the transposition scheme is related you can say to how to carry the BGP NLRI in the control plane / data plane. However in that WGLC thread the overall issue was not the transposition as that was a red herring, however the crux of the WGLC discussion was security related to route leaks to the Internet termed "mis routing".
        - https://www.mail-archive.com/bess@ietf.org/msg07020.html
        - https://www.mail-archive.com/bess@ietf.org/msg07185.html

F3-CT-Issue-5: Issues to Load in Issues Tracker in github

Update in CT document to disallow SRv6 Transposition for SAFI 76 in BGP-CT

- Additional details for SRv6 only example discussed by Kaliraj and Ketan – for R4 set in MPLS label field.
- Discussion on Migration from BGP-LU (RFC3101) to RC8277 to CT
    - Indicate that LU 3107 to RFC8277 is not in the scope of this document.
    - Add a Section on Interaction with LU that includes:
        - Deployment of CT along with LU
        - Transition from LU to CT
    - RFC-9252-bis:
        - Write document indicating the issues of problems with SRv6 transport for RFC8277 NLRIs
        - Any conflict between using SRv6 and MPLS-centric NLRIs

F3-CT-Issue-6: CT’s claims on the benefits of using unique RDs for advertised routes

Alternate title: CT’s claims on the benefits of using unique RDs for advertised routes – the claimed benefits being troubleshooting, enabling forwarding diversity etc.

[Editor’s comment: A discussion of Unique RDs may have to be included in both CAR and CT drafts]

[Author: DJ [Dhananjaya Rao]:

Introduction:
CT strongly recommends using unique RDs for advertised routes – the claimed benefits being troubleshooting, enabling forwarding diversity etc.   I’m starting a new question on one of the basic issues that the CT RD based approach is exposed to, as suggested by Nats in the thread :
https://mailarchive.ietf.org/arch/msg/idr/Quqc-Z3--lfmyfgzBXRNB-b9730/

This is not a new issue, though it wasn’t discussed as part of Q3. And a couple of options Nats responded with raise new questions. This issue impacts all use-cases discussed and has a greater impact on the popular Anycast scenario.

CT strongly recommends using unique RDs for advertised routes – the claimed benefits being troubleshooting, enabling forwarding diversity etc.

But the use of unique RDs create e2e LSPs to the originating nodes that do not provide local domain convergence and multipath. Withdraws of CT routes need to be propagated up to ingress PEs for traffic re-convergence.

For this, CT defines a “RD stripping”/import procedure [See draft-kaliraj-idr-bgp-classful-transport-planes-17#section-10.4]. However, it still does not provide failure localization/suppression at intermediate transit

Topology:
For illustration, we take the same topology from the other thread, filled out slightly:

```
                  Domain2
                     --------   BN2 -- -- -- E2 (IP1, C1)
                     |
       E1  -- -- -- BN1
                     |
                     --------   BN3 -- -- -- E3 (IP1, C1)
          Domain1               Domain3
```

As described in the CT draft referred section, at a transit node, such as BN1 in Domain1, both routes come together via the RD stripping / import procedure to form a multipath in C1 RIB.

Discussion to above text by CAR author (DJ) and CT authors (Nats, Kaliraj)

- **[Nat-reply (NV3)]** As per Section 8, there is nothing that is “**strongly” recommended**. Customers will be provided the flexibility to configure how they want their endpoint/BN and how they would want their RD under it. Please read on for further explanation of label allocation modes. https://www.ietf.org/archive/id/draft-kaliraj-idr-bgp-classful-transport-planes-17.html#name-use-of-route-distinguisher
    - **[DJ-2]:** Literally read, -Section 8- does not include “strongly” recommend. But it describes potential benefits of using unique RDs, without describing any of the

limitations/disadvantages. However, other sections in the latest version (-17) of the CT draft have following statements:

  - Section 10.9: "Deploying unique RDs is strongly RECOMMENDED because it helps in troubleshooting by uniquely identifying originator of a route and avoids path-hiding. "
  - Section 10.1: "Unique RD SHOULD be used by the originator of a Classful Transport route to disambiguate the multiple BGP advertisements for a transport end point."
  - Section 10.2: "Unique RD SHOULD be used by the originator of a Classful Transport route to disambiguate the multiple BGP advertisements for a transport end point."

- **[DJ-2]:** I didn't check further. And of course, we have your many statements on the list and in IDR presentations. Now, even though you did not address the specific issues below, from the couple of statements you made about operator flexibility to configure, you appear to be leaning towards using the same RD values as a potential solution to the quandary of slow convergence, providing basic BGP multipath/local repair without redundant routes and churn.
  - I couldn't find a single mention of configuring the same RD values on originators in the -17 version of the CT draft. Apologies if I missed it. Or perhaps that will be in the next version.
- [**DJ-2]** Using the same RDs on originators undermines all benefits you have repeatedly stated for having an RD in NLRI. But if that's really the only practical option, then it begs the question, why is an RD, and its subsequent requirement for a VPN-like import at every hop, needed at all in the transport NLRI?
- **[Kaliraj-reply-DJ (Kaliraj-1)]:** About the churn argument, not true. The advertised CT label does not change when local failure events happen and nexthop changes from one nexthop to another. Because of Per-Prefix/Per-TC-Prefix label allocation mode. So Churn is not propagated further. About RD:EP routes propagated further, that is correct, but it would not churn on local failure events. And the 'further propagation' can be advantageous in some cases, explained towards end of this email
- **[Kaliraj-1]:** About the scale argument: In any transport-network, the number of routes an Ingress PE will see is a function of the following parameters:

(a) (NUM_PFX) Number of Egress-nodes originating the transport-prefix
(CT: IP-prefix with Unique-RDs).

Unicast prefixes (announced by pe11 or pe12) – this number will be 1.
Anycast prefix (announced by pe11 or pe12) –
If unique RD used on pe11, pe12: This number value will be 2
If same RD used on pe1, pe12: This number will be 1.

See example topology in section 18
[draft-kaliraj-idr-bgp-classful-transport-planes-17]

(b) (NUM_BN) Number of Border-nodes fanout for ingress-PE in its domain, that the route is traversing with "nexthop-self".

1 - This number will be 2 for this example topology.
Note: Add paths is enabled with nexthop unchanged at RRs.
Not enabled at EBGP-boundary or at nexthop-self BNs.

So the formula is (NUM_PFX * NUM_BN) for this scenario.
And I think that holds good for LU, CT, CAR.

For a unicast prefix, the number of CT routes at ingress-PE pe25 will be 2.
For an anycast prefix, if unique-RD is used, number of CT routes at pe25 will be 4.
if same-RD is used, number of CT routes at pe25 will be 2.

For unicast-prefixes, it is same scale for CT and CAR.

(This is assuming other conditions remain equal like
'addpath-EBGP/addpath-with-nhs is not used'.
For any solution using addpath-send in contiguous domains,
this number will be higher. So we want to use addpath judiciously
and not everywhere).

For anycast-prefixes:
When using same-RD, the scale in CT is same as CAR
(provisioned with same IP:Color on all anycast-sites).

When using unique-RD, the scale in CT is more, but bounded.
It's a function of the variable: "number of anycast-sites".

- [**End of Kaliraj-1 discussion]**

Further points made by DJ on CT draft:

The draft claims that this keeps churn about failure of node E2 from propagating beyond BN1. However, this is not accurate.

Since these are 2 unique routes (with RD2 and RD3), each has its own bestpath and both will be advertised out from BN1 towards E1 by default. This is despite having same local label allocated by BN1 and same next-hop (BN1) being advertised, hence being redundant advertisements that serve no purpose beyond BN1. So, the churn about the failure of either of them will be unnecessarily advertised to E1. And note, with Anycast, there could be routes from many such egress nodes across domains.

The draft does not elaborate whether or how the redundant routes are suppressed.

But now, there are two possible options briefly proposed by CT co-authors in the above Q3 thread.

a) "NV2> RD is flexible enough that for cases like Anycast, where multiple paths are not desired, RD may be set to the same value across routers."

   Configuring same RD is possible. But it contradicts the heavy emphasis CT has placed on advertising unique RDs for troubleshooting and identifying originator. Is that not needed here?

   Also, given that this Anycast IP is intended to originated by routers across different providers (color/admin domains), how do the providers coordinate the configuration of the same RD values on their respective Anycast routers?

b) "NV2> Install a local multipath route for forwarding and originate a single new route. This leaves the local router in control of multipath."

   This is another interesting option. Since there are two or more RD routes that contribute to the multipath for the stripped/imported route multipath, which RD route will be selected to be sent to peers? And what is the selection criteria?

> The statement above uses the term ‘originate’, which could suggest it is a route from the intermediate node, e.g BN1. If so, is that with a local RD?
>
> That again contradicts the emphasis on keeping the original RD. Besides, the co-authors have stated earlier that there is no RD rewrite. The above logic would effectively be an RD rewrite, which also has implications.

[Kaliraj-1 reply to DJ]: And yes, this is another possible option, and the RD will identify the Aggregating BN, that is originating the new route. And please note, the further propagation of RD:AnycastEP routes can be advantageous as-well in following ways: (These are what are not possible with CAR, because it doesn’t support the a.2.i variation above):

- It gives the ingress an idea of how many sites are currently advertising the anycast prefix. If all but one anycast-site are lost, just looking at the route-list can raise an alarm, and operator in ingress-domain can take corrective action to avoid an impending loss of connectivity to the anycast-prefix.
- When per-prefix-label is used (based on RD:EP) instead of per-tc-prefix-label (based on TRDB EP). The ingress can get TE control to direct/ecmp/replicate traffic towards the different anycast sites. In this case the label allocated/carried on these RD:EP CT routes will be unique. And even here, the local failure churn Will Not be propagated beyond local-BN, because per-prefix-label is in use.

[Kaliraj-1-reply]: In essence, our emphasis has been on the fact that RD allows all this flexibility, variations. We expect Unique-RD will be used in most Cases. Same-RD may also be used in some cases based on customer need. We don’t disagree with that, but we do recommend use of unique RDs, for the above reasons. If you got a feeling from reading the draft that only one of the variations MUST be used, that is not the case. We will refine the text in the next version.

[Kalirja-1-reply]: IMO, a TE solution is all about ability to provide better visibility. CT allows for that, while letting [the] user control how much visibility is needed. Thanks for all the discussion.

Questions/Request:
It will be useful if the authors clarify on these options and the operational/protocol considerations we’ve described above.

[Nats-Response (NV3)]: Thanks for raising a separate question as per my request.
[NV3]: This claim is not true. CT allows for multiple fail-safe mechanisms including local repair. Let us explore your use case further based on section 18.4.2
https://www.ietf.org/archive/id/draft-kaliraj-idr-bgp-classful-transport-planes-17.html#name-local-repair-of-primary-pat

[NV3] There are two ways in which operators can choose to allocate label for BGP-CT on SNs and BNs. Again, these are conscious operator choices based on the use-cases the customer is trying to solve.

1. Per-Prefix/Per-Transport-Class (Unique Label for IP1/Transport-Class)
2. Per-Prefix (Unique Label for RDx:IP1)

[NV3] Which the customer can couple with on SNs and BNs

1. Same RD for a transport-class
2. Unique RD for a transport-class

[NV3] The above building blocks provide the operator the flexibility to achieve a broad set of the use-cases (including the ones mentioned above) without getting affected by path selection pinch points, allowing for local-repair at Transit BNs as well as availability of unique paths up until the Ingress "E1" node.

[NV3] If it is not clear in the draft, then we can take an action item to add text to the draft to update this detail on various label allocation and RD usage modes as part of Section 8. I will add this to the presentation slides for BGP-CT as action items requested by Susan for the next 4 months. Once this is clarified in the draft, I hope we can come to a consensus.

[NV3] Thanks again for bringing this out.

F3-CT-Issue-6: Issues to Load in Issue Tracker in github

- Update in CT that provides examples given by discussion and discusses scaling and churn (appendix or main text)
- Above update should include
    - MPLS forward and SR forwarding with single or multiple RDs,
    - Cases when RD rewrite when nexthop is modified at BR.

2: Question on CAR draft:

F3-CAR-Issue-1: BGP-CAR Appendix A.7. Anycast EP Scenario

https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w/

Please find our first question below. Each question will start With a Marker: SUB Q3.CAR-Q<n>: <title>. The format is text in your draft with inline questions.

Text: SUB: Q3.CAR-Q1: BGP-CAR Appendix A.7. Anycast EP Scenario

https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#appendix-A.7

Topology:

A.7. Advertising BGP CAR routes for shared IP addresses

```
       +-------------+      +--------------+
       |             |      |         +----|
       |             |------|         | E2 |(IP1)
       |----+        |      |         +----|
       | E1 |        |      |  Domain 2    |
       |----+        |      +--------------+
       |             |      +--------------+
       |             |      |         +----|
       |  Domain 1   |------|         | E3 |(IP1)
       +-------------+      |         +----|
                            |  Domain 3    |
                            +--------------+
```

Figure 11: BGP CAR advertisements for shared IP addresses

---Snip---

Example-2: Anycast with egress domain visibility at ingress PE

- E2 advertises (IP1, C1) and E3 advertises (IP1, C2) CAR routes for the Anycast IP IP1.
- An ingress PE E1 receives the best path(s) propagated through BGP hops across the network for both (IP1, C1) and (IP1, C2).
- The two CAR routes do not get merged at any intermediate node, providing E1 control over path selection and load-balancing of traffic across these routes.
- Traffic for colored service routes steered at E1 is forwarded to either E2 or E3 (or load-balanced across both) as determined by E1.

[Dhananjaya Rao (DJ) response]: Please also note that there is an example-1 which is for traditional anycast forwarding across domains This example-2 is a case where the operator wants diversity of the egress domains visible at ingress PEs for load-balancing. Hence they assign different colors to different domains.

Questions: (Nats)

With that background:

1. How (IP1, C1) and (IP1,C2) resolve over the same "color" tunnel?
   - **[DJ]:** If both routes need to resolve over a common color in the network that would be achieved by use of Color-ExtComm on these routes, details we have answered in other two questions.
     - **[Nats-2]:** Note: That the operator has to co-ordinate now C1, C2,… CN and the real transport color "C" carried in the color-extended-community. In addition to this, LCM is another variable that may come into play. This complicates the equation further In the BGP-CT operational model, there is no need to generate and manage artificial color values.
       - **[DJ-2]** The operator is making the choice for diversity. It is clear we will not agree on the operational model. Color is a flexible construct, and we don't see a need to restrict its usage.

2. What is the "Color" carried in the Service Route to map to this anycast-NH?
   - **[DJ]:** The service route from each domain would also carry the appropriate color for that domain – e.g. C1 from Domain1.
     - **[Nats2]:** If the Service route carries C1, it breaks anycast since the service can no longer reach Domain 3 which only understands C2.
       - **[DJ-**2]: There will be similarly be a service route from Domain 2 with C2. So, both these routes at ingress PE will form a multipath to enable any load-balancing and recurse onto the CAR routes from Domain2 and Domain3.

3. Observation:
   - **[Nats]** This approach could become a combinatorial explosion burning more colors proportional to number of anycast sites making it [NumColors * NumSites])
     - **[DJ]:** There isn't any combinatorial explosion. The operator has control. In above example, the number of colors would be = #domains offering the Anycast service.
       - **[Nats2]:** For each anycast IP there needs to a per site per SLA color.
         - For ex. In the context of IP1, there are 3 SLAs
         - SLA A = {A1-Domain2, A2-Domain3},
         - SLA B = {B1-Domain2, B2-Domain3},
         - SLA C = {C1-Domain2, C2-Domain3}
       - **[Nats2]** This is NumSLA(C)NumSites = 3(C)2 for the above example which is burning 6 colors per IP.
         - **[DJ-2]:** If there are 3 Anycast services, each with requirement to have diversity across 2 domains, then there will be 6 colors. That is not an issue. The color space is 32-bits. And as described above, the end to end intent can be decoupled from the underlying transport domain intents.
           - What actually matters is the number of transport routes that needs to be distributed from these routers offering the Anycast service. And that number is exactly the same in CAR and CT.
           - However, CAR provides the automatic multipath and redundant paths suppression when there are multiple routes for a color. Whereas in CT, you get the e2e propagation of all individual routes.

- **[Nats]** Not carrying color in NLRI as in BGP-CT avoids this problem.
  - **[DJ]:** It will be good to identify the problem first.
  - **[nats2]:** I hope the above example allows for you to establish clarity.

4. How does ECMP/Protection work for (IP1,C1) & (IP1,C2) at BNs in Domain 1 since they are two unique CAR prefixes?

- **[DJ]:** This example is about a conscious operator choice to assign different colors for each domain. It is logical for the network design to have redundancy and availability at each hop / BN for a color. Just as an example, 2 BNs in Domain-1 connected to 2 BNs in Domain-3.
  - **[Nats2]** This is agreed. However, ECMP/Protection still needs to happen between (IP1,C1) and (IP1,C2) at Ingress Node in E1 to support anycast.
  - **[Nats2]** How does this happen provided the service route carries either C1 or C2 as in your statement above and CAR path selection uses NLRI key?
    - **[DJ-2]** Answered above.

- **[DJ]:** Note though BGP-CT does have a well-understood problem for this scenario. With unique-RD, there is the slow convergence issue ala VPN IAS-OpB. Or if the proposed "RD stripping" multipath procedure is used, then the forwarding diversity is lost, as discussed previously on list.
  - https://mailarchive.ietf.org/arch/msg/idr/q213Pj4nFrFuWOmtyVQpFkqwDW4/
  - **[Nats-2]:** If you have specific questions, please formulate a question that adheres to rules of the Q3 thread. RD allows for all flexibilities as provided by Balaji's response shared below: (https://mailarchive.ietf.org/arch/msg/idr/mZEyLrvb2ooXDvGMryznZBac7BE/)
    - **[DJ-2]** Since you have raised this question, it's only fair to answer to CT's behaviors at the same time. From your comments on CAR above, it's clear you expect the routes from the Anycast nodes in the different egress domains to all have the same Transport-Class. So, at a midpoint like BN of Domain-1, all the routes will merge into a single LSP.
      - Now, even though all these individual routes are going to be advertised onwards towards PE1, they really don't provide any differentiated load-balancing ability or control to the PE, since they all map to a single outgoing label and path.
      - Or alternatively, one can probably not enable the RD-stripping/multipath merge on BN1 through configuration, so there is no merge – but then there is no multipath/protection at any intermediate nodes, and slower convergence.
    - **[DJ-2]** These are suboptimal choices.

  - **[Nats-2]:** 2. I prefer RD to AddPath ID for the following reasons:
    - RD lends itself better to troubleshooting than AddPath ID's, since the originator sets it, and it remains the same across the entire path.
    - RD is flexible enough that for cases like Anycast, where multiple paths are not desired, RD may be set to the same value across routers.
    - I don't believe that usage of RD and VPN-like mechanisms would lead to any noticeable overhead. These mechanisms have been shown to scale to millions of prefixes in the VPN world, and transport routes are usually a fraction in number.
    - I also don't believe that usage of RD affects end-to-end convergence. It, in fact, offers more flexibility in multi-path. With RD, a router has the liberty to do any one of the following:
    - Provide an upstream node with end-to-end visibility of a path by simply propagating the received NLRI with RD intact. This allows the receiving router

to perform multipath as it wishes. It can choose any subset of the paths for multi-path.

- Install a local multipath route for forwarding, and originate a single new route. This leaves the local router in control of multipath.

F3-CAR-Issue-1 Summary: Issues to Load in Issue Tracker in github

- Revision of sections in A.7 and B.2 to address any unclear issues.
    - Since authors believe these two sections are clear, the WG will be queried to determine if anyone feels these two sections are unclear.
- Discussion of CAR view on the use of color to indicate egress domain visibility

F3-CAR-Issue-2: BGP-CAR – Consensus on the need for resolution-schemes

CAR Text: Appendix B. Color Mapping Illustrations

https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#appendix-B

"There are a variety of deployment scenarios that arise w.r.t different color mappings in an inter-domain environment. This section attempts to enumerate them and provide clarity into the usage of the color related protocol constructs."

---Snip---

B.2. Single color domain containing network domains with N:M color

Distribution

- Certain network domains may not be enabled for some of the colors, but may still be required to provide transit.
- When a (E, C) route traverses a domain where color C is not available, the operator may decide to use a different intent of color c that is available in that domain to resolve the next-hop and establish a path through the domain.
  - The next-hop resolution may occur via paths of any intra-domain protocol or even via paths provided by BGP CAR.
  - The next-hop resolution color c may be defined as a local policy at ingress or transit nodes of the domain

[Nats] Note: BGP-CT defines resolution-scheme as a construct to realize this "Local policy" in a well-defined and operator friendly manner. We see the above bulleted item as an agreement from BGP-CAR that such a construct is required in some form or the other. Do you agree?

F3-CAR-Issue-2 Responses:

[Dhananjaya Rao (DJ) reply] 7/20/2022

Email thread: https://mailarchive.ietf.org/arch/msg/idr/QvPiBXj2bJc5SaOIej_Jclpftyk/

[DJ-reply1]: Thanks for the query. BGP CAR leverages already established steering and nexthop resolution techniques for both service steering and for resolution of CAR nexthops. [*Editor: techniques from SR*]

- **[text]:** *[References for techniques]*
  - draft-ietf-spring-segment-routing-policy (section 8.4 and 8.6) has defined Color extended community driven automated steering. It also describes the use of a local policy to control BGP nexthop resolution and steering.
    - Both mechanisms are supported by multiple implementations, in steering BGP services over SR policy/Flex Algo, and to fall back to best effort.
  - The BGP-CAR draft references the terminology in
    - https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#section-1.1

    and describes the usage for CAR nexthop resolution
    - https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#section-2.5
  - and the usage for service steering in

  - https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#section-3.

[DJ-reply1]: We don't agree BGP-CT defines any new construct that is operator-friendly. Rather, we think defining new constructs such as mapping communities adds ambiguity and operational overhead.

F3-CAR-Issue-2: Issues to Load in Issue Tracker in github

- CAR Sections 1.1 needs indicate that local BGP policy can customize or adjust the route validation (section 2.4), route resolution (2.5), and AIGP (2.6).
- Section 2.10 should cover any issues regarding conflicts caused by local policy.

F3-CAR-Issue-3: CAR-Q3- Handling [of] LCM and Color Extended Communities

[IDR mail thread]: https://mailarchive.ietf.org/arch/msg/idr/w5ROKVQPtVcI_BTBXfnKpKB4h4k/

Reference: https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#appendix-B

Text:

Appendix B. Color Mapping Illustrations

"There are a variety of deployment scenarios that arise w.r.t
different color mappings in an inter-domain environment. This
section attempts to enumerate them and provide clarity into the usage
of the color related protocol constructs."

---Snip---

B.2. Single color domain containing network domains with N:M color
distribution

---Snip---

* It may also be automatically signaled from egress border nodes
by attaching a color extended community with value c to the BGP
CAR routes.

o Hence, routes of N colors may be resolved via a smaller set of M
colored paths in a transit domain, while preserving the original
color-awareness end-to-end.

[Nats] For the above bulleted item,
1. Why is yet another method required in presence of LCM?
2. Is the Color carried in Color Extended Community different from the one carried in LCM (if any)?
3. How does this Color Extended Community interact with LCM and color carried in NLRI?
4. If LCM and Color Extended Community are handled differently, those procedures are under-specified. Conversely, if it is handled in the same way, this seems redundant.

Note:

- It is confirmed that the CAR route can have only one LCM. Does the same rule apply for the set of mapping-communities (LCM, Color Extended Community)?
- BGP-CT defines mapping-community as the abstraction to represent any such community. Do you agree that BGP needs a logical mapping-community construct such that such rules can be specified against the same?

F3-CAR-Q3 Responses:

Response 1: Dhananjaya response (CAR author)

Mail thread: https://mailarchive.ietf.org/arch/msg/idr/AKSQrklNV2BYR9uQXdQYvDr2fm8/

[DJ-reply]: Firstly, LCM Extended Community and Color Extended Community are for different purposes and should not be mixed. Using draft-hr-spring-intentaware-routing-using-color for reference, there are 2 distinct requirements to be supported:

2) Domains with different intent granularity (section 6.3.1.9)
3) Network Domains under different administration (section 4.1)

- **[Kaliraj-reply1-to DJ]:** What happens when there is scenario with both requirements 1 'and' 2?
  - Ref: https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#appendix-B.3

    "When the routes are distributed between domains with different color- to-intent mapping schemes, both N:N and N:M cases are possible, although an N:M mapping is more likely to occur."

    Your first statement above seems to contradict with possibility of N:M scenario mentioned in appendix-B.3 of the CAR draft.

    - **[DJ-reply-2]:** The two are not mutually exclusive. I just stated that the two requirements are distinct, did not say they are mutually exclusive. The statement from Appendix-B.3 you pasted above already states both are possible.

[DJ-reply1-continued] Requirement 1 is the case where within the same administrative or color domain, BGP CAR routes for N end-to-end intents may need to traverse across an intermediate domain where only M intents are available, N >= M.

- Example: Multi-domain network is designed as Access-Core-Access. The core may have the most granular N intents, whereas the access only has fewer M intents. So, the BGP nexthop resolution for a CAR route must be via a color-aware path for one of these M intents in the access domain. For this case, LCM-EC is not applicable. Instead, as described in the CAR draft (https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#appendix-B.2). Color Ext-comm is used to automate the resolution.

[DJ-reply1-continued] For requirement 2, where CAR routes traverse across different color domains, LCM-EC is used as already described in Section 2.8 and Appendix.B.3 of draft-dskc-bess-bgp-car. I Hope this clarifies the difference.

- **[Kaliraj-reply1-to DJ]:** The procedures described in appendix-B.3 assume N:N. How is N:M handled, with multiple color domains? Nats' questions [in original question] are related to that scenario, which need to be answered.
  - **[DJ-reply-2]:** The color in LCM-EC represents the e2e NLRI intent in a different color domain. When traversing a N:M transit within that color domain, the Color Ext-Comm is also attached and used for nexthop resolution, as per Appendix B.2 described above.

F3-CAR-Issue-3: Issues to Load in Issue Tracker in github

- Appendix B.2 should be clarified after F3-CAR-Issue-2 has been expanded to include:
    - CAR Sections 1.1 needs indicate that local BGP policy can customize or adjust the route validation (section 2.4), route resolution (2.5), and AIGP (2.6).
    - Section 2.10 should cover any issues regarding conflicts caused by local policy.

F3-CAR-Q4: BGP-CAR – Mis-Routing in Non-agreeing color-domains for Anycast EPs

Reference: https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#appendix-B.3

B.3. Multiple color domains

"When the routes are distributed between domains with different color- to-intent mapping schemes, both N:N and N:M cases are possible, although an N:M mapping is more likely to occur."

Reference topology:

```
D1 ----- D2 ----- D3
C1        C2       C3
```

- C1 in D1 maps to C2 in D2 and to C3 in D3
- BGP CAR is enabled in all three color domains

The reference topology above is used to elaborate on the design described in Section 2.8.

When the route originates in color domain D1 and gets advertised to a different color domain D2, following procedures apply:

- The original intent in the BGP CAR route is preserved; i.e. route is (E, C1)
- A BR of D1 attaches LCM-EC with value C1 when advertising to a BR in D2
- A BR in D2 receiving (E, C1) maps C1 in received LCM-EC to local color, say C2
  - A BR in D2 may receive (E, C1) from multiple D1 BRs which provide equal cost or primary/backup paths
- Within D2, this LCM-EC value of C2 is used instead of the Color in CAR route NLRI (E, C1). This applies to all procedures described in the earlier section for a single color domain, such as next-hop resolution and service steering.

[Nats]: In the anycast scenario,

```
        SLA1=200,SLA2=100   SLA1=100,SLA2=200
D1 ----- D2 --------------------- D3
         |
         |
         D4
        SLA1=200,SLA2=100
```

Observation: ECMP/Protection is non-deterministic

```
    @D2
       (IP1,100), LCM=200 (from D3)
       (IP1,100)          (from D4)

       (IP1,200), LCM=100 (from D3)
       (IP1,200)          (from D4)
```

The Effective Resolution Key when LCM is in play will be (E, LCM).
However, ECMP/Protection is based on path selection of NLRI key
as mentioned in BGP-CAR Section 2.7.
(https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#section-2.7)

Problem:
The traffic for SLA1 can be mis-routed to SLA2 (or vice versa) at D2

Note: CT defines the Transport Route DB construct to collect transport routes pertaining to an associated transport-class. Path Selection works on this repository in the context of its corresponding color resulting in non-ambiguous ECMP/protection.

Response:
[DJ]: As you indicated [in the report above] … the same case where an Anycast service has been coordinated between two different administrative domains to be operated as a shared service. A common Anycast IP address is being used for the purpose.

- As we'd indicted earlier, our expectation is that the administrative domains at the same time also coordinate the use of an unused color in each domain to be used along with the Anycast IP. [This] removes any possibility of misrouting. Operationally, this is identical to not using the same IP that's been used for Anycast, for some other purpose within the two administrative domains.

- The CT approach of importing the routes into a local color "VRF" to achieve multipath leads to the highly suboptimal behavior where all these unique RD routes from the different Anycast BRs are still redundantly propagated downstream all the way to all the ingress PEs in all domains. This causes not only unnecessary increase in control plane state on all downstream BGP nodes, but also requires them to deal with the churn of all path changes to all of these unique RD routes. I had referred to this in a previous message to the list.

[For my earlier response, please see]

- [Please see]: https://mailarchive.ietf.org/arch/msg/idr/R-ockSQHvDojiaEu__0gLBd0xBk/

The CAR design is intentional with the purpose of not incurring any of this runtime overhead over the lifetime of these or any transport routes.

[Shunwan]: Within D2, since the LCM-EC value of the Color is used instead of the Color in CAR route NLRI, so (E, LCM-EC Color) of the CAR route should be used for Path Selection. If I understand the case correctly, this will solve the mis-routing problem that may exist.

[Nats]: Perhaps you missed this part. ECMP/Protection is based on path selection using (E, C) as described in section 2.7. (even when LCM is in use).

https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-04#section-2.7

> "The (E, C) route inherently provides availability of redundant paths at every hop. For instance, BGP CAR routes originated by two egress ABRs in a domain are advertised as multiple paths to ingress ABRs in the domain, where they become equal-cost or primary-backup paths. A failure of an egress ABR is detected and handled by ingress ABRs locally within the domain for faster convergence, without any necessity to propagate the event to upstream nodes for traffic restoration.
>
> BGP ADD-PATH should be enabled for BGP CAR to signal multiple next hops through a transport RR."

[Nats]: Note that BGP-CT use path selection based on "E" in the context of the Transport Route Database (Transport Route-Target). This is similar to the "path selection based on (E, LCM) or essentially (E, Effective Color)" model that you are proposing. That's why it works in BGP-CT.

- **[Shunwan-reply-2]:** Thanks for your pointer! I'll re-read the relevant sections.

F3-CAR-Issue-4: Issues to Load in Issue Tracker in github

- Clarify paragraph 2 in Section 10 to include assumptions regarding coordination of shared ANYCAST service used across multiple color domains.
- Link revised paragraph 2 in section 10 to Appendix A.7
- Revise Appendix A.7 (or create a new) to specifically detail how an ANYCAST Address will operate.

F3-CAR-Q5: Update Packing Observations

IDR mail link: https://mailarchive.ietf.org/arch/msg/idr/xlOuRX4WSK3s6UNMEILoamYk1_A/
Author: Natrajan Venkataraman natv@juniper.net (nats)
Date: 7/19/2022

F3-CAR-Q5 Introduction:

As per the BGP-CAR draft section 2.9.2.3:

https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#section-2.9.2.3

---Snip---

- SRv6 SID Information: field of size as indicated by the length that either carries the SRv6 SID(s) for the advertised color-aware route as one of the following:
    - A single 128-bit SRv6 SID or a stack of 128-bit SRv6 SIDs
    - A transposed portion (refer [I-D.ietf-bess-srv6-services]) of the SRv6 SID that MUST be of size in multiples of one octet and less than 16.

---Snip---

The BGP color-aware route update for SRv6 MUST include the BGP
Prefix-SID attribute along with the TLV carrying the SRv6 SID information as specified in [I-D.ietf-bess-srv6-services] when using the transposition scheme of encoding for packing efficiency of BGP updates.

The Prefix-SID attribute is mandatory even when SRv6 SID Information
is carried as part of the BGP-CAR NLRI

F3-CAR-Q5: Observations (Nats)

1. The update size increases by 14 bytes per CAR-NLRI when compared with CT-NLRI
2. This is due to duplication of SRv6 SID Information per NLRI
3. The packing efficiency is also the same since Prefix-SID attribute iss in use

- **[swadesh-reply-1]:** HI Nats: Thanks for the query. Please find my response to specific points inline. But you have raised question about how update packing efficiency claimed is achieved. So it’s worth mentioning that BGP-CAR NLRI design maintains update packing benefits in multiple ways listed below:
    1. Optimizes BGP update packing by carrying per-prefix information as part of NLRI. e.g. label-index is carried in NLRI. This allows packing multiple NLRIs in single update message. Whereas in BGP-CT, label-index is carried in an attribute.
    2. Allows to signal multiple per-prefix encapsulation types & values as part of NLRI. e.g. MPLS Label, SRv6 SID and any other encapsulations. CT can only signal a single label field
- **[Swadesh-reply]** The CT SAFI also suffers from a related robustness issue that was raised in IDR about overloading the MPLS label field to carry SID information : https://mailarchive.ietf.org/arch/msg/idr/R98YR27K31ZlbdDvsbMX1f-liTY/

F3-CAR-Q5: Observations with responses

1. The update size increases by 14 bytes per CAR-NLRI when compared with CT-NLRI
   - **[Swadesh (SA) reply]** CAR draft provides flexibility to signal variable part of SRv6 SID in NLRI and rest in prefix SID attribute.
     - Thus does not increase NLRI size by 14 bytes. More importantly, please also note that it enables packing multiple NLRI in same update message and avoids repetition of attributes per NLRI.
     - Attribute (mostly size > 80 bytes) repetition is worse and that is the case in CT with per prefix SRv6 SID carried in Prefix SID attribute. In case transposing scheme is planned to be used by CT, it has the same robustness issue found with existing SAFIs.
     - Please refer https://mailarchive.ietf.org/arch/msg/idr/R98YR27K31ZlbdDvsbMX1f-liTY/
2. This is due to duplication of SRv6 SID Information per NLRI
   - **[Swadesh (SA) reply]** Its not duplicate information. Its variable part of per prefix SRv6 SID.

3. The packing efficiency is also the same since Prefix-SID attribute [is] in use
   - **[Swadesh (SA) reply]** There is no update packing in CT with per prefix SRv6 SID contained in Prefix-SID attribute. If transposition scheme is used, it suffer with robustness issue mentioned above.

F3-CAR-Q5 Question (with responses)

- How does this improve the CAR packing efficiency as claimed?
  [Swadesh (SA) reply]: Answered in the top post and response to point 1, 2 and 3.

NLRI and ATTR comparison between CAR and CT for SRv6 use case

```
--------CAR-NLRI-START----------
             0                   1                   2                   3
          0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         |  NLRI Length  |  Key Length   |   NLRI Type   |Prefix Length  |
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         |               IP Prefix (variable)                          //
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         |               Color (4 octets)                                |
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+


       Total Bytes: 12 bytes

      SRv6 SID TLV - CAR format (without EP behaviors)
          0                   1                   2                   3
          0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         |      Type     |    Length     |   SRv6 SID Info (variable)  //
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

      Total Bytes: 18 bytes (1 SID)
```

```
--------CAR-NLRI-END-----------
```

[Nats]: Conclusion:
Total Size consumed in BGP Update =
(30 bytes * Num NLRI) + sizeof (Prefix-SID-ATTR-srv6)

[Swadesh(SA) reply]: SRv6 SID TLV will carry variable part of SRv6 SID (typically 2-4 bytes) and that make total size per NLRI as 16-18 bytes and not 30.

- Also Prefix-SID attribute will not have variable part SRv6 SID and hence can be shared across NLRIs providing update packing.

```
--------CT-NLRI-START--------
             0                   1                   2                   3
             0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
            +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
            |    Length     |                 Label                 |Rsrv |S|
            +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
            ~                     Route Distinguisher (8 bytes)              |
            |                                                                |
            +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
            |                     IPv4/IPv6 Prefix                           ~
            |                                                                |
            +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

      Total Bytes: SAFI-76 + IPv4 = 16 bytes
--------CT-NLRI-END--------
```

[Nats]: Total Size consumed in BGP Update =
(16 bytes * Num NLRI) + sizeof (Prefix-SID-ATTR-srv6)

[Swadesh (SA) reply]: CAR carries variable part of SRv6 SID as information in NLRI and hence prefix SID attribute is shared across NLRIs. That is not the case with BGP CT with per prefix SRv6 SID in Prefix SID attribute. If transposition scheme is used by CT, it suffer[s] with robustness issue mentioned in top post and response to point

```
"Prefix-SID attribute is common for CT and CAR"
------PREFIX SID ATTR START -------
      SRv6 SID Sub-TLV
          0                   1                   2                   3
          0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         | SRv6 Service  |    SRv6 Service               |                |
         | Sub-TLV       |    Sub-TLV                    |                |
         | Type=1        |    Length                     |  RESERVED1     |
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         |  SRv6 SID Value (16 octets)                                   //
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         | Svc SID Flags |   SRv6 Endpoint Behavior      |   RESERVED2    |
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         |  SRv6 Service Data Sub-Sub-TLVs                               //
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
      SRv6 SID data sub-sub-TLV

          0                   1                   2                   3
          0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         | SRv6 Service  |    SRv6 Service               | Locator Block |
         | Data Sub-Sub  |    Data Sub-Sub-TLV           | Length        |
         | -TLV Type=1   |    Length                     |               |
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         | Locator Node  | Function      | Argument      | Transposition |
         | Length        | Length        | Length        | Length        |
         +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
         | Transposition |
         | Offset        |
         +-+-+-+-+-+-+-+-+

------PREFIX SID ATTR END -------
```

F3-CAR-Issue-5: Issues to Load in Issue Tracker in github

- Section 6 on Scaling needs to be expanded to include:
  - Bytes added to BGP UPDATE message for CAR NLRI with SRv6
    - A single 128-bit SRv6 SID or a stack of 128-bit SRv6 SIDs
    - A transposed portion (refer [I-D.ietf-bess-srv6-services]) of the SRv6 SID that MUST be of size in multiples of one octet and less than 16.
  - Compression allowed due to signal multiple per-prefix encapsulation types & values as part of NLRI. e.g. MPLS Label, SRv6 SID and any other encapsulations.

- Section 2.9.2.3 needs to be upgraded to point to examples in Appendices of carrying single 128-bit SRv6 SID and Stack of SRv6-SIDs. The examples in the appendices should also reference back to scaling in section 6.

3: WG questions

F3-BOTH-Issue-1: New Address Families [Shunwan Zhuang]

IDR mail thread: https://mailarchive.ietf.org/arch/msg/idr/4T3-b4_ckpGu3BwjwuESqpYsoFk/

[Question]: Both drafts introduce new BGP address families. Should we consider the compatibility and incremental deployment for the new BGP address family extensions in multi-domain networks?

For example, the new address family is only supported in the first and the third domain, while the transit domain does not support it.

If yes, can the authors please describe how to do it?

Responses

[CT-Response (Nats)]: https://mailarchive.ietf.org/arch/msg/idr/mf09fpqZUlejoO00N928ZMn4H2I/

Thanks for your question. Please let us know if you need more clarification.
SUB- Q3-CT-ANSWER-1: Passing through domains that does not support BGP-CT

To answer your question let us consider the following topology highlighting Control and Data plane flows.

Let us assume two SLAs, namely
Gold=100, Bronze=200
TRT:0:100 = Transport-Class route-target for Gold SLA
TRT:0:200 = Transport-Class route-target for Bronze SLA

NOTE: The below model preserves End-To-End SLA using per SLA Tunnel Endpoint in the BGP-LU domain (AS2). However, this is not a requirement and using a single loopback makes all SLA to use the same intradomain transport tunnel in AS2.

```
                +-------EBGP-MHOP-CT--------+
                |                           |
  [PE3--------ASBR3]--[ASBR22----ASBR21]--[ASBR1------PE1]
  <-----BGP-CT----->  <-----BGP-LU----->  <----BGP-CT---->
          AS3                 AS2                AS1
         --------Packet Forwarding Direction------->
```

Control Plane Flow: SAFI-4/SAFI-76 Transport
From PE1 to ASBR1
(CT-1)   <--RD1:PE1,TRT:0:100
                             PNH=PE1,Label: CT-L0
                          <--RD2:PE1,TRT:0:200

```
                                   PNH=PE1,Label: CT-L1
```

From ASBR1 to ASBR3 (via EBGP Multihop for transport layer)

```
(CT-2)      <---------------------------RD1:PE1,TRT:0:100
                                          PNH=ASBR1,Label: CT-L2
           <---------------------------RD2:PE1,TRT:0:200
                                         PNH=ASBR1,Label: CT-L3
```

From ASBR21 to ASBR22

```
(LU-3)                      <--ASBR1/32,
                               PNH=ASBR21-LPBK-Gold, Label: LU-L0, Comm: Gold
                           <--ASBR1/32,
                              PNH=ASBR21-LPBK-Bronze, Label: LU-L1, Comm: Bronze
```

From ASBR22 to ASBR3

```
(LU-4)          <--ASBR1/32,
                   PNH=ASBR22,Label: LU-L2, Comm: Gold
                 <--ASBR1/32,
                    PNH=ASBR22,Label: LU-L3, Comm: Bronze
```

From ASBR3 to PE3

```
(CT-5) <--RD1:PE1,TRT:0:100
           PNH=PE1,Label: CT-L4
      <--RD2:PE1,TRT:0:200
          PNH=PE1,Label: CT-L5
```

Route processing at ASBR3:

1. Install BGP-LU route for ASBR1 to Gold or Bronze
   Transport Class Route DB based on Comm: Gold or Bronze
2. RD1:PE1 Route with TRT:0:100 resolve over ASBR1/32 Route in Gold TRDB
3. RD2:PE1 Route with TRT:0:200 resolve over ASBR1/32 Route in Bronze TRDB
4. Advertise BGP-CT routes to PE3

ASBR3 Gold TRDB:

ASBR1/32   PUSH LU-L2, via ASBR3-ASBR22-Link

ASBR3 Bronze TRDB:

ASBR1/32   PUSH LU-L3, via ASBR3-ASBR22-Link

ASBR3 MPLS FIB:

Gold SLA:

CT-L4   SWAP CT-L2, PUSH LU-L2, via ASBR3-ASBR22-Link

Bronze SLA:

CT-L5   SWAP CT-L3, PUSH LU-L3, via ASBR3-ASBR22-Link

Rest of the BGP-CT nodes behave as described in Section 18. of the BGP-CT Draft. https://www.ietf.org/archive/id/draft-kaliraj-idr-bgp-classful-transport-planes-17.html#name-illustration-of-procedures-

[Shunwan-reply-2]: https://mailarchive.ietf.org/arch/msg/idr/tTFM-R55GCBZWyd-nQsPjnc7HVU/
Thank you for your detailed explanation. If I understand correctly, IBGP-LU Add-Path needs to be supported between ASBR22 and ASBR21, and EBGP-LU Add-Path needs to be supported between ASBR3 and ASBR22. EBGP-LU Add-Path deployment may be a big challenge.

Response 2: DJ on CAR
[DJ-reply]: https://mailarchive.ietf.org/arch/msg/idr/mY6aPqkJ2PE_1Vmylv35Us57Rhw/

[DJ: Thank you for your question. The BGP-CAR draft has an example for this scenario. Could you please take a look ?

- https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#appendix-A.4

[Shunwan-reply-2] Hi DJ - Thanks for the guidance! It does have an example in the BGP-CAR draft. I think that it works for incremental deployments.

F3-WG-Issue-1: Issues to Load in Issue Tracker in github

[Both]: Provide an example of incremental deployment in domains 1, 2, and 3. Suppose that only domains 1 and 3 have been enhanced to

- CAR/CT/WG – agree that CAR A.4 contains is common incremental example,
- CT - Add example of incremental deployment to draft.
- CAR – enhance the A.4 example based on lists discussion.

F3-Both-Issue-2: Support for SRv6

[Author] Jingrong Xie **[xiejingrong@huawei.com]**

IDR mail link: https://mailarchive.ietf.org/arch/msg/idr/7C7dlvIgZuNNx3rLorC6S24Ta50/

Both the drafts say they support SRv6 as well as MPLS data plane, but neither one has clear illustration how it may support. Can the authors please provide an illustration of SRv6 data plane (e.g., E2E SRv6 & intra-domain SRv6) based on a sample topology?

My feeling is that the difference in the encapsulation of SRv6 and SR-MPLS/MPLS may need to be considered and may have impact on the choice.

[The following is a sample topology I am used to understand multi-domain network

where BR is ASBR, [AC1-P1-BR1] is an IGP domain,

[BR3-BR4] is an IGP domain, and

[BR2-P2-AC2] is an IGP domain)

[AC1----P1----BR1]----[BR3----BR4]----[BR2----P2 ----AC2]

Thank you very much.

Jingrong

Follow-up by Jingrong:

- I have received the response from Rajesh about the SRv6 data-plane of CT, very much appreciated.
- I am still waiting for response and discussion from CAR about the detail of SRv6 data-plane support.

At the meantime, we detailed the thoughts and submitted a draft on this: https://datatracker.ietf.org/doc/draft-wang-idr-cpr/

Hope it is helpful to the discussion about the SRv6 data-plane support for e2e intent aware paths.

Responses:
[Rajesh-CT-reply]: https://mailarchive.ietf.org/arch/msg/idr/7C7dlvIgZuNNx3rLorC6S24Ta50/

Thanks for bringing [this] valid point.

- Regarding CT , this has been well explained in below draft at "section 5.1. Option C Transport Interworking".
    - https://datatracker.ietf.org/doc/draft-salih-spring-srv6-inter-domain-sids/
- The example shows IBGP-CT connection between border routers in each domain and single hop EBGP-CT for inter-domain connections.
- Please note, CT can support any srv6 requirement.

[Jingrong-reply]: https://mailarchive.ietf.org/arch/msg/idr/8Zqs8TgQctmg61V30l7oEi8Vf-A/

Thank you very much for providing the reading guidance. I have read the related draft <draft-salih-spring-srv6-inter-domain-sids>, and the following is my understanding of the data plane:

(use the simple topology in the draft section 5.1 and 3.1)

{[1]----[2]----[4]}......{[6]----[8]----[10]}......{[12]----[15]----[16]}

1->2: (S=1, D=2)(16.DT4, B:4:REP::1, SL=2) (Customer-pkt)

2->4: (S=1, D=B:4:REP::1)(16.DT4, B:4:REP::1, SL=1) (Customer-pkt)

4->6: (S=1, D=B:6:REP6::1)(16.DT4, B:4:REP::1, SL=1) (Customer-pkt)

6->8: (S=6, D=8)(10, SL=1), (S=1, D=B:10:REP::1)(16.DT4, B:4:REP::1, SL=1) (Customer-pkt)

8->10: (S=6, D=10)(10, SL=0), (S=1, D=B:10:REP::1)(16.DT4, B:4:REP::1, SL=1) (Customer-pkt)

10->12: (S=1, D=B:12:REP6::1)(16.DT4, B:4:REP::1, SL=1) (Customer-pkt)

12->14: (S=12, D=14)(B:16:END::1, SL=1), (S=1, D= B:16:END::1)(16.DT4, B:4:REP::1, SL=1) (Customer-pkt)

14->16: (S=12, D= B:16:END::1)(B:16:END::1, SL=0), (S=1, D= B:16:END::1)(16.DT4, B:4:REP::1, SL=1) (Customer-pkt)

Representation remark:

(1) The red text is encap/decap procedure.

(2) The SRH is in "Reduced variant" mode as defined in RFC8754 section 4.1.

(3) the representation of SRH is using the representation of RFC8754 section 6.2.

Please let me know if the above understanding is correct.

[CAR-response-DJ]: https://mailarchive.ietf.org/arch/msg/idr/-NdYWDgco_9sQpNwqBXSjpmz0-M/

Thank you for your question, apologies for the delay in getting to it.

- BGP-CAR SAFI is designed to support multiple transports consistently, and the current version describes the encoding mechanisms for SRv6.
- As you probably gauged from the illustrations in the current version, the BGP-CAR control/dataplane flows for MPLS maintain consistency with the model established and deployed with BGP-LU. The approach would be similar for SRv6, where CAR leverages base SRv6 routing mechanisms already described in SRv6-overlay (RFC9252) and SRv6-MPLS interworking draft (draft-agrawal-spring-srv6-mpls-interworking).

However, there are additional options and operational considerations that do need to be described. We plan to include them along with illustrations in the next version of the draft. Thank you for the draft reference. We will also review it.

[Jiangrong-reply-2]: I Look forward to the next revision of the car draft as well as the review comments.

F3-WG-Issue-2: Issues to Load in Issue Tracker in github

- CAR/CT: Provide an illustration of SRv6 data plane (e.g., E2E SRv6 & intra-domain SRv6) based on a sample topology?:
- CAR: (DJ) Add additional options and operational considerations that do need to be described. We plan to include them along with illustrations in the next version of the draft.

F3-Both-Q3: Key Operational differences between the CAR and CT drafts (Bruno)

Author: Bruno Decraene

IDR mail thread: https://mailarchive.ietf.org/arch/msg/idr/-N9CncTl8JtwDLmGZEJ1RLqzMSM/

[Editor’s Note: Bruno and the immediate CT/CAR authors are presented below. Other members of the IDR WG are presented in the next section. Bruno’s text is base text in the next section.]

F3-Both-Q3-1: [Bruno] Introduction:

I think that a key point is the NLRI key, i.e. the Network Layer Reachability Information, aka the reachability that we are interested in. I’m assuming that it is clear that the reachability that we want to advertise is (EndPoint, Color). IOW, we want (at least) a path to each (EndPoint, Color).

BGP CAR uses exactly that key:

“NLRI Key: IP Prefix, Color” https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#section-2.1

```
“    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |  NLRI Length  |  Key Length   |   NLRI Type   |Prefix Length  |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               IP Prefix (variable)                           //
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |               Color (4 octets)                                |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+ »
```

https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#section-2.9.2

BGP CT uses a different key, namely (RD, EndPoint):

“When AFI/SAFI is 1/76, the Classful Transport NLRI Prefix consists of an 8-byte RD followed by an IPv4 prefix.”

[Section 7.0]

```
“     0                   1                   2                   3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |    Length     |                 Label                 |Rsrv |S|
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     ~                     Route Distinguisher (8 bytes)             |
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
     |                     IPv4/IPv6 Prefix                          ~
     |                                                               |
     +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

         Fig 2: SAFI 76 "Classful Transport" NLRI”
```

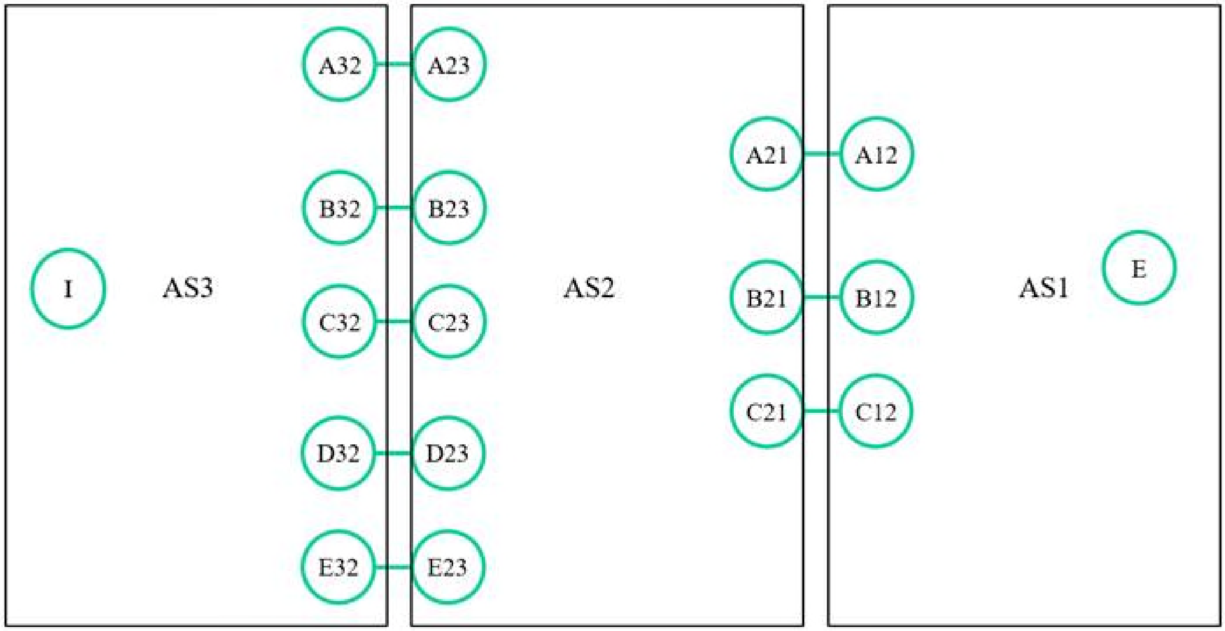
```
“Route Distinguisher:

     8 byte RD as defined in [RFC4364 Sec 4.2].”
```

https://www.ietf.org/archive/id/draft-kaliraj-idr-bgp-classful-transport-planes-17.html#name-bgp-classful-transport-fami

- **[Kaliraj]** yes, and the color is carried in the RT. [RDs] don't carry any meaning, just a distinguisher. So path-selection happens for just 'EP' in context of a Color (Transport class Route DB).
  - "When AFI/SAFI is 1/76, the Classful Transport NLRI Prefix consists of an 8-byte RD followed by an IPv4 prefix." (CT – section x.x)

F3-Both-Q3-2 [Bruno] Provide a sample topology.



F3-Both-Q3 [Bruno Decraene] Discuss the pros/cons clearly stating whether your post is a: pro, con, or clarifying question.

BGP CAR uses the key/NLRI that we do want to propagate: (EndPoint, Color). That's a good fit.

BGP CT uses, as key, (RD, Endpoint), with RD as defined in RFC4364.

Two observations:

a) (RD, Endpoint) is not the NLRI that we want to reach. We want to reach (EndPoint, color)

- **[Kaliraj-reply]:** I feel like you may not be considering Anycast EP? Please see following threads that discuss possibility of Mis-routing, ECMP, and Color management problems with Anycast-EP deployments, when using Color in NLRI.
  - https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w/
  - https://mailarchive.ietf.org/arch/msg/idr/OOZOBSyjdAYBar8NxvOqo6-5fAc/

- **[Kaliraj-reply]:** EP with Transport-Target:0:<color> gives that info. RD is just the messenger, [that disambiguates] (disambiguator) in BGP updates.

b) In VPN/RFC4364, the purpose of the RD is to _*distinguish*_ EndPoint (IP prefix) because they are not unique across VPNs. That is not needed for the BGP color use case because EndPoint are Public (or a minimum agreed upon by a set of consenting adults)

- **[Kaliraj-reply]:** Even here, an EP has a different personality in the context of a transport-layer color. E.g. in your topology, the reachability info to reach EP 'E' via a certain Color Gold is different from the reachability info to reach same EP via a different color Bronze. So though EP E is a provider-space 'public' IP-address, it has a different per color persona/instance, which need to be distinguished in a BGP update. CAR uses color to disambiguate those instances in a BGP-update, CT uses RD.

"a" and "b" are two arguments against using (RD, EndPoint) as the NLRI.key.

There has been some argument that RD may be useful to advertise multiple path for a destination (Endpoint, Color) or even to identify the source of the advertisement.

- This is not inline with the definition of RD / RFC4364 which states

  "An RD is simply a number, and it does not contain any inherent information; it does not identify the origin of the route or the set of VPNs to which the route is to be distributed. The purpose of the RD is solely to allow one to create distinct routes to a common IPv4 address prefix." https://datatracker.ietf.org/doc/html/rfc4364#section-4.1)

  - **[Kaliraj-reply]:** RD does not identify destination or source of route, for route leaking purposes. But _when_ unique RDs are in in use, it does aid in troubleshooting. That is what we meant. We do recommend using **unique-RDs, but not mandatory**. Same-RDs MAY be used when path-information need to be filtered out. See CT section 10.9.
    - https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-10.9
  - **[Kalirja-reply]:** Deploying unique RDs is strongly RECOMMENDED because it helps in troubleshooting by uniquely identifying originator of a route and avoids path-hiding. [The] following email from Balaji nicely summarizes the benefits of using RD :
    - https://mailarchive.ietf.org/arch/msg/idr/mZEyLrvb2ooXDvGMryznZBac7BE/

- That RD purpose is absolutely not needed in the use case that we are discussing, because the IP address prefix is unique/has a single meaning. So really no need to re-add a RD with public name space, which would typically contain …. an IP address in most discussions (and possibly the same IP address of the EndPoint if the route is sourced by the destination…).
  - **[Kaliraj-reply]:** Consider Anycast use-cases, where the same EP may be used on multiple nodes, so not unique.

- So in summary, the NLRI.key proposed in BGP CT (RD, EndPoint) is not a good fit. While the NLRI.key proposed in BGP CAR (EndPoint, Color) is the right fit.
  - **[Kaliraj-reply]:** Please see thread referred to in beginning on this email that describe problems because of color is carried in NLRI.

Going further, let’s assume that we have a RD field because… we have one. What could be its use?

- One proposal was to advertise the source. But this explicitly contradicts RD definition in RFC4364 (cf above). This is also not the typical operational model, both for Internet and VPN, where the source is typically indicated using a community or extended community (e.g., site of origin).
    - **[Kaliraj-reply]:** Please see the threads referred to in beginning on this email [thread?] that describe problems because of color is carried in NLRI.
        - https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w/
        - https://mailarchive.ietf.org/arch/msg/idr/OOZOBSyjdAYBar8NxvOqo6-5fAc/
    - **[Kaliraj-reply]:** It is good to hear consensus that things like SOO Community are in common use, which affect the update packing arguments.

- One proposal was to advertise multiple paths for the same destination. However:
    - RD(s) is(are) chosen by the originator ie. the egress domain. There is no reason that this AS choose the right number of paths as best fists all others (ingress, transit) ASes. E.g. in above figure, AS1 may advertise 3 RD/paths, while may be AS3 would be willing to have 5 RD/paths.
        - **[Kaliraj-reply]:** The RDs chosen by egress-domain are propagated as-is by transit domain to ingress. This model does not need additional Colors in those transport-networks. The Egress may originate RD per service-function or stats-group. I agree that these functions will be confined to egress-domain, as you note.
            - [For Example] for the ‘per-ASBR statistics usecase’ in topology above, node E can use three RDs to get stats of traffic coming from three ASBRs (A12, B12, C12) in its domain, irrespective of which ASBRs in other domains this traffic traversed.
            - Another example is: different service-funcions (SF1, SF2) can be attached to the node E, such that RD1:E, RD2:E can be used to advertise these service functions with different UHP labels in BGP-CT, while using same Transport-class RT. So provisioning new colored tunnels in the transport network is not required.
        - **[Kaliraj-reply**] In CAR, for such usecases, operator will have the provisioning and management overhead of using distinct Colors (because Color is the distinguisher in CAR NLRI), and will have to co-ordinate to resolve them over the ‘same SLA tunnel’ in the transport-layer. And again, using multiple colors to represent same SLA prevents ECMP/Protection between the CAR NLRIs. (See https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w/)

    - It’s very possible that the number of RD be not even chosen by the egress AS. Based on Seamless MPLS work, I would assume that the route be either originated by the egress PE (in which case a single route/RD) is used, or by the ASBRs of the egress AS (in which case the number of route/RD is the number of ASBR of this AS, e.g. 3 in the above figure, irrespectively of other considerations).
- If a distinguisher field would be needed for some deployments, a priori I would personally not be opposed to have (EndPoint, Color, distinguisher) with distinguisher being a simple unsigned

integer field. That been said, with a 32-bits color field, I feel like we already have plenty of room without adding a distinguisher.

Other Discussion Threads (Outside of Bruno and CAR/CT authors)

Thread 1: Jeff Haas/Bruno Discussion

IDR mail thread: https://mailarchive.ietf.org/arch/msg/idr/XR61gdzPKNNnbr322f_th-wxGIE/

[Jeff-reply-1]: Thanks for the detailed mail and good questions. I had one clarification to add as we wait for the -CT authors to respond to your various points:

- In your topology, above, presume the color domains are not identical and remapping is required. In -CAR, this would be via LCM.
    - Presume that in AS1, the advertised endpoint is (E, C1).
    - Presume that at AS3, the received endpoint is (E, C1) LCM C2.

    [Bruno-reply-1]: OK

- **[Jeff-reply-1]:** Presumably in AS3, operators would be more concerned with the operations of C2 rather than remote C1?
    - **[Bruno-reply-1]:** I'm very sorry, but I'm not seeing what you have in mind. Could you please elaborate?
    - **[Bruno-reply-1]:** In the meantime, although it's possibly not the subject, I'll try to elaborate. Assume that this color C1 is for "low delay" "intent". What we want to reach is (E, low delay). We need that NRLI to be propagated. If there are three "intents" to reach E (e.g. low delay, encrypted, high bandwidth) we need three NLRIs. That's what we get by encoding the intent in the NLRI. In CAR the intent is encoded as a color. (with a bijection)
        - **[Jeff-reply-2]:** agreed.
    - **[Bruno-reply-1]** I'm not sure to see what would not work if that color number be overwritten by the LCM: we still have one NLRI per intent.
        - **[Jeff-reply-2]:** But that "low delay" intent varies per domain and may not share the same colors for it.
            - **[Bruno reply-2]:** Not sharing the same color does not seem an issue to me. Let's assume that BGP-CAR LCM community is always added. In this case, this seems exactly like BGP-CT Transport Class ID, no? (unless I'm missing something in BGP-CT.
                - **[Jeff-reply-3]:** The question is partially motivated by that observation.
            - **[Bruno reply-2]:** Not sharing the same intent across ASes seems like a bigger issue. Intent needs to match, at least roughly, along the path. (unless this is agreed to be "best effort"... which is probably already the case for "low delay" 😉 (low delay may means different delays between different competing ASes) Seems a priori like similar to CoS: codepoint may be different but per hop behavior needs to be roughly aligned.
                - **[Jeff reply-3]** And similar to CoS troubleshooting, being able to locally determine the intent needed on receipt along with the needed code point further downstream to get the desired behavior.

- **[Bruno-reply-1]:** Trying another angle, my email focuses on the NLRI.key. I believe we need one NLRI.key per (EndPoint, Intent) so that BGP propagates a path for each (EndPoint, Intent). To me, this is rather orthogonal to where is carried the "color" used for the indirection/route resolution. If the WG wants to carry that color in a community (e.g. an always present LCM or Route Target Extended community) I don't think that this change[s] my email on the NLRI.key.

- **[Jeff-reply-1]:** My question is operationally whether the receiver of an NLRI from a remote color domain cares more about "original intent" or about "local intent". If locally 100 is my "low delay", anything remote will require you to do some sort of assisted lookup to figure out what that intent was.
  - **[Bruno-reply-2]:** I'd say local. But very likely I'm missing the issue that you have in mind.
    - **[Jeff reply-3**]: No, I think you're confirming it.
  - **[Bruno-reply-2]:** Not sure what you mean by "assisted lookup". One need to look in the community (LCM or TC). Assuming the use of LCM is mandated in all cases, this point seems very similar between CAR and CT.
    - **[Jeff**-reply-3] This is effectively what I'm trying to tease out of the operational picture. The operational need, that I believe you're confirming, is you care at any given BGP Speaker what the local intent happens to be. For -CAR, it can move in the update depending on whether LCM is in use or not. Yet having the "original intent" seems to be the core selling point of the -CAR encoding.
    - **[Jeff-reply-3]:** Note that I'm over a year from trying to change someone's mind on the preference. Protocol-wise, aside from issues with route selection pinch points and error handling considerations, the protocol mechanisms are largely identical. However, I'm still trying to understand the operators' mental model vs. how things operationally must behave in a multi-color-domain environment.
    - **[Jeff-reply-3]:** thanks for entertaining my question.

Thread 2: Kaliraj Feedback on Anycast EP

IDR mail thread: https://mailarchive.ietf.org/arch/msg/idr/OIpprDKR_BjN-qQmUnM4TMEjcXQ/

[Kaliraj-reply1]: Hi Bruno, I feel like you may not be considering Anycast EP? Please see following threads that discuss possibility of Mis-routing, ECMP, and Color management problems with Anycast-EP deployments, when using Color in NLRI.

> https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w/
> https://mailarchive.ietf.org/arch/msg/idr/OOZOBSyjdAYBar8NxvOqo6-5fAc/

- **[Bruno-reply]:** - without "EP": my reading is that implicitly, you means that Anycast (without "EP") worst just fine with BGP CAR. So let's make this explicit.
  - **[Kaliraj**-2]: No. I meant 'Anycast Endpoint'. I meant, CAR can mis-route for 'Anycast Endpoints' in non-agreeing color domains.
  - - with "EP": I'm not completely certain what "EP" means. From BGP-CT, "EP : End point, a loopback address in the network.". Anycast EP is not self-explicit to me, but from

below email, I'll assume that it's "anycast" but not really "any" i.e. Ingress wants visibility of each path and be able to select the one it want.

- **[Kalirjaa-2]:** Anycast EP is an ‘Anycast IP-address’, that would be used as nexthop in BGP service route updates.
- **[Kaliraj-2]:** The mis-routing observation is for non-agreeing color domains, using anycast endpoints. Please see following threads that discuss possibility of Mis-routing, ECMP, and Color management problems with Anycast-EP deployments, when using Color in NLRI. ***(good reference missing)***
- https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w

- **[Bruno-reply]** Honestly, I'm not seeing any "Mis-routing, ECMP, and color management problems". I'm seeing some questions been asked. I don't feel that my email and your above thread are discussion the same point:
  - ***My original email*** is on NLRI.key.
    - One want (at least) one path per (Intent, EndPoint). Hence IMO as per BGP (4271) BGP rule, the natural encoding of the NLRI.key is to include (EndPoint, Intent).
  - ***Your thread*** seems to be about color resolution/indirection between the service route and the transport route.
    - On this point, my co-authors would be much more competent to reply. A priori, it seems to me that the use of Local-Color-Mapping (LCM) Extended Community addresses your point (mapping to a third color if needed; which represent the intent as seen by the service). And, speaking for myself only, if you/the WG really want to have this resolution color always encoded in a community, one could possibly always attach this community (up to mandating this in the draft is needed). That seems like an optimization/minor point to me (i.e. nothing fundamental). https://mailarchive.ietf.org/arch/msg/idr/OOZOBSyjdAYBar8NxvOqo6-5fAc/
    - **[Kaliraj-2]:** It is the second email-thread that talks about Misrouting. Pasting link again here: https://mailarchive.ietf.org/arch/msg/idr/yI9y1iik3hO-dATSrvi4ST4K9CY/
    - **[Kaliraj-2]** Shunwan seems to understand the problem. And DJ as-well. All DJ is saying is the color needs to be coordinated, and when doing so, the color-management problem appears for the ‘improving visibility to ingress’ case.
    - **[Kaliraj-2]** They are related. I’ll try to connect the dots. Please read on. [in this thread]

- **[Bruno-reply]:** IP anycast, by definition, mandates coordination between the endpoints (domains) to agree on the IP to use. So coordination is granted and IMO they should also coordinate on the color to use. (if it were me, I'd say that the "owner" of the IP address is the one choosing/allocating the color to use). With that, I think that the problem does not exist.
  - **[Kaliraj-**2]: The scenario being discussed here is administrative-domains which don’t have one Color-namespace across all the domains. Sure, the problem can be solved in different places. CT solves it in the protocol. In CAR operators coordinate that problem doesn’t happen. All customer networks may not be able to reach such an agreement, to have single color namespace. Especially across network mergers.

        - **[Bruno-**2]: One color-name space is not required across all the domains. What is required is the use of a common color for the IP anycast address advertised by all egress ASes. But specifically for anycast, regardless of the new coloring feature, coordination _is_ (already) required for the selection of this IP address. So coordination there is and this can be extended to the selection of color.
            - **Jeff reply to Bruno-**2]: To distill this one point from your conversation with Kaliraj … ***If this is required for correct forwarding in such scenarios***, it's important that this get noted in the [CAR] document's procedures. Seems like a good Appendix entry.
            - **Robert reply to Bruno-**2] Not sure I follow this. If I have PI space what is there to coordinate if I want to advertise /24 from that space via multiple upstream ASNs and treat it as anycast service address block?
                - **[Bruno-**3]: No problem for this case. In your use case, you are the only one advertising that /24 (i.e. a single AS advertise the anycast address). I'm assuming that you will agree with yourself and for a given intent (e.g. low delay) you will select and advertise the same color to all upstream ASNs. On my side, I was considering the most complicated case where ISP X and ISP Y were offering the same service over the same anycast address. So two different domains which need to coordinate.
                    - **[Robert-**2]: Well, I thought the discussion is not about how I advertise my prefix to immediately connected ISP. I was under impression the discussion is how those ISPs will pass it to their peers and upstreams. Of course assuming they are all willing to play the rainbow game.

- **[Bruno-**reply]: **note** that one could probably find a similar example with BGP-CT. With RD type 1, the administrator field is an IP address. An easy choice seems to use the IP address of the endpoint. In which case, it seems to me that with BGP-CT one could have (derived from your email)

```
      @D2
          (01:IP1:O1:IP1), TCID=200 (from D3)
          (01:IP1:O1:IP1), TCID=100 (from D4)

         (01:IP1:O2:IP1), TCID=100 (from D3)
          (01:IP1:O2:IP1), TCID=200 (from D4)
```

    - Note: 01: IP1:01:IP is RD:Endpoint and more precisely "RD type":" Administrator subfield": Assigned Number subfield:EndPoint IP address
- **[Bruno-reply]** With that BGP-CT have the same problem (same NRLI.index for different intents). In both cases, the root cause is that the owner of IP1 needs to coordinate/maintain the sub allocations spaces (color for BGP-CAR, Assigned Number subfield for BGP-CT)

    - **[Kalirja-2]** So in the example (non-agreeing color domains),

```
@D2
(IP1,100), LCM=200 (from D3)
(IP1,100)             (from D4)
```

- **[Kaliraj-2]:** The two routes indicate different 'effective-intent', even though they have same Color value in the NLRI. So path-selection uses these two routes to do ECMP/protection.
  - Ref : https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-04#section-2.7
- **[Kaliraj-2]:** Nope. CT doesn't have a problem. Because the endpoint routes are organized in a Transport-Route-DB, which is per TC. So, we will have :

```
 TC100 TRDB :
        (IP1), (from D4)
        (IP1), (from D3)
TC200 TRDB :
        (IP1), (from D3)
        (IP1), (from D4)
```

- **[Kaliraj-2]** So the path selection that happens in the TRDB (context of TCID) does not mix different SLAs, irrespective of what the NLRI key contains.

[Responding to Bruno's above comment:] *Your thread seems to be about color resolution/indirection between the service route and the transport route.*

- **[Kalirja-2]** It is about the resolution of the Transport-route(CAR) over the correct intent. because the 'effective intent' of these paths is not the same (one of them has an LCM) they resolve over different intents. But path-selection NLRI.key does not take this into account. So traffic intended for SLA 100 can get routed into SLA 200, or the reverse. Hence, mis-routing is possible.

- **[Bruno-reply]:** On this point, my co-authors would be much more competent to reply. A priori, it seems to me that the use of Local-Color-Mapping (LCM) Extended Community addresses your point (mapping to a third color if needed; which represent the intent as seen by the service). And, speaking for myself only, if you/the WG really want to have this resolution color always encoded in a community, one could possibly always attach this community (up to mandating this in the draft is needed). That seems like an optimization/minor point to me (i.e. nothing fundamental)
  - **[Kaliraj-2]**: Kaliraj raises his concerns regarding Anycast or Non-agreeing Color-domains handling in CAR. (per https://mailarchive.ietf.org/arch/msg/idr/OOZOBSyjdAYBar8NxvOqo6-5fAc)

F3-WG-Issue-3a: Issues to Load in Issue Tracker in github

- CAR/CT – Create common example for shared ANYCAST Service across multiple domains. Bruno will be asked to create topologies for these services.
- CAR/CT - Based on common example, add text to draft.
  - Note: This action item should be added to CAR github repository with a note that this issue overlaps with F3-CAR-Issue-4.

Thread 3: Robert-Kaliraj Feedback on Anycast EP

Quick reference: https://mailarchive.ietf.org/arch/msg/idr/OIpprDKR_BjN-qQmUnM4TMEjcXQ/

Kaliraj's comment (from above): In CAR, for such use cases, operator will have the provisioning and management overhead of using distinct Colors (because Color is the distinguisher in CAR NLRI), and will have to co-ordinate to resolve them over the 'same SLA tunnel' in the transport-layer. And again, using multiple colors to represent same SLA prevents ECMP/Protection between the CAR NLRIs. [For reference see: https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w/)

- **[Robert Raszuk-reply**]: I think you are now mixing mapping with transport. Why [can] different service-functions [not] be mapped to the same color? Why do you say that CAR would need to use multiple colors? Again it is my understanding that mapping flows to color-ed transport is something completely orthogonal to establishing such transport itself and should not be mixed.
    - **[Kaliraj-replies-to-Robert]:** Because CAR NLRI is Color:IP, unless different color is used, updates will not pass thru path-selection pinch points. Using different IP-addresses defeats the purpose of not having to use different loopbacks (that can be done with LU itself). That's the disadvantage of overloading the same field (Color in NLRI) to carry the color and also act as a disambiguator. This is what RFC-4364 solves nicely by using RD and RT, which BGP-CT also adopts.
        - **[Robert-r2]:** Sorry but you have missed my point. If the goal is to use the same color across different service functions. I do not see any reason to send multiple transport signallings one per each service function. Color aware transport should be decoupled from direct service relation. Indirection and hierarchy is here at your disposal. **[editorial note:** Robert's email text is unclear. Transport signalling" probably means transport signalling actions/messages. "Tranport Signallings one per service function is also unclear".]
            - **Aijun:** There maybe some slices of the network share the same color. Don't you think using "RD:Node" (CT Solutions)to differentiate these slices and use "transport class"/color to group them will be more flexible than encoding the color directly into the NLRI(CAR).
            - [Robert-r3]: What I am saying that transport should not carry service irrespective of the solution. Indirection is the key here to map any service to any colors as it seems fit by a specific ingress. (final text comment removed).

F3-WG-Issue-3b: Issues to Load in Issue Tracker in github

- CT: Clarify what type of indirection and hierarchy in CT provides scaling in Section 14 as a high-level introduction to the section.
- CAR: Section 2.5 provides two comments on route resolution that need to be clarified:
    - "When multiple resolutions are possible, the default preference should be: IGP Flex-Algo, SR Policy, RSVP-TE, BGP Car, [and] BGP-LU."
        - This description uses the word should which implies that local policy can interfere. This should be clarified.
        - This description does not include the inclusion of LCM or Extended-Color Community or Color in the Tunnel Attribute.
    - "Resolution may be automated using Color-EC as illustrated in Appendix B.2." This comment does not provide a normative set of results for route resolution.

- CAR/CT – Should include discussion on impact on anycast endpoints, non-agreeing color-domains.

[Kaliraj-reply2]: Sub: Q3-CT-ANSWER-2: Clarification to Bruno on BGP-CT NLRI format.
[IDR mail thread: https://mailarchive.ietf.org/arch/msg/idr/Q7c38zA24CkFMWyLks5GKMl-HbY/]

Moving the following discussion:

https://mailarchive.ietf.org/arch/msg/idr/Qvq9ij1WUpCUkGwN4cuDADBtfNI/

to this Part3 email-thread, as suggested by Susan.

To answer your question:

- **[Kaliraj-reply2-Part1]** Yes: BGP-CT follows RFC-8277, and advertises a single label, when "Multiple Labels Capability" is not negotiated. We don't foresee the need to use Multiple Labels Capability with BGP-CT.
  - **[Bruno-reply-1]** It's interesting that you mention a BGP capability to advertise multiple labels. RFC3107 did not require one. RFC8277 was obliged to define one because the same AFI/SAFI than RFC3107 is used and there is a need to distinguish between implementation compliant with 3107 (i.e. accepting a label stack) from implementation pretending to be compliant (not accepting a stack). The situation is different for BGP-CT as a new AFI/SAFI is defined hence there is a priori no need to define two NLRI encodings (stack vs single label) and a capability. Why do we need this extra complexity?
    - **[Kaliraj-reply-2]:** Correct Bruno. By following procedures defined in RFC-8277, BGP-CT can advertise, receive label-stack in the NLRI.
    - **[Kaliraj-reply-2]** About the capability, yes. [Because] CT follows RFC-8277, it uses the “Multiple Labels Capability” to indicate this support. This capability specifies a “Count”: (see [RFC8277] https://datatracker.ietf.org/doc/html/rfc8277#section-2.1)

      > “If one of the triples is <AFI, SAFI, Count>, the Count is the maximum number of labels that the BGP speaker sending the Capability can process in a received UPDATE of the specified AFI/SAFI.”

    - **[Kaliraj-reply-2]** If BGP-CT doesn’t use the capability and implicitly assumes it, it cannot have this control on ‘Count’. Authors of RFC-8277 may have put this ‘Count’ after careful considerations of various implementations. So I think BGP-CT may benefit from abiding by those rules.

    - **[Kaliraj-reply-2]:** As a side note, that is somewhat related to this discussion: I see there is a picture being painted by some that RFC-8277 and RFC-4364 are ancient and should not be followed, extended.
      - It is my opinion that RFC-8277 and RFC-4364 are one of the most extensible, generic, stable and time-tested networking design-patterns IDR and BESS WGs have created.
      - So it is good to build on top of them, re-using what is good in them, and fixing what is not. Instead of abandoning them and starting from scratch

to define a brand new set of mechanisms, that are specific to one set of technologies, and don't cater to some known deployment scenarios.

[Kaliraj-reply-1- Part2]: So, Section 7 of the BGP-CT draft illustrates the BGP-CT NLRI with single-label.

https://www.ietf.org/archive/id/draft-kaliraj-idr-bgp-classful-transport-planes-17.html#section-7

see: "For better readability, the following figure illustrates a BGP Classful Transport family NLRI when single Label is advertised:"

- Procedures with respect to whether single-label or label-stack is encoded is same as SAFI-4 or SAFI-128.
- Procedures with respect to encoding Route-Distinguisher:Prefix is same as SAFI-128.
- BGP-CT is nothing but "BGP-LU with RD and RT". A VPN-ized BGP-LU Transport.
    - **[Bruno-reply-1]** "BGP-LU" allows advertising multiple labels from day one (both RFC3107 and 8277). So I understand that BGP-CT allows the advertisement of multi labels in the NLRI. Please correct me if I'm wrong.
    - **[Kaliraj-reply]:** See https://mailarchive.ietf.org/arch/msg/idr/7sjUuTVP9-ejnZe4VjVNthjAWAU/
      I feel like you may not be considering Anycast EP.
      [Bruno-reply2]:
      - without "EP": my reading is that implicitly, you means that Anycast (without "EP") worst just fine with BGP CAR. So let's make this explicit.
      - with "EP": I'm not completely certain what "EP" means. From BGP-CT,
          - "EP : End point, a loopback address in the network.".
          - Anycast EP is not self-explicit to me, but from below email, I'll assume that it's "anycast" but not really "any" i.e. Ingress wants visibility of each path and be able to select the one it want.

[Kaliraj]: Please see following threads that discuss possibility of Mis-routing, ECMP, and Color management problems with Anycast-EP deployments,
when using Color in NLRI.

- https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w/
- https://mailarchive.ietf.org/arch/msg/idr/OOZOBSyjdAYBar8NxvOqo6-5fAc/

  [Bruno-reply-2]: Honestly, I'm not seeing any "Mis-routing, ECMP, and color management problems". I'm seeing some questions been asked.
    - [Kaliraj-2 (KV2)]: It is the second email-thread that talks about Misrouting. Pasting link again here:
        - https://mailarchive.ietf.org/arch/msg/idr/yI9y1iik3hO-dATSrvi4ST4K9CY/
    - [Kaliraj-2] Shunwan seems to understand the problem. And DJ as-well. All DJ is saying is the color needs to be coordinated, and when doing so, the color-management problem appears for the 'improving visibility to ingress' case.

  [Bruno-reply-2]: I don't feel that my email and your above thread are discussion the same point:
    - [Kaliraj (KV2)]: They are related. I'll try to connect the dots.

[Bruno-reply-2] My original email is on NLRI.key. One want (at least) one path per (Intent, EndPoint). Hence IMO as per BGP (4271) BGP rule, the natural encoding of the NLRI.key is to include (EndPoint, Intent).

- [Kaliraj (KV2)]: So in the example (non agreeing color domains),

  @D2

  (IP1,100), LCM=200 (from D3)

  (IP1,100) (from D4)

- **[kaliraj (KV2)]:** The two routes indicate different 'effective-intent', even though they have same Color value in the NLRI. So path-selection uses these two routes to do ECMP/protection.
  - Ref : https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-04#section-2.7

- **[Bruno-reply-2]** Your thread seems to be about color resolution/indirection between the service route and the transport route.
    - **[Kaliraj (Kv2)]:** It is about the resolution of the Transport-route (CAR) over the correct intent because the 'effective intent' of these paths is not the same (one of them has an LCM) they resolve over different intents. But path-selection NLRI.key does not take this into account. So traffic intended for SLA 100 can get routed into SLA 200, or the reverse. Hence, mis-routing is possible.
  - **[Bruno-reply-2]** On this point, my co-authors would be much more competent to reply. A priori, it seems to me that the use of Local-Color-Mapping (LCM) Extended Community addresses your point (mapping to a third color if needed; which represent the intent as seen by the service).
  - **[Bruno-reply-2]** And, speaking for myself only, if you/the WG really want to have this resolution color always encoded in a community, one could possibly always attach this community (up to mandating this in the draft is needed). That seems like an optimization/minor point to me (i.e. nothing fundamental)
    - **[Kaliraj (KV2)]:** From my analysis of CAR so far, I *do not see the procedures specified do involve Anycast or Non-agreeing Color-domains. It may be best to describe these as Caveats in the CAR draft.*

      (Editorial note: text in italics is an editorial summation with some text removed.)
    - **[Kaliraj (KV2)]:** BGP-CT handles all these cases without exception because it follows RFC-4364 procedures.

- **[Bruno-reply-2**]: IP anycast, by definition, mandates coordination between the endpoints (domains) to agree on the IP to use. So coordination is granted and IMO they should also coordinate on the color to use. (if it were me, I'd say that the "owner" of the IP address is the one choosing/allocating the color to use). With that, I think that the problem does not exist.
  - **[Kaliraj (KV2)]**: The scenario being discussed here is administrative-domains which don't have one Color-namespace across all the domains.
    - Sure, the problem can be solved in different places. CT solves it in the protocol. In CAR operators coordinate that problem doesn't happen.
    - All customer networks may not be able to reach such an agreement, to have single color namespace. Especially across network mergers.
  - **[Bruno-2 (reply-3)]** One color-name space is not required across all the domains. What is required is the use of a common color for the IP anycast address advertised by all egress ASes.
  - **[Bruno-2 (reply-3)]:** But specifically for anycast, regardless of the new coloring feature, coordination _is_ (already) required for the selection of this IP address. So coordination there is and this can be extended to the selection of color.

        - **[Jeff-Haas]:** If this is required for correct forwarding in such scenarios, it's important that this get noted in the document's procedures. Seems like a good Appendix entry.
        - **[Robert Raszuk]:** “Not sure I follow this. If I have PI space what is there to coordinate if I want to advertise /24 from that space via multiple upstream ASNs and treat it as anycast service address block?
            - **[Bruno-reply**]: No problem for this case. In your use case, you are the only one advertising that /24 (i.e. a single AS advertise the anycast address). I'm assuming that you will agree with yourself and for a given intent (e.g. low delay) you will select and advertise the same color to all upstream ASNs.
            - **[Brun**o-reply] On my side, I was considering the most complicated case where ISP X and ISP Y were offering the same service over the same anycast address. So two different domains which need to coordinate.
                - **[Robert-reply2]**: Well I thought the discussion is not about how I advertise my prefix to immediately connected ISP. I was under impression the discussion is how those ISPs will pass it to their peers and upstreams. Of course assuming they are all willing to play the rainbow game.

- **[Bruno-reply-2]** note that one could probably find a similar example with BGP-CT. With RD type 1, the administrator field is an IP address. An easy choice seems to use the IP address of the endpoint. In which case, it seems to me that with BGP-CT one could have (derived from your email)
    - Example
      @D2
      (01:IP1:O1:IP1), TCID=200 (from D3)
      (01:IP1:O1:IP1), TCID=100 (from D4)

      (01:IP1:O2:IP1), TCID=100 (from D3)
      (01:IP1:O2:IP1), TCID=200 (from D4)

    - **Note:** 01:IP1:01:IP is RD:Endpoint and more precisely "RD type":" Administrator subfield": Assigned Number subfield:EndPoint IP address

- **[Bruno-reply-2]:** With that BGP-CT have the same problem (same NRLI.index for different intents). In both cases, the root cause is that the owner of IP1 needs to coordinate/maintain the sub allocations spaces (color for BGP-CAR, Assigned Number subfield for BGP-CT)
    - **[Kaliraj (KV**-2)] : Nope. CT doesn’t have a problem. Because the endpoint routes are organized in a Transport-Route-DB, which is per TC.
    - **[KV**-2] So we will have:
      TC100 TRDB :
      (IP1), (from D4)
      (IP1), (from D3)
      TC200 TRDB :
      (IP1), (from D3)
      (IP1), (from D4)

      So the path selection that happens in the TRDB (context of TCID) does not mix different SLAs, irrespective of what the NLRI key contains.

F3-WG-Issue-3c: Issues to Load in Issue Tracker in github

[Kaliraj and Bruno add issues here]

- CAR: Add Discussion on Non-agreeing color-domains for Anycast endpoints to error handling and manageability section (section 10). This issue overlaps with F3-CAR-Issue-4 and F3-Wg-Issue-3a.
- CT: Add description of procedures when IP address has collision, or with same RD. (See Bruno's comments for details).

[per discussion above – Kaliraj has inline comments from https://mailarchive.ietf.org/arch/msg/idr/7sjUuTVP9-ejnZe4VjVNthjAWAU/

[Bruno] I'm assuming that it is clear that the reachability that we want to advertise is (EndPoint, Color). IOW, we want (at least) a path to each (EndPoint, Color).

Part 1: [Bruno]

- BGP CAR uses exactly that key:
- BGP CT uses a different key, namely (RD, EndPoint):
    - **[Kaliraj]:** yes, and the color is carried in the RT. RD [do not] carry any meaning, just a distinguisher. So path-selection happens for just 'EP' in context of a Color (Transport class Route DB).

Part 2: [Bruno]

- BGP CAR uses the key/NLRI that we do want to propagate: (EndPoint, Color). That's a good fit.
- BGP CT uses, as key, (RD, Endpoint), with RD as defined in RFC4364.
    - a) (RD, Endpoint) is not the NLRI that we want to reach. We want to reach (EndPoint, color)
        - **[Kaliraj]:** EP with Transport-Target:0:<color> gives that info. RD is just the messenger, disambiguator BGP updates.
    - b) In VPN/RFC4364, the purpose of the RD is to _distinguish_ EndPoint (IP prefix) because they are not unique across VPNs. That is not needed for the BGP color use case because EndPoint are Public (or a minimum agreed upon by a set of consenting adults)
        - **[Kaliraj]:** Even here, an EP has a different personality in context of a transport-layer color.
            - E.g. in your topology, the reachability info to reach EP 'E' via a certain Color Gold is different from the reachability info to reach same EP via a different color Bronze.
            - So though EP E is a provider-space 'public' IP-address, it has a different per color persona/instance, which need to be distinguished in a BGP update. CAR uses color to disambiguate those instances in a BGP-update, CT uses RD
    - **[Bruno]** "a" and "b" are two arguments against using (RD, EndPoint) as the NLRI.key.

- **[Bruno]** There has been some argument that RD may be useful to advertise multiple path for a destination (Endpoint, Color) or even to identify the source of the advertisement.
    - This is not inline with the definition of RD / RFC4364 which states
        - "*An RD is simply a number, and it does not contain any inherent information; it does not identify the origin of the route or the set of VPNs to which the route is to be distributed. The purpose of the RD is solely to allow one to create distinct routes to a common IPv4 address prefix.*"
    - **[Kaliraj]** RD does not identify destination or source of route, for route leaking purposes. But _when_ unique RDs are in use, it does aid in troubleshooting. That is what we meant. We do recommend using unique-RDs, but not mandatory. Same-RDs MAY be used when path-information need to be filtered out.
        - See: https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-10.9
        - Text: "Deploying unique RDs is strongly RECOMMENDED because it helps in troubleshooting by uniquely identifying originator of a route and avoids path-hiding."
    - **[Kaliraj]** The following email from Baliaji nicely summarize the benefits of using RD:
        - https://mailarchive.ietf.org/arch/msg/idr/mZEyLrvb2ooXDvGMryznZBac7BE/

- **[Bruno]:** That RD purpose is absolutely not needed in the use case that we are discussing, because the IP address prefix is unique/has a single meaning.
    - **[Kaliraj]:** Consider Anycast use-cases, where the same EP may be used on multiple nodes, so not unique.

- **[Bruno]:** So really no need to re-add a RD with public name space, which would typically contain …. an IP address in most discussions (and possibly the same IP address of the EndPoint if the route is sourced by the destination…).
- **[Bruno]:** So in summary, the NLRI.key proposed in BGP CT (RD, EndPoint) is not a good fit. While the NLRI.key proposed in BGP CAR (EndPoint, Color) is the right fit.
    - **[Kaliraj]:** Please see threads referred to in beginning on this email that describe problems because of color is carried in NLRI.

- **[Bruno]** Going further, let's assume that we have a RD field because… we have one. What could be its use?
    - **[Kaliraj]:** It is good to hear consensus that things like SOO Community are in common use, which affect the update packing arguments.

- **[Bruno]:** One proposal was to advertise the source. But this explicitly contradicts RD definition in RFC4364 (cf above). This is also not the typical operational model, both for Internet and VPN, where the source is typically indicated using a community or extended community (e.g., site of origin).
    - **[Kaliraj]:** The RDs chosen by egress-domain are propagated as-is by transit domain to ingress.
        - This model does not need additional Colors in those transport-networks. The Egress may originate RD per service-function or stats-group.
        - I agree that these functions will be confined to egress-domain, as you note. E.g. for the 'per-ASBR statistics use case' in topology above, node E can use three

RDs to get stats of traffic coming from three ASBRs (A12, B12, C12) in its domain, irrespective of which ASBRs in other domains this traffic traversed.

- [**Kaliraj]:** Another example is : different [service-functions] (SF1, SF2) can be attached to the node E, such that RD1:E, RD2:E can be KV> used to advertise these service functions with different UHP labels in BGP-CT, while using same Transport-class RT. So provisioning new colored tunnels in the transport network is not required.

- [**Kaliraj]:** In CAR, for such use cases, operator will have the provisioning and management overhead of using distinct Colors (because Color is the distinguisher in CAR NLRI), and will have to co-ordinate to resolve them over the 'same SLA tunnel' in the transport-layer. And again, using multiple colors to represent same SLA prevents ECMP/Protection between the CAR NLRIs. (see reference below)
  - https://mailarchive.ietf.org/arch/msg/idr/nAj25sX0x_lp09VEqUDSCxmDR_w/

- **[Robert] (responding to Kaliraj]:** I think you are now mixing mapping with transport.
  - Why different service-functions [cannot] be mapped to the same color?
  - Why do you say that CAR would need to use multiple colors?
  - Again it is my understanding that mapping flows to color-ed transport is something completely orthogonal to establishing such transport itself and should not be mixed.
  - **[Kaliraj reply]:** Because CAR NLRI is Color:IP, unless different color is used, updates will not pass thru path-selection pinch points. Using different IP-addresses defeats the purpose of not having to use different loopbacks (that can be done with LU itself).
    - That's the disadvantage of overloading the same field (Color in NLRI) to carry the color and also act as a disambiguator.
    - This is what RFC-4364 solves nicely by using RD and RT, which BGP-CT also adopts.
  - **[Robert]:** Sorry but you have missed my point. If the goal is to use the same color across different service functions I do not see any reason to send multiple transport signaling one per each service function. Color aware transport should be decoupled from direct service relation. Indirection and hierarchy is here at your disposal.
    - **[Aijun Wang]:** There maybe some slices of the network share the same color. Don't you think using "RD:Node" (CT Solutions)to differentiate these slices and use "transport class"/color to group them will be more flexible than encoding the color directly into the NLRI(CAR).
      - **[Robert]:** What I am saying that transport should not carry service irrespective of the solution. Indirection is the key here to map any service to any colors as it seems fit by a specific ingress. If any proposal carries say slide_id directly in the transport signalling I would dismiss it.

F3-Both-Q3-all: Issues to add to Issue Tracker in github

This section is a summary of the issue raised above:

- CAR/CT – Create common example for shared ANYCAST Service across multiple domains. Bruno and Jeff Haas will be asked to create topologies for these services.
- CAR/CT - Based on common example, add text to draft.
  - Note: This action item should be added to CAR github repository with a note that this issue overlaps with F3-CAR-Issue-4.
- CT: In the introduction to Section 14, clarify what paradigm is used for scaling in CT. (e.g. indirection and hierarchy). The text should include the following:
  - descriptions of how CT's paradigms for indirection and hierarchy are used in the network's transport and service topology,
  - descriptions should include unique RDs and same RDs in stable and changing topologies (e.g. routing churn).
- CAR: In the introduction to section 6, clarify what paradigm is used for scaling in CAR (e.g., indirection and hierarchy). This text should include the following:
  - descriptions of how CAR's paradigms for indirection and hierarchy are used in the network's transport and service topology,
  - how scaling is impacted by NLRI changes are handled in route withdraws, refreshes, and updates.
- CAR: Section 2.5 provides two comments on route resolution that need to be clarified:
  - "When multiple resolutions are possible, the default preference should be: IGP Flex-Algo, SR Policy, RSVP-TE, BGP Car, [and] BGP-LU."
    - This description uses the word should which implies that local policy can interfere. This should be clarified.
    - This description does not include the inclusion of LCM or Extended-Color Community or Color in the Tunnel Attribute.
  - "Resolution may be automated using Color-EC as illustrated in Appendix B.2." This comment does not provide a normative set of results for route resolution.
- CAR/CT – Should include discussion on impact on anycast endpoints, non-agreeing color-domains.
- CAR: Add Discussion on Non-agreeing color-domains for Anycast endpoints to error handling and manageability section (section 10). This issue overlaps with F3-CAR-Issue-4 and F3-Wg-Issue-3a.
- CT: Add description of procedures when IP address has collision, or with same RD. (See Bruno's comments for details).

F3-Both-Issue-4: Intent at the Service (e.g. VPN Layer)

Originator: Ketan Talaulikar:

IDR thread link: https://mailarchive.ietf.org/arch/msg/idr/hHto6CYV6zWeTju7gHWLH1qRsOA/

Description: The BGP-CAR solution set also includes support for extending color (i.e. "intent") awareness into the VPN service layer as well such that end customers can indicate more than a single SLA for their prefix to their service providers. E.g. a local site with /24 IPv4 would have hosts running different applications requiring different SLAs. The resolution happens similarly to the BGP CAR transport - i.e., based on color.

- **[Aijun Wang]:** Why don't use the color extended communities to achieve the same effect for these services?

Text: More details can be found at:

- https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-problem-statement-05#section-3.2
- https://datatracker.ietf.org/doc/html/draft-dskc-bess-bgp-car-05#section-8

[Editor's note: draft-hr-spring-intentaware-routing-using-color-00 has replaced draft-dskc-bess-bgp-car-problem-statement].

Therefore, the BGP-CAR solution goes beyond addressing the transport requirement alone and also naturally extends and integrates into the service layer as well.

- **[Aijun Wang]:** In my opinions, VPN solutions in CAR just extend the transport slice to the customer sites. It has no relation with the service layer. The intent or color of one service should always be expressed in the "color extend communities"

Note: I have not seen any solution from the proponents of the BGP-CT draft for this allied problem space.

Question: Is this not another reason to believe that the two solutions are not really "functionally" identical?

Specific topology: (missing)

Response: (need to query for response.) Spring – document issue

F3-Both-Issue-4: Issues to add to Issue tracker in github

- CT: Add a section to discuss how color is implement in the VPN service layer.
- CT: Add a definition of intent that aligns with Spring and other IETF/IRTF WGs
- CAR: Add a section to discuss how color is implemented in the VPN service layer
- CAR: Clarify the definition of intent to align with Spring and other IETF/IRTF WGs.

F4-Both-Issue-5: Technology BGP-CT and CAR are based on and Implications

[Jeffrey Zhang]: Whether BGP-CT/CAR are based on VPN or BGP-LU and which one is better to go forward.

[text]: Following section explains the relationship and distinction between SAFI 76 and SAFI 4, SAFI 128. https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-9

- "SAFI 128 (Inet-VPN) is an RFC8277<https://datatracker.ietf.org/doc/html/rfc8277> encoded family that carries service prefixes in the NLRI, where the prefixes come from the customer namespaces and are contexualized into separate user virtual service RIBs called VRFs using RFC4364 [https://datatracker.ietf.org/doc/html/rfc4364] procedures."
- SAFI 4 (BGP LU) is an RFC8277 [https://datatracker.ietf.org/doc/html/rfc8277] encoded family that carries transport prefixes in the NLRI, where the prefixes come from the provider namespace.

[Added by Kaliraj]: Just FYI: This was also explained in the following Part2 email thread: https://mailarchive.ietf.org/arch/msg/idr/GYXaStRs2xVs5r6MT9bFszU5E_g/

[Jeffrey Zhang: Discussion]:

- Both VPN and BGP-LU use the <label+prefix> encoding in RFC3107. VPN has additional RD/RT mechanisms and there is not much difference between VPN and BGP-LU other than that.
- The RT is used to control the propagation and importation of NLRIs, and the RD is used to distinguish different prefixes (perhaps I say "different NRLIs for the same prefix bits") - not only for distinguishing (overlapping) prefixes from different VPNs, but also for distinguishing different paths from different PEs for the same prefix in the same VPN. Nothing prevents it from being used for different NLRIs from the same PE for the same prefix in the same VPN.
- As some have pointed out, BGP-CT is based on mature VPN mechanisms. Considering above, I would say BGP-CT is based on both BGP-LU and VPN.
- Some people say BGP-CAR is based on BGP-LU and is better. I suppose the rational for saying that is not because color is explicitly encoded in the NLRI but that it does not use RD/RT. For that I have two points:
    1. Encoding transport class (TC) as a RT has advantages. It does not make packing bad because many will share the same TC, and the TC RT can be used to optimize the propagation and importation of the routes to only where the TC is used.

    2. Using RD/RT is not a bad thing. Rather, RD is more flexible (that's why in VPN design RD is just opaque and does not have any VPN semantics) and RT optimizes route propagation and importation. VPN technology is mature and widely deployed, and it is not obsolete. It can be used for other use cases like BGP-CT, which is a modern technology that works for both non-SR and SR use cases. Using "transport class" term instead of "color" does not mean it does not work for SR. A color can be treated as a TC and encoded in the TC RT, and encoding that way does not conflict with that SR policy encodes color in NLRI - they're different SAFI anyway.

Questions: [missing]
[Robert Raszuk Reply]: Assume I would like to extend this concept of differentiated forwarding to public cloud operators and Internet.

- [Editor's note: The following text in italics has been edited to provide specific points rather than innuendo.]
    - *How do new users who have never run VPN based signaling BGP services (e.g. L3VPNs or L2VPNs) handle CT paradigms?*

  - *I think the NLRI paradigm and TLV format of the NLRI has nice extension properties. Removing the label terminology makes sense where labels are no longer used.*

F3-Both-Issue-5: Issues to add to Issue tracker in github

- CT: Clarify the interaction with RDs and RTCs by discussing how CT handles RDs, RTCs, labels and other VPN signaling information that sent to domain with CT.
- CT discuss how efficient CT is in domains which do not handle SR-MPLS or VPNs.
- CAR: Clarify the interaction with RDs and RTCs by discussing how CAR handles RDs, RTCs, labels and other VPN signaling information that sent to domain with CAR.
- CAR discuss how efficient CAR is domains which do not handle SR-MPLS or VPNs.

F3-WG-Issue-6: Benefits of Route Targets

Thread: https://mailarchive.ietf.org/arch/msg/idr/v9f1wKjalFFOBq-3NmtN1Cg_2eQ/

[author]: Swadesh Agrawal
[subject]: CT authors have listed a benefit of route-targets as “To draw venn diagrams of transport domains that allows flexible expression of intent and handling non-agreeing color domains gracefully”.

Reference: https://mailarchive.ietf.org/arch/msg/idr/I3p0tT1BVu8MxpwQEnknzzKy8VM/

[Discussion:]
We have a question on the practicality of attaching multiple TCs for distribution to non-agreeing color domains.

A TC value used for an intent in one color domain cannot be used in any other color domain for another intent otherwise it will result in import to wrong TC that results in mis-routing and forwarding.

Please consider 3 color domains (CD1, CD2 and CD3). Each is administered independently and have TC to intent mapping as shown below. CT extended among them and originator attaches set of TCs for remote color domains intent will cause mis routing and forwarding in local and remote color domain.

```
        --- CD2
       /
CD1---/
      \
       \
        --- CD3
```

Color Domain 1 mappings:
TC1 mapped to low latency(LL)
TC2 mapped to high bandwidth(HBW)
TC3 mapped to plane A

Color Domain 2 mappings:
TC2 mapped to LL
TC3 mapped to HBW
TC4 mapped to plane A

Color Domain 3 mappings:
TC3 mapped to LL
TC2 mapped to plane A
TC5 mapped to HBW

PE in Color Domain 1 originates LL CT route attaching 3 Route-Targets{TC1, TC2, TC3}. TC1 for LL intent in CD1. TC2 for LL intent in CD2. TC3 for LL intent in CD3.
Such a route will be imported not just in LL TC in CD1 but also in HBW and plane A TC because TC2 and TC3 route-targets are used locally as well for different intents causing mis-routing and forwarding. Similarly such route is not just imported in LL TC in CD2 but also in HBW.

- **[Kaliraj-reply]:** No. PE in Domain1 originates CT route attaching only TC1. Which gets rewritten to TC2 when entering Domain2, and to TC3 when entering Domain3.

In summary, color domains are administratively managed by separate entities and will have independent TC to intent mapping. In future, when need to extend CT among them, using a Venn diagram of route targets requires globally distinct TC values which mandates rework of TC value across color domains that is not practical.

So, the only practical model is to rewrite the TC extended community at color domain boundaries,, which the CT draft also mentions. That is, in above example, a CT route from a PE in CD1 has TC1, which then gets rewritten to TC2 when advertised to a router in CD2, and to TC3 when advertised to a router in CD3.

- **[Kalriaj-reply]:** That's right. This is how it works as explained in the draft. So can you explain what was your question?
  - If your question was, where is the venn-diagram analogy used? An example is mentioned in https://datatracker.ietf.org/doc/html/draft-kaliraj-idr-bgp-classful-transport-planes-17#section-14.2 where: the Ingress-SNs subscribe tp "Egress-SN:TC-n" BGP-CT routes. But core BNs subscribe to "TC-n" BGP-CT routes. So BNs have the super-set of BGP-CT routes, but ingress-SNs have only a subset of those routes, providing 'On-Demand-Nexthop' functionality. This is one example of how the venn-diagrams formed with BGP-CT can be put to use
- **[Swadesh-2]** We just wanted to understand what the use of "Venn Diagrams" of RTs entailed specifically w.r.t multiple color domains. Since you had referenced it in the thread https://mailarchive.ietf.org/arch/msg/idr/I3p0tT1BVu8MxpwQEnknzzKy8VM/ as well as in the last IDR session. Thanks for confirming it doesn't apply. As I indicated in my example, trying to attach multiple RTs for colors of different color domains would not be practical as it would lead to incorrect import and misrouting.

F3-WG-Issue-6: Issues to add to Issue tracker in github

- CT: Provide normative text and examples for non-agreeing color domains with examples on how transport class is used. This example should include the example in F3-WG-issue-6 – which includes attaching multiple RTs to be used in different color domains is not practical
- CAR: Provide normative text and examples for non-agreeing color domains. Normative may require the authors to add additional text. The examples should include the topology (similar or exactly like the topology above) with three color domains.

F3-WG-Issue-7: Compatibility of BGP-CT and BGP-CAR to SR-PCE

IDR mail thread: https://mailarchive.ietf.org/arch/msg/idr/zWqlGvaL3zS2NqTsDAAk9L0iH-Q/

Issue author: Shraddha Hegde <shraddha@juniper.net> Wed, 27 July 2022

Introduction:

A number of adoption posts for CAR/CT have expressed concern over BGP-CT's "mapping community" as not being compatible to SR-PCE architecture. I would like to clarify that BGP-CT is very much compatible to SR-PCE architecture. "mapping community" is an abstraction that carries the desired intent.

In terms of protocol extensions, it could be regular BGP community or Extended color community.

CT Text: Section 2 Terminology section in https://datatracker.ietf.org/doc/draft-kaliraj-idr-bgp-classful-transport-planes/.

Topology: Here is example topology and related description from CT authors

```
                                    Gold = 100 Bronze = 200
                                          |-------RSVP-TE-----|
                                +--------[ASBR2     (AS2)     PE2]--+
        Gold = 100 Bronze = 200 |                                   |
CE1-----[PE1      (AS1)     ASBR1]                                 CE2 (1.1.1.1)
          |--COL-SR-TUNNEL---|  |                                   |   <--SvcPfx1 (color:0:100)
                                +--------[ASBR3     (AS3)     PE3]--+   <--SvcPfx2 (color:0:200)
                                         |-----FLEX-ALGO-----|          <--SvcPfx3
(color:0:100200)
                                        Gold = 100 Bronze = 200
```

For BGP-CT routes, transport-target:0:<color> will be the mapping community as well
The resolution scheme for BGP-CT routes will be auto-created as the following in all relevant BNs and SNs

BGP-CT resolution-schemes:
Resolution Scheme Gold (Auto Created)
Mapping-community: transport-target:0:100
Transport Route DB: TC[100]

Resolution Scheme Bronze (Auto Created)
Mapping-community: transport-target:0:200
Transport Route DB: TC[200]

Service Route resolution-schemes in PE1
Resolution Scheme Gold_fallback_BE
Mapping-community: color:0:100
Transport Route DB: TC[100, BE]

Resolution Scheme Bronze_fallback_BE
Mapping-community: color:0:200
Transport Route DB: TC[200, BE]

Resolution Scheme Gold_fallback_Bronze
Mapping-community: color:0:100200

Transport Route DB: TC[100, 200]

For deployments that want to make use of extended color community, "mapping community" is nothing but extended color community.

While "mapping community" adopts itself well to the SR-PCE architecture it also works well for deployments that used intent-aware paths and used technologies that existed prior to SR-PCE. "mapping community" is also flexible enough to satisfy evolving usecases. Example scenarios described below.

1) A number of [the] RSVP deployments use regular BGP communities carried in service prefixes to represent the intent-route that the service prefix should resolve on. These deployments would like to extend the intent-awareness multi-domain without having to go through the pain of re-mapping all their service prefixes to start using extended color community.   "mapping community" caters well to this usecase.

2) The intent-awarenesss usecases and requirements have evolved from SR-PCE days.   Section 6 of [draft-hr-spring-intentaware-routing-using-color] describes a number describes a number of fallback requirements where

> "Fallback schemes should be decoupled from primary. For example, different service routes using same primary but different fallback schemes.

(see https://datatracker.ietf.org/doc/draft-hr-spring-intentaware-routing-using-color/)

In this usecase intent is more than just color. BGP-CT solves this by carrying "mapping community" in service prefix that represents the intent which specified primary color and fallback color.

BGP-CT's approach for carrying intent is consistent. It's always in the "mapping community" no matter what kind of intent is being carried.

Question: How are use cases 1 and 2 solved in BGP-CAR?

F3-WG-Issue-7: Issues to add to Issue tracker in github

- CT: Consider how CT implements or interoperates with all the constructs in RFC 9256. Provide a short section in your document regarding support.
- CT: Describe the limits of any community, extended community, wide-community regarding color when interacting with CT's mapping community.
- CAR: Consider how CAR implements or interoperates with all the constructs in RFC9256. CAR: Provide a short section in your document regarding support.
- CAR: Describe the limits of any community, extended community or wide-community regarding color. Describe how any of these limits interact with LCM.

F3-Both-Issue-8 – Scaling and Expected Route size

[Robert Raszuk]: posts the following as follow-on to Jeff message sizes, but it is a different thread. [IDR message thread: https://mailarchive.ietf.org/arch/msg/idr/v8kkDGmr3ViPIR4UEmOPJbJ8B44/]

Technical diagram: Missing Diagram and description of use case.
References: Section 2.9.2 of CAR draft.

- [Editor: Need section for CT draft]

[Robert]: Relative to the discussion about scale and stability of CAR vs CT proposal I would like to bring a very important difference.

- As part of NLRI, CAR defines a prefix as a real IPv4 or IPv6 prefix with length. See section 2.9.2 of CAR draft.
- CT however in its NLRI defines prefix as address of 32 or 128 bits and there is no length. See section 7 of CT draft. That means that even if the operator wishes to aggregate all 300K PEs into one prefix (per ASN or per POP or per pod/cluster etc ...) for a given color before it sends it over BGP with CT it is not an option. While in CAR it is.

Leave alone that adding RD to the CT NLRI makes it even more impossible to aggregate. This is a fundamental difference especially when /32s or /128s are not needed to be sprayed to other ASNs.

- Till that is fixed I recommend that CT draft goes back to the drawing board and would not be even accepted as experimental. Its experiment has just concluded as a failure.

[Dhananjaya Rao (DJ)-reply]: Robert: I agree that the scaling requirements are indeed much higher than what's been deployed with BGP-LU. We do think a hierarchical design is a must if the transport layer is extended for such scale. The CAR draft describes a couple of hierarchical MPLS designs (Section 6) and a brief analysis of trade-offs (Section 6.3).

- from: https://mailarchive.ietf.org/arch/msg/idr/oLEvSCrzxzJFnNX86AlZ07V9708

[Srihari Sangli-reply]: Could you please describe the use case why would a customer want to do this. Also, what are we optimizing by doing this. Please also consider it is not just the sender that needs to do this work. Are you trying to optimize the bits on the wire? Given different attributes these prefixes may have, I would say this can be a theoretical exercise.

- **[Robert-2]:** 2.5 min of BGP propagation time for 1.5 million routes (as Jeff stated). Do you think it is not worth to consider an aggregation in such a case? Imagine I have POD or Cluster with 1000 PEs (compute as PEs) and I see no value in advertising all 1000s each with 5 colors especially considering that such 5 colors will share identical links and paths to those compute. Same if you replace compute by PEs.

  [Chair (Jeff Haas) notes that Robert needs to follow the format. It appears that Robert believes he has referred to a section in CT, but the editor (Susan Hares) is unclear about what section he refers to.]

- **[Robert-3]:** I quoted sections from both drafts ? I do think this fear of moderation - when just pointing out operational differences and asking clarification questions - simply inappropriate.
- [Jeffs correction]: https://mailarchive.ietf.org/arch/msg/idr/J-iWDz2dthqz25V6qWUOwD26aZ8/

  [Conversation continues on at:

https://mailarchive.ietf.org/arch/msg/idr/J-iWDz2dthqz25V6qWUOwD26aZ8/

  - [Jeff-2]: The 1.5M routes is from the problem statement in SPRING. I'd suggest you take it up with that WG as to whether they think these were intended to be aggregate-able. Discussion of what aggregation could look like in a proposal is appropriate for IDR.
  - **[Jeff-2]** It's reasonable to ask the authors how they intend to address aggregation. It's reasonable to point out text you think supports your thinking.
  - **[Jeff-2**] Unreasonable statements: “*Till that is fixed I recommend that CT draft goes back to the drawing board and would not be even accepted as experimental. Its experiment has just concluded as a failure."*

- [Shraddha]: As per the current version of: [draft-hr-spring-intentaware-routing-using-color]
  https://datatracker.ietf.org/doc/draft-hr-spring-intentaware-routing-using-color/
  - Ability to aggregate the intent-aware route itself has not been listed as a specific requirement. While ability to aggregate subscription routes has been listed as a requirement.
  - In case of MPLS forwarding plane, aggregation of intent aware routes is not feasible as you need distinct label for each <endpoint, color>. In case of SRv6, its feasible to summarize the intent-aware routes but it takes away visibility into other granular information such as E2E metric, PE reachability etc and such a requirement needs more discussion.
  - A pull model via automatic subscription routes has been listed as a scaling requirement for the solution in Sec 6.3.2.1 of https://datatracker.ietf.org/doc/draft-hr-spring-intentaware-routing-using-color/.
  - This is a much better scalability option IMO.

  - [Robert-reply-Shraddha]: Is pull model really going to be deployed inter-as here?
    - [Shraddha-2] Pls refer [to text below] from section 6.1.1 in https://datatracker.ietf.org/doc/draft-hr-spring-intentaware-routing-using-color/:
      - 6.1.1. Transport Network Intent Requirements
      - "The requirements described in this document are mostly applicable to network under a single administrative domain that are organized into multiple network domains. The requirements are also applicable to multi-AS networks with closely cooperating administration."
    - [Uttaro] +1 to Shraddha
      - [Robert-2]: Apologies but this is not what I was pointing at. I was asking if you are really going to deploy RTC like pull mechanism on the inter-as boundaries (even under cooperating administration) ?
        - Especially here in CT proposal RTC pull model will be almost ineffective if what you request is color and there are a total 5 colors in the game. So let's make it clear that the current scheme does not provide much for scale optimization.
        - [Jeff-2]: Possibly so. I keep hoping the -CAR authors will fill out their TBD covering filtering so we can have start having a discussion on how (E,C) filtering will generally behave in terms of scale. (There are also interesting questions (There are also interesting questions that will occur around how subscriptions work across color domains.)
      - [Robert-2] (Editor: Italics contains post without inflammatory text) *The only option I could perhaps see helping with scale would be ORF like extension where ASN_X could signal ASN_Y which next hops are of specific interest.*

But then we run into again into the RD wall where it makes it even less possible to filter as you would have to wildcard front of the NLRI ,,, not good.

- **[Jeff-2]** My expectation is the proposals will eventually work toward subscriptions for sets of endpoints vs. the "effective color". This is the property we're looking at for route resolution in both proposals we're interested in. Components of NLRI for RIB key (and BGP route comparison purposes) isn't necessarily what we're interested in.

- **[Robert-2]** (Editor italics text was edited to remove inflammatory information*): Concerns regarding CT and scale.*
  - **[Srihari-2a]: I** was not hinting that we "not worry" about it, I was questioning the practicality of such a aggregation from the domain BR. Each specific endpoint address can have uniqueness (described via many attributes not to mention the new intent), and I would argue the ingress PE might be more interested in knowing this detail than just point to the aggregate route (very similar to one using default vs more specific routing).
  - **[Srihari-2b]** But I don't refute your claim that aggregation may be needed. Sure, it will help scalability, we all know that and I see both proposals have ability to do that. Nats responded in one of the prev emails.
  - **[Jeff-2]:** I'm unclear how CAR aggregates better. While I'd prefer to see the response from the -CAR authors as to how they expect aggregation to work, please feel free to speculate. Ideally with citation from the specification, if possible. Please also note the responses from the -CT authors that the IP Prefix is a prefix, not a host address.
  - **[Robert-**3a]: Hi Jeff, Many thx for your note ! **Btw I am not claiming CAR aggregates or summarizes any better :)**
    - **[Jeff-3a**] Ah. That wasn't clear in prior discussion.
  - **[Robert-**3b] So thinking about it more I did some more reading. It looks like the CT intention is following which sorry is not very obvious. CT routes as stated in CT draft will contain two RTs:
    - *An egress SN MAY advertise BGP CT route for RD:eSN with two Route Targets: transport-target:0:<TC> and a RT carrying <eSN>:<TC>.
    - **[Jeff-3b]:** The text in this section could use a few more caveats as well, but at least sketches in some of the details in a way that provides a point of discussion. I think the most salient point to the filtering discussion is called out in one sentence: "*However, the amount of state carried in RTC family may become proportional to number of PNHs in the network.* " I suspect this will be true for any filtering discussion we have. More than anything else, I think this point is a core pivot for the scaling properties of these LU-like routes with colors.

- **[Robert-3c]:** Then in turn *extension to RTC [draft-zzhang-idr-bgp-rt-constrains-extension-00] is needed to filter on both. (https://datatracker.ietf.org/doc/html/draft-zzhang-idr-bgp-rt-constrains-extension-00).
    - [Jeff-3c]: I suspect that the observation comes back to my prior response in this thread: We need a subscription mechanism based on a subset of endpoints for a given color. The -CT route-target procedures are an attempt to do this binding, but I think it still misses some of the core points.
    - [Jeff-3d] I'll leave it to the -ct authors to expound on where they think this goes, but it's a detail we'll want to clarify in each of the specifications as we work toward RFC. I suspect there may even be some level of commonality that lets a single mechanism be used for both proposals.
- **[Robert-3d]** Well I was misled by claims made that CT is using today's VPN distribution - it is clearly not. So with the above I see why now you really need RD and why some scaling could be established. Not that I really accept this as good solution, it is very cumbersome to me - but I see how it could be deployed. CT authors pls do correct me if the above understanding is incorrect.
- **[Robert-3e]** PS. I am not a huge fan of controllers between cooperating domains but it seems that this entire CAR/CT problem space could be easily solved by such an upper controller layer instead of pushing it to the dynamic routing protocol.
    - [Jeff-3e]: If we had been able to declare victory with only having a single color domain, we would have been mostly done with this work three years ago. :-)
    - [Robert-4a]: I am not sure I follow this. I was not suggesting we need single color domains. I was saying that the color translation and mapping can be done at the controller layer which in turn distribute policies down to network elements (SNs). We already have service routes in BGP, we have forwarding in each domain provided one way or the other. I am not sure I really see huge value in adding BGP to the mix as some distributed mapping scheme.
        - [Jeff-4a]: The short form of it is there was a desire for BGP-LU like behavior for distributing the necessary paths that had differentiated forwarding that can be used for service route nexthop resolution. If you're not interested in deploying such a solution, there's other options. Clearly with a god-box controller and sufficiently flexible policy, you can do what you like without much protocol help.
    - **[Robert-4b]** I am not even sure how CAR/CT would work with controllers generating path policies in each domain. Are we assuming that there is none such elements where we deploy CAR/CT?

- **[Jeff-4b]:** I think you'll find that each of the proposals have a rich set of circumstances where they pick up their differentiated forwarding from. How they do so is just another part of the BGP saga of being the glue that ties things together

F3-Both-Issue-8: Issues to add to Issue tracker in github

- CAR: Discuss in scale section how CAR scales to:
  - limits in draft-hr-spring-intentaware-routing-using-color-00
  - Jeff Haas' rough route calculation: 1.5 million routes, given 10K update, about 2.5 minutes of convergence
  - Robert's use case: transient route problems every 5-10 sections every 50 seconds
- CT: Discuss in scale section how CAR scales to:
  - limits in draft-hr-spring-intentaware-routing-using-color-00
  - Jeff Haas' rough route calculation: 1.5 million routes, given 10K update, about 2.5 minutes of convergence
  - Robert's use case: transient route problems every 5-10 sections every 50 seconds