BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2020

A. Farrel
Old Dog Consulting
J. Drake
E. Rosen
Juniper Networks
K. Patel
Arrcus, Inc.
L. Jalil
Verizon

March 11, 2020

Gateway Auto-Discovery and Route Advertisement for Segment Routing Enabled Domain Interconnection
draft-ietf-bess-datacenter-gateway-05

Abstract

Data centers are critical components of the infrastructure used by network operators to provide services to their customers. Data centers are attached to the Internet or a backbone network by gateway routers. One data center typically has more than one gateway for commercial, load balancing, and resiliency reasons.

Segment Routing is a popular protocol mechanism for use within a data center, but also for steering traffic that flows between two data center sites. In order that one data center site may load balance the traffic it sends to another data center site, it needs to know the complete set of gateway routers at the remote data center, the points of connection from those gateways to the backbone network, and the connectivity across the backbone network.

Segment Routing may also be operated in other domains, such as access networks. Those domains also need to be connected across backbone networks through gateways.

This document defines a mechanism using the BGP Tunnel Encapsulation attribute to allow each gateway router to advertise the routes to the prefixes in the Segment Routing domains to which it provides access, and also to advertise on behalf of each other gateway to the same Segment Routing domain.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This Internet-Draft will expire on September 12, 2020.

Copyright Notice

Table of Contents

1. Introduction

Data centers (DCs) are critical components of the infrastructure used by network operators to provide services to their customers. DCs are attached to the Internet or a backbone network by gateway routers (GWs). One DC typically has more than one GW for various reasons including commercial preferences, load balancing, and resiliency against connection of device failure.

Segment Routing (SR) [RFC8402] is a popular protocol mechanism for use within a DC, but also for steering traffic that flows between two DC sites. In order for a source (ingress) DC that uses SR to load balance the flows it sends to a destination (egress) DC, it needs to know the complete set of entry nodes (i.e., GWs) for that egress DC from the backbone network connecting the two DCs. Note that it is assumed that the connected set of DCs and the backbone network connecting them are part of the same SR BGP Link State (LS) instance ([RFC7752] and [I-D.ietf-idr-bgpls-segment-routing-epe]) so that traffic engineering using SR may be used for these flows.

SR may also be operated in other domains, such as access networks. Those domains also need to be connected across backbone networks through gateways.

Suppose that there are two gateways, GW1 and GW2 as shown in Figure 1, for a given egress SR domain and that they each advertise a route to prefix X which is located within the egress SR domain with each setting itself as next hop. One might think that the GWs for X could be inferred from the routes' next hop fields, but typically it is not the case that both routes get distributed across the backbone: rather only the best route, as selected by BGP, is distributed. This precludes load balancing flows across both GWs.

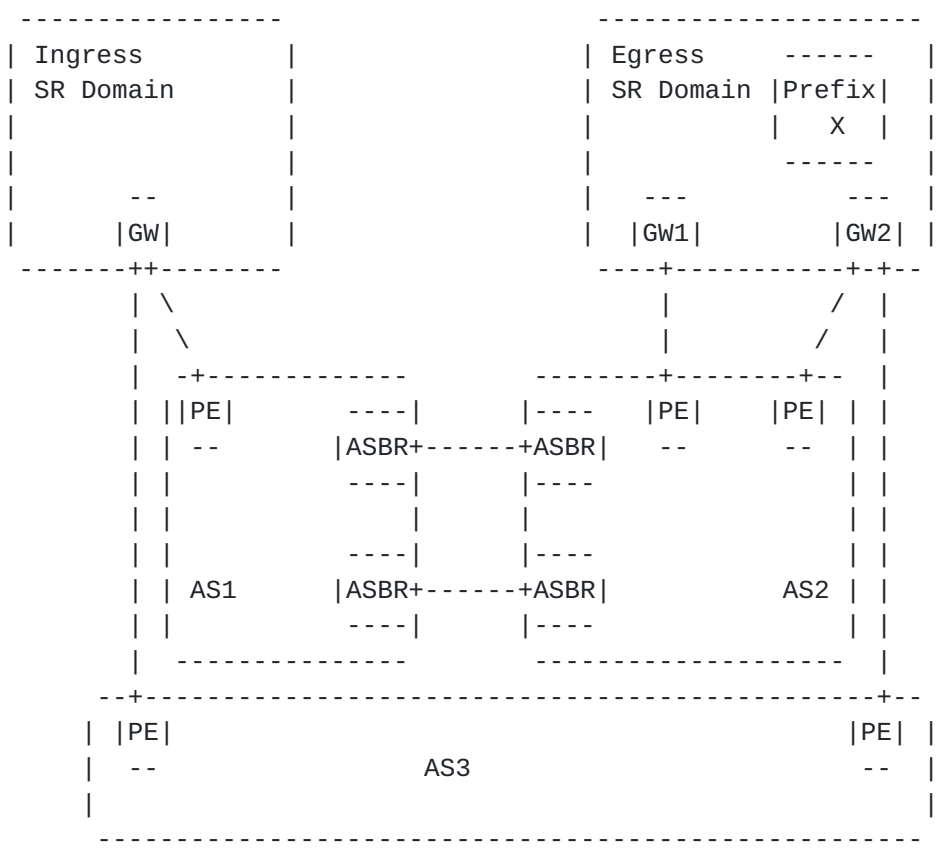

Figure 1: Example Segment Routing Domain Interconnection

The obvious solution to this problem is to use the BGP feature that allows the advertisement of multiple paths in BGP (known as Add-Paths) [RFC7911] to ensure that all routes to X get advertised by BGP. However, even if this is done, the identity of the GWs will be lost as soon as the routes get distributed through an Autonomous System Border Router (ASBR) that will set itself to be the next hop. And if there are multiple Autonomous Systems (ASes) in the backbone, not only will the next hop change several times, but the Add-Paths technique will experience scaling issues. This all means that the Add-Paths approach is limited to SR domains connected over a single AS.

This document defines a solution that overcomes this limitation and works equally well with a backbone constructed from one or more ASes. The solution uses the Tunnel Encapsulation attribute [I-D.ietf-idr-tunnel-encaps] as follows:

> We define a new tunnel type, "SR Tunnel". When the GWs to a given SR domain advertise a route to a prefix X within the SR domain, they will each include a Tunnel Encapsulation attribute with

multiple tunnel instances each of type "SR Tunnel" (value 17), one for each GW, and each containing a Remote Endpoint sub-TLV with that GW's address.

In other words, each route advertised by a GW identifies all of the GWs to the same SR domain (see Section 3 for a discussion of how GWs discover each other). Therefore, even if only one of the routes is distributed to other ASes, it will not matter how many times the next hop changes, as the Tunnel Encapsulation attribute (and its remote endpoint sub-TLVs) will remain unchanged.

To put this in the context of Figure 1, GW1 and GW2 discover each other as gateways for the egress SR domain. Both GW1 and GW2 advertise themselves as having routes to prefix X. Furthermore, GW1 includes a Tunnel Encapsulation attribute with a tunnel instance of type "SR tunnel" for itself and another for GW2. Similarly, GW2 includes a Tunnel Encapsulation for itself and another for GW1. The gateway in the ingress SR domain can now see all possible paths to the egress SR domain regardless of which route advertisement is propagated to it, and it can choose one, or balance traffic flows as it sees fit.

The protocol extensions defined in this document are put into the broader context of SR domain interconnection by [I-D.farrel-spring-sr-domain-interconnect]. That document shows how other existing protocol elements may be combined with the extensions defined in this document to provide a full system.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. SR Domain Gateway Auto-Discovery

To allow a given SR domain's GWs to auto-discover each other and to coordinate their operations, the following procedures are implemented:

- Each GW is configured with an identifier for the SR domain. That identifier is common across all GWs to the domain (i.e., the same identifier is used by all GWs to the same SR domain), and unique across all SR domains that are connected (i.e., across all GWs to all SR domains that are interconnected).

- A route target ([RFC4360]) is attached to each GW's auto-discovery route and has its value set to the SR domain identifier.

- Each GW constructs an import filtering rule to import any route that carries a route target with the same SR domain identifier that the GW itself uses. This means that only these GWs will import those routes, and that all GWs to the same SR domain will import each other's routes and will learn (auto-discover) the current set of active GWs for the SR domain.

The auto-discovery route that each GW advertises consists of the following:

- An IPv4 or IPv6 NLRI containing one of the GW's loopback addresses (that is, with an AFI/SAFI pair that is one of 1/1, 2/1, 1/4, or 2/4).

- A Tunnel Encapsulation attribute containing the GW's encapsulation information, which at a minimum consists of an SR Tunnel TLV (type TBD1 to be allocated by IANA) with a Remote Endpoint sub-TLV as specified in [I-D.ietf-idr-tunnel-encaps].

To avoid the side effect of applying the Tunnel Encapsulation attribute to any packet that is addressed to the GW itself, the GW SHOULD use a different loopback address for the two cases.

As described in Section 1, each GW will include a Tunnel Encapsulation attribute for each GW that is active for the SR domain (including itself), and will include these in every route advertised externally to the SR domain by each GW. As the current set of active GWs changes (due to the addition of a new GW or the failure/removal of an existing GW) each externally advertised route will be re-advertised with the set of SR tunnel instances reflecting the current set of active GWs.

If a gateway becomes disconnected from the backbone network, or if the SR domain operator decides to terminate the gateway's activity, it withdraws the advertisements described above. This means that remote gateways at other sites will stop seeing advertisements from this gateway. It also means that other local gateways at this site will "unlearn" the removed gateway and stop including a Tunnel Encapsulation attribute for the removed gateway in their advertisements.

4. Relationship to BGP Link State and Egress Peer Engineering

When a remote GW receives a route to a prefix X it can use the SR tunnel instances within the contained Tunnel Encapsulation attribute to identify the GWs through which X can be reached. It uses this information to compute SR Traffic Engineering (SR TE) paths across the backbone network looking at the information advertised to it in SR BGP Link State (BGP-LS) [I-D.ietf-idr-bgp-ls-segment-routing-ext] and correlated using the SR domain identity. SR Egress Peer Engineering (EPE) [I-D.ietf-idr-bgpls-segment-routing-epe] can be used to supplement the information advertised in BGP-LS.

5. Advertising an SR Domain Route Externally

When a packet destined for prefix X is sent on an SR TE path to a GW for the SR domain containing X, it needs to carry the receiving GW's label for X such that this label rises to the top of the stack before the GW completes its processing of the packet. To achieve this we place a Prefix SID sub-TLV [I-D.ietf-idr-tunnel-encaps] for X in each SR tunnel instance in the Tunnel Encapsulation attribute in the externally advertised route for X.

Alternatively, if the GWs for a given SR domain are configured to allow remote GWs to perform SR TE through that SR domain for a prefix X, then each GW computes an SR TE path through that SR domain to X from each of the currently active GWs, and places each in an MPLS label stack sub-TLV [I-D.ietf-idr-tunnel-encaps] in the SR tunnel instance for that GW.

6. Encapsulation

If the GWs for a given SR domain are configured to allow remote GWs to send them a packet in that SR domain's native encapsulation, then each GW will also include multiple instances of a tunnel TLV for that native encapsulation in externally advertised routes: one for each GW and each containing a remote endpoint sub-TLV with that GW's address. A remote GW may then encapsulate a packet according to the rules defined via the sub-TLVs included in each of the tunnel TLV instances.

7. IANA Considerations

7.1. Tunnel Encapsulation Tunnel Type

IANA maintains a registry called "Border Gateway Protocol (BGP) Parameters" with a sub-registry called "BGP Tunnel Encapsulation Attribute Tunnel Types." The registration policy for this registry is First-Come First-Served [RFC8126].

IANA has assigned the value 17 from this sub-registry for "SR Tunnel".

7.2. Tunnel Encapsulation Sub-TLVs

IANA maintains a registry called "Border Gateway Protocol (BGP) Parameters" with a sub-registry called "BGP Tunnel Encapsulation Attribute Sub-TLVs." The registration policy for this registry is Standards Action.[RFC8126].

IANA is requested to assign a codepoint from this sub-registry for "SR Tunnel TLV" (TBD1). The next available value may be used and reference should be made to this document.

8. Security Considerations

From a protocol point of view, the mechanisms described in this document can leverage the security mechanisms already defined for BGP. Further discussion of security considerations for BGP may be found in the BGP specification itself [RFC4271] and in the security analysis for BGP [RFC4272]. The original discussion of the use of the TCP MD5 signature option to protect BGP sessions is found in [RFC5925], while [RFC6952] includes an analysis of BGP keying and authentication issues.

The mechanisms described in this document involve sharing routing or reachability information between domains: that may mean disclosing information that is normally contained within a domain. So it needs to be understood that normal security paradigms based on the boundaries of domains are weakened. Discussion of these issues with respect to VPNs can be found in [RFC4364], while [RFC7926] describes many of the issues associated with the exchange of topology or TE information between domains.

Particular exposures resulting from this work include:

- Gateways to a domain will know about all other gateways to the same domain. This feature applies within a domain and so is not a substantial exposure, but it does mean that if the BGP exchanges within a domain can be snooped or if a gateway can be subverted then an attacker may learn the full set of gateways to a domain. This would facilitate more effective attacks on that domain.

- The existence of multiple gateways to a domain becomes more visible across the backbone and even into remote domains. This means that an attacker is able to prepare a more comprehensive attack than exists when only the locally attached backbone network (e.g., the AS that hosts the domain) can see all of the gateways

to a site. For example, a Denial of Service attack on a single GW is mitigated by the existence of other GWs, but if the attacker knows about all the gateways then the whole set can be attacked at once.

- o A node in a domain that does not have external BGP peering (i.e., is not really a domain gateway and cannot speak BGP into the backbone network) may be able to get itself advertised as a gateway by letting other genuine gateways discover it (by speaking BGP to them within the domain) and so may get those genuine gateways to advertise it as a gateway into the backbone network. This would allow the malicious node to attract traffic without having to have secure BGP peerings with out-of-domain nodes.

- o If it is possible to modify a BGP message within the backbone, it may be possible to spoof the existence of a gateway. This could cause traffic to be attracted to a specific node and might result in black-holing of traffic.

All of the issues in the list above could cause disruption to domain interconnection, but are not new protocol vulnerabilities so much as new exposures of information that SHOULD be protected against using existing protocol mechanisms. Furthermore, it is a general observation that if these attacks are possible then it is highly likely that far more significant attacks can be made on the routing system. It should be noted that BGP peerings are not discovered, but always arise from explicit configuration.

9. Manageability Considerations

The principal configuration item added by this solution is the allocation of an SR domain identifier. The same identifier MUST be assigned to every GW to the same domain, and each domain MUST have a different identifier. This requires coordination, probably through a central management agent.

It should be noted that BGP peerings are not discovered, but always arise from explicit configuration. This is no different from any other BGP operation.

10. Acknowledgements

Thanks to Bruno Rijsman and Stephane Litkowsji for review comments, and to Robert Raszuk for useful discussions.

11. References

11.1. Normative References

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Talaulikar, K., Filsfils, C., Patel, K., Ray, S., and J. Dong, "BGP-LS extensions for Segment Routing BGP Egress Peer Engineering", draft-ietf-idr-bgpls-segment-routing-epe-19 (work in progress), May 2019.

[I-D.ietf-idr-tunnel-encaps]
Patel, K., Velde, G., and S. Ramachandra, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-15 (work in progress), December 2019.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

11.2. Informative References

[I-D.farrel-spring-sr-domain-interconnect]
Farrel, A. and J. Drake, "Interconnection of Segment Routing Domains - Problem Statement and Solution Landscape", draft-farrel-spring-sr-domain-interconnect-05 (work in progress), October 2018.

[I-D.ietf-idr-bgp-ls-segment-routing-ext]
Previdi, S., Talaulikar, K., Filsfils, C., Gredler, H., and M. Chen, "BGP Link-State extensions for Segment Routing", draft-ietf-idr-bgp-ls-segment-routing-ext-16 (work in progress), June 2019.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, DOI 10.17487/RFC6952, May 2013, <https://www.rfc-editor.org/info/rfc6952>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC7926] Farrel, A., Ed., Drake, J., Bitar, N., Swallow, G., Ceccarelli, D., and X. Zhang, "Problem Statement and Architecture for Information Exchange between Interconnected Traffic-Engineered Networks", BCP 206, RFC 7926, DOI 10.17487/RFC7926, July 2016, <https://www.rfc-editor.org/info/rfc7926>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

Authors' Addresses

Adrian Farrel
Old Dog Consulting

Email: adrian@olddog.co.uk

John Drake
Juniper Networks

Email: jdrake@juniper.net

Eric Rosen
Juniper Networks

Email: erosen52@gmail.com

Keyur Patel
Arrcus, Inc.

Email: keyur@arrcus.com

Luay Jalil
Verizon

Email: luay.jalil@verizon.com