BESS Working Group
Internet-Draft
Intended status: Informational
Expires: April 19, 2017

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

October 16, 2016

Gateway Auto-Discovery and Route Advertisement for Segment Routing Enabled Data Center Interconnection
draft-drake-bess-datacenter-gateway-02

Abstract

Data centers have become 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 operating 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.

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 data center site to which it provides access, and also to advertise on behalf of each other gateway to the same data center site.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

Copyright Notice

Table of Contents

1. Introduction

Data centers (DCs) have become 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) [I-D.ietf-spring-segment-routing] is a popular protocol mechanism for operating within a DC, but also for steering traffic that flows between two DC sites. In order for an ingress DC that uses SR to load balance the flows it sends to an 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.

Suppose that there are two gateways, GW1 and GW2 as shown in Figure 1, for a given egress DC and that they each advertise a route to prefix X which is located within the egress DC 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.

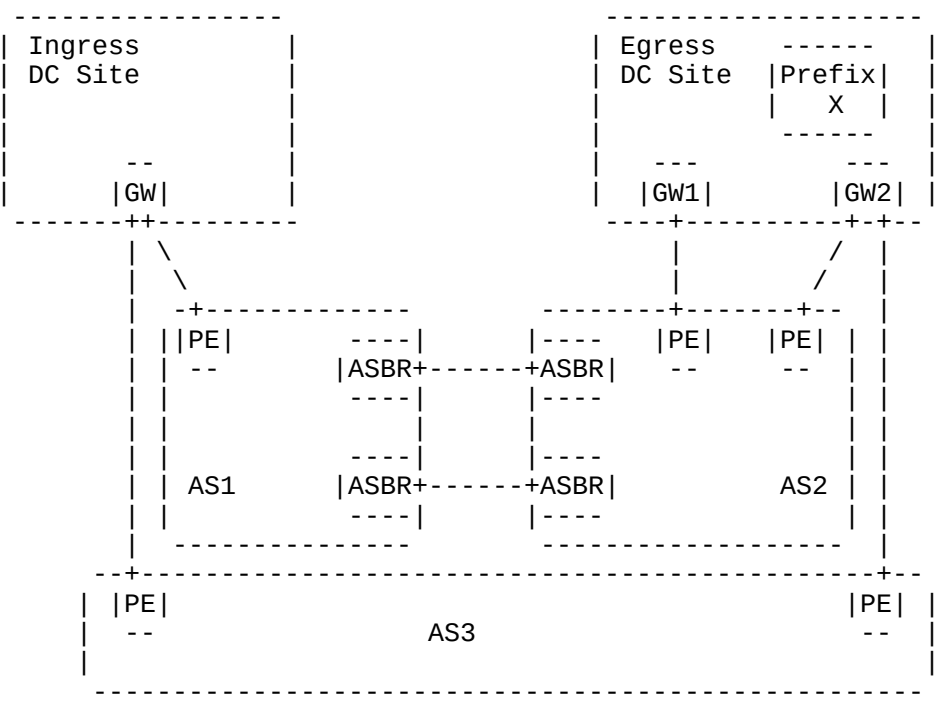

Figure 1: Example Data Center Interconnection

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 this approach is limited to DC sites 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. This 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 DC advertise a route to a prefix X within the DC, they will each include a Tunnel Encapsulation attribute with multiple tunnel

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

In other words, each route advertised by any GW identifies all of the GWs to the same DC (see Section 2 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 data center site. 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 data center site can now see all possible paths to the egress data center site regardless of which route advertisement is propagated to it, and it can choose one or balance traffic flows as it sees fit.

2. DC Gateway Auto-Discovery

To allow a given DC'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 DC that is common across all GWs to the DC (i.e., across all GWs to all DC sites that are interconnected) and unique across all DCs that are connected.

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

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

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

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

- o A Tunnel Encapsulation attribute containing the GW's encapsulation information, which at a minimum consists of an SR tunnel TLV (type 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 DC site (including itself), and will include these in every route advertised externally to the DC site 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 DC 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.

3. 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 TE paths across the backbone network looking at the information advertised to it in SR BGP Link State (BGP-LS) [I-D.gredler-idr-bgp-ls-segment-routing-ext] and correlated using the DC identity. SR Egress Peer Engineering (EPE) [I-D.ietf-idr-bgpls-segment-routing-epe] can be used to supplement the information advertised in the BGP-LS.

4. Advertising a DC Route Externally

When a packet destined for prefix X is sent on an SR TE path to a GW for the DC site 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 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 DC are configured to allow remote GWs to perform SR TE through that DC for a prefix X, then each GW computes an SR TE path through that DC 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.

5. Encapsulation

If the GWs for a given DC are configured to allow remote GWs to send them a packet in that DC'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.

6. IANA Considerations

IANA maintains a registry called "BGP parameters" with a sub-registry called "BGP Tunnel Encapsulation Tunnel Types." The registration policy for this registry is First-Come First-Served.

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

[[Note: This text is likely to be replaced with a specific code point value once FCFS allocation has been made.]]

7. Security Considerations

TBD

8. Manageability Considerations

TBD

9. Acknowledgements

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

10. References

10.1. Normative References

[I-D.ietf-idr-bgpls-segment-routing-epe]
Previdi, S., Filsfils, C., Ray, S., Patel, K., Dong, J., and M. Chen, "Segment Routing BGP Egress Peer Engineering BGP-LS Extensions", draft-ietf-idr-bgpls-segment-routing-epe-05 (work in progress), May 2016.

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "The BGP Tunnel Encapsulation Attribute", draft-ietf-idr-tunnel-encaps-02 (work in progress), May 2016.

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

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

[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, <http://www.rfc-editor.org/info/rfc7752>.

10.2. Informative References

[I-D.gredler-idr-bgp-ls-segment-routing-ext]
Previdi, S., Psenak, P., Filsfils, C., Gredler, H., Chen, M., and j. jefftant@gmail.com, "BGP Link-State extensions for Segment Routing", draft-gredler-idr-bgp-ls-segment-routing-ext-03 (work in progress), July 2016.

[I-D.ietf-spring-segment-routing]
Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", draft-ietf-spring-segment-routing-09 (work in progress), July 2016.

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

Authors' Addresses

John Drake
Juniper Networks

Email: jdrake@juniper.net

Adrian Farrel
Juniper Networks

Email: adrian@olddog.co.uk

Eric Rosen
Juniper Networks

Email: erosen@juniper.net

Keyur Patel
Arrcus, Inc.

Email: keyur@arrcus.com

Luay Jalil
Verizon

Email: luay.jalil@verizon.com