BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 9, 2021

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

June 7, 2021

Gateway Auto-Discovery and Route Advertisement for Segment Routing Enabled Site Interconnection
draft-ietf-bess-datacenter-gateway-12

Abstract

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. Other sites, such as access networks, also need to be connected across backbone networks through gateways.

This document defines a mechanism using the BGP Tunnel Encapsulation attribute to allow data center gateway routers to advertise routes to the prefixes reachable in the site, including advertising them on behalf of other gateways at the same site. This allows segment routing to be used to identify multiple paths across the Internet or backbone network between different gateways. The paths can be selected for load-balancing, resilience, and quality purposes.

Status of This Memo

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 (sites) are interconnected by a backbone network, which consists of any number of private networks and/or the Internet, by gateway routers (GWs). One DC typically has more than one GW for various reasons including commercial preferences, load balancing, or resiliency against connection or device failure.

Segment Routing (SR) [RFC8402] is a protocol mechanism that can be used within a DC, and also for steering traffic that flows between two DC sites. In order for a source site (also known as an ingress site) that uses SR to load balance the flows it sends to a destination site (also known as an egress site), 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 DC sites and the border nodes in the backbone network on the paths that connect the DC sites 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.

Other sites, such as access networks, also need to be connected across backbone networks through gateways. For illustrative purposes, consider the ingress and egress sites shown in Figure 1 as separate ASes (noting that the sites could be implemented as part of the ASes to which they are attached, or as separate ASes). The various ASes that provide connectivity between the ingress and egress sites could each be constructed differently and use different technologies such as IP, MPLS using global table routing information from native BGP, MPLS IP VPN, SR-MPLS IP VPN, or SRv6 IP VPN. That is, the ingress and egress sites can be connected by tunnels across a variety of technologies. This document describes how SR identifiers (SIDs) are used to identify the paths between the ingress and egress sites.

The solution described in this document is agnostic as to whether the transit ASes do or do not have SR capabilities. the solution uses SR to stitch together path segments between GWs and through the ASBRs. Thus, there is a requirement that the GWs and ASBRs are SR-capable. The solution supports the SR path being extended into the ingress and egress sites if they are SR-capable.

The solution defined in this document can be seen in the broader context of site interconnection in [I-D.farrel-spring-sr-domain-interconnect]. That document shows how other existing protocol elements may be combined with the solution defined in this document to provide a full system, but is not a necessary reference for understanding this document.

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

```
           -----------------                    ---------------------
          | Ingress         |                  | Egress     ------   |
          | Site            |                  | Site      |Prefix|  |
          |                 |                  |           |   X  |  |
          |                 |                  |            ------   |
          |       --        |                  |   ---         ---  |
          |      |GW|       |                  |  |GW1|        |GW2| |
           -------++--------                    ----+-----------+-+--
                  | \                               |          /  |
                  |  \                              |         /   |
                  |  -+-------------        --------+--------+--  |
                  | ||ASBR|     ----|      |----  |ASBR| |ASBR| | |
                  | | ----     |ASBR+------+ASBR|  ----   ----  | |
                  | |           ----|      |----                | |
                  | |               |      |                    | |
                  | |           ----|      |----                | |
                  | | AS1      |ASBR+------+ASBR|           AS2 | |
                  | |           ----|      |----                | |
                  |  ---------------        --------------------  |
                --+-----------------------------------------------+--
               | |ASBR|                                       |ASBR| |
               |  ----               AS3                       ----  |
               |                                                     |
                -----------------------------------------------------
```

Figure 1: Example Site 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 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 using the Tunnel Encapsulation attribute [RFC9012] as follows:

> When a GW to a given site advertises a route to a prefix X within that site, it will include a Tunnel Encapsulation attribute that contains the union of the Tunnel Encapsulation attributes advertised by each of the GWs to that site, including itself.

In other words, each route advertised by a GW identifies all of the GWs to the same site (see Section 3 for a discussion of how GWs discover each other). I.e., the Tunnel Encapsulation attribute advertised by each GW contains multiple Tunnel TLVs, one or more from each active GW, and each Tunnel TLV will contain a Tunnel Egress Endpoint Sub-TLV that identifies the GW for that Tunnel TLV. 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 will remain unchanged.

To put this in the context of Figure 1, GW1 and GW2 discover each other as gateways for the egress site. Both GW1 and GW2 advertise themselves as having routes to prefix X. Furthermore, GW1 includes a Tunnel Encapsulation attribute which is the union of its Tunnel Encapsulation attribute and GW2's Tunnel Encapsulation attribute. Similarly, GW2 includes a Tunnel Encapsulation attribute which is the union of its Tunnel Encapsulation attribute and GW1's Tunnel Encapsulation attribute. The gateway in the ingress site can now see all possible paths to X in the egress site regardless of which route is propagated to it, and it can choose one, or balance traffic flows as it sees fit.

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. Site Gateway Auto-Discovery

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

- o A route target ([RFC4360]) MUST be attached to each GW's auto-discovery route (defined below) and its value MUST be set to a value that indicates the site identifier. The rules for constructing a route target are detailed in [RFC4360]. It is RECOMMENDED that a Type x00 or x02 route target be used.

- o Site identifiers are set through configuration. The site identifiers MUST be the same across all GWs to the site (i.e., the same identifier is used by all GWs to the same site), and MUST be unique across all sites that are connected (i.e., across all GWs to all sites that are interconnected).

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

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

- An IPv4 or IPv6 Network Layer Reachability Information (NLRI) [RFC4760] containing one of the GW's loopback addresses (that is, with an AFI/SAFI pair that is one of IPv4/NLRI used for unicast forwarding (1/1), IPv6/NLRI used for unicast forwarding (2/1), IPv4/NLRI with MPLS Labels (1/4), or IPv6/NLRI with MPLS Labels (2/4)).

- A Tunnel Encapsulation attribute [RFC9012] containing the GW's encapsulation information encoded in one or more Tunnel TLVs.

To avoid the side effect of applying the Tunnel Encapsulation attribute to any packet that is addressed to the GW itself, the address advertised for auto-discovery MUST be a different loopback address than is advertised for packets directed to the gateway itself.

As described in Section 1, each GW will include a Tunnel Encapsulation attribute with the GW encapsulation information for each of the site's active GWs (including itself) in every route advertised externally to that site. 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 a new Tunnel Encapsulation attribute which reflects the current set of active GWs.

If a gateway becomes disconnected from the backbone network, or if the site operator decides to terminate the gateway's activity, it MUST withdraw the advertisements described above. This means that remote gateways at other sites will stop seeing advertisements from or about this gateway. Note that if the routing within a site is broken (for example, such that there is a route from one GW to another, but not in the reverse direction), then it is possible that incoming traffic will be routed to the wrong GW to reach the destination prefix - in this degraded network situation, traffic may be dropped.

Note that if a GW is (mis)configured with a different site identifier from the other GWs to the same site then it will not be auto-

discovered by the other GWs (and will not auto-discover the other GWs). This would result in a GW for another site receiving only the Tunnel Encapsulation attribute included in the BGP best route; i.e., the Tunnel Encapsulation attribute of the (mis)configured GW or that of the other GWs.

4. Relationship to BGP Link State and Egress Peer Engineering

When a remote GW receives a route to a prefix X, it uses the Tunnel Egress Endpoint Sub-TLVs in the containing 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 site 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 a Site Route Externally

When a packet destined for prefix X is sent on an SR TE path to a GW for the site containing X (that is, the packet is sent in the ingress site on an SR TE path that describes the whole path including those parts that are within the egress site), it needs to carry the receiving GW's SID for X such that this SID becomes the next SID that is due to be processed before the GW completes its processing of the packet. To achieve this, each Tunnel TLV in the Tunnel Encapsulation attribute contains a Prefix-SID sub-TLV [RFC9012] for X.

As defined in [RFC9012], the Prefix-SID sub-TLV is only for IPv4/IPV6 labelled unicast routes, so the solution described in this document only applies to routes of those types. If the use of the Prefix-SID sub-TLV for routes of other types is defined in the future, further documents will be needed to describe their use for site interconnection consistent with this document.

Alternatively, if MPLS SR is in use and if the GWs for a given egress site are configured to allow GWs at remote ingress sites to perform SR TE through that egress site for a prefix X, then each GW to the egress site computes an SR TE path through the egress site to X, and places each in an MPLS label stack sub-TLV [RFC9012] in the SR Tunnel TLV for that GW.

Please refer to Section 7 of [I-D.farrel-spring-sr-domain-interconnect] for worked examples of how the SID stack is constructed in this case, and how the advertisements would work.

6. Encapsulation

If the GWs for a given site are configured to allow remote GWs to send them a packet in that site'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 Tunnel Egress 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 TLVs.

7. IANA Considerations

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 previously assigned the value 17 from this sub-registry for "SR Tunnel", referencing this document. IANA is now requested to mark that assignment as deprecated. IANA may reclaim that codepoint at such a time that the registry is depleted.

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 sites: that may mean disclosing information that is normally contained within a site. So it needs to be understood that normal security paradigms based on the boundaries of sites are weakened and interception of BGP messages may result in information being disclosed to third parties. 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 sites.

Particular exposures resulting from this work include:

- Gateways to a site will know about all other gateways to the same site. This feature applies within a site and so is not a substantial exposure, but it does mean that if the BGP exchanges

within a site can be snooped or if a gateway can be subverted then an attacker may learn the full set of gateways to a site. This would facilitate more effective attacks on that site.

- The existence of multiple gateways to a site becomes more visible across the backbone and even into remote sites. 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 site) 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.

- A node in a site that does not have external BGP peering (i.e., is not really a site 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 site) 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-site nodes.

- An external party intercepting BGP messages anywhere between sites may learn information about the functioning of the sites and the locations of end points. While this is not necessarily a significant security or privacy risk, it is possible that the disclosure of this information could be used by an attacker.

- 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 site interconnection, but are not new protocol vulnerabilities so much as new exposures of information that SHOULD be protected against using existing protocol mechanisms such as securing the TCP sessions over which the BGP messages flow. 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.

Given that the gateways and ASBRs are connected by tunnels that may run across parts of the network that are not trusted, data center operators using the approach set out in this network MUST consider using gateway-to-gateway encryption to protect the data center traffic. Additionally, due consideration MUST be given to encrypting

end-to-end traffic as it would be for any traffic that uses a public or untrusted network for transport.

9. Manageability Considerations

The principal configuration item added by this solution is the allocation of a site identifier. The same identifier MUST be assigned to every GW to the same site, and each site 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.

The site identifiers that are configured and carried in route targets (see Section 3) are an important feature to ensure that all of the gateways to a site discover each other. It is, therefore, important that this value is not misconfigured since that would result in the gateways not discovering each other and not advertising each other.

9.1. Relationship to Route Target Constraint

In order to limit the VPN routing information that is maintained at a given route reflector, [RFC4364] suggests the use of "Cooperative Route Filtering" [RFC5291] between route reflectors. [RFC4684] defines an extension to that mechanism to include support for multiple autonomous systems and asymmetric VPN topologies such as hub-and-spoke. The mechanism in RFC 4684 is known as Route Target Constraint (RTC).

An operator would not normally configure RTC by default for any AFI/ SAFI combination, and would only enable it after careful consideration. When using the mechanisms defined in this document, the operator should consider carefully the effects of filtering routes. In some cases this may be desirable, and in others it could limit the effectiveness of the procedures.

10. Acknowledgements

Thanks to Bruno Rijsman, Stephane Litkowski, Boris Hassanov, Linda Dunbar, Ravi Singh, and Daniel Migault for review comments, and to Robert Raszuk for useful discussions. Gyan Mishra provided a helpful GenArt review, and John Scudder and Benjamin Kaduk made helpful comments during IESG review.

11. References

11.1. Normative References

[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>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[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>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

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-18 (work in progress), April 2021.

[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.

[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>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <https://www.rfc-editor.org/info/rfc4684>.

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <https://www.rfc-editor.org/info/rfc5291>.

[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