LSVR K. Patel
Internet-Draft Arrcus, Inc.
Intended status: Informational A. Lindem
Expires: January 3, 2019 Cisco Systems
S. Zandi
G. Dawra
Linkedin
July 2, 2018

Usage and Applicability of Link State Vector Routing in Data Centers
draft-keyupate-lsvr-applicability-02.txt

Abstract

This document discusses the usage and applicability of Link State Vector Routing (LSVR) extensions in the CLOS architecture of Data Center Networks. The document is intended to provide a simplified guide for the deployment of LSVR extensions.

Table of Contents

1. Introduction

This document complements [I-D.ietf-lsvr-bgp-spf] by discussing the applicability of the technology in a simple and fairly common deployment scenario, which is described in Section 4.

After describing the deployment scenario, Section 5 will describe the reasons for BGP modifications for such deployments.

Once the control plane routing protocol requirements are described, Section 6 will cover the LSVR protocol enhancements to BGP to meet these requirements and their applicability to Data Center CLOS networks.

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. Recommended Reading

This document assumes knowledge of existing data center networks and data center network topologies [CLOS]. This document also assumes knowledge of data center routing protocols like BGP [RFC4271], BGP-SPF [I-D.ietf-lsvr-bgp-spf], OSPF [RFC2328], as well as, data center OAM protocols like LLDP [RFC4957] and BFD [RFC5580].

4. Common Deployment Scenario

Within a Data Center, a common network design to interconnect servers is done using the CLOS topology [CLOS]. The CLOS topology is fully non-blocking and the topology is realized using Equal Cost Multipath (ECMP). In a CLOS topology, the minimum number of parallel paths between two servers is determined by the width of a tier-1 stage as shown in the figure 1.

The following example illustrates multistage CLOS topology.

```
                                      Tier-1
                                     +-----+
                                     |NODE |
                                  +->| 12  |--+
                                  |  +-----+  |
                          Tier-2  |           |   Tier-2
                         +-----+  |  +-----+  |  +-----+
           +------------>|NODE |--+->|NODE |--+--|NODE |-------------+
           |       +-----|  9  |--+  | 10  |  +--| 11  |-----+       |
           |       |     +-----+     +-----+     +-----+     |       |
           |       |                                         |       |
           |       |     +-----+     +-----+     +-----+     |       |
           | +-----+---->|NODE |--+  |NODE |  +--|NODE |-----+-----+ |
           | |     | +---|  6  |--+->|  7  |--+--|  8  |---+ |     | |
           | |     | |   +-----+  |  +-----+  |  +-----+   | |     | |
           | |     | |            |           |            | |     | |
         +-----+ +-----+          |  +-----+  |          +-----+ +-----+
         |NODE | |NODE | Tier-3   +->|NODE |--+   Tier-3 |NODE | |NODE |
         |  1  | |  2  |             |  3  |             |  4  | |  5  |
         +-----+ +-----+             +-----+             +-----+ +-----+
           | |     | |                                     | |     | |
           A O     B O            <- Servers ->            Z O     O O
```

Figure 1: Illustration of the basic CLOS

5. Justification for BGP SPF Extension

Many data centers use BGP as a routing protocol to create an overlay as well as an underlay network for their CLOS Topologies to simplify layer-3 routing and operations [RFC7938]. However, BGP is a path-vector routing protocol. Since it does not create a fabric topology, it uses hop-by-hop EBGP peering to facilitate hop-by-hop routing to create the underlay network and to resolve any overlay next hops. The hop-by-hop BGP peering paradigm imposes several restrictions within a CLOS. It severely prohibits a deployment of Route Reflectors/Route Controllers as the EBGP sessions are inline with the data path. The BGP best path algorithm is prefix-based and it prevents announcements of prefixes to other BGP speakers until the best path decision process is performed for the prefix at each intermediate hop. These restrictions significantly delay the overall convergence of the underlay network within a CLOS.

The LSVR SPF modifications allow BGP to overcome these limitations. Furthermore, using the BGP-LS NLRI format [RFC7752] allows the LSVR data to be advertised for nodes, links, and prefixes in the BGP routing domain and used for SPF computations.

6. LSVR Applicability to CLOS Networks

With the BGP SPF extensions [I-D.ietf-lsvr-bgp-spf], the BGP best path computation and route computation are replaced with OSPF-like algorithms [RFC2328] both to determine whether an BGP-LS NLRI has changed and needs to be re-advertised and to compute the routing table. These modifications will significantly improve convergence of the underlay while affording the operational benefits of a single routing protocol [RFC7938].

Data center controllers typically require visibility to the BGP topology to compute traffic-engineered paths. These controllers learn the topology and other relevant information via the BGP-LS address family [RFC7752] which is totally independent of the underlay address families (usually IPv4/IPv6 unicast). Furthermore, in traditional BGP underlays, all the BGP routers will need to advertise their BGP-LS information independently. With the BGP SPF extensions, controllers can learn the topology using the same BGP advertisements used to compute the underlay routes. Furthermore, these data center controllers can avail the convergence advantages of the BGP SPF extensions. The placement of controllers can be outside of the forwarding path or within the forwarding path.

Alternatively, as each and every router in the BGP SPF domain will have a complete view of the topology, the operator can also choose to configure BGP sessions in hop-by-hop peering model described in

[RFC7938] along with BFD [RFC5580]. In doing so, while the hop-by-hop peering model lacks inherent benefits of the controller-based model, BGP updates need not be serialized by BGP best path algorithm in either of these models. This helps overall network convergence.

6.1. Usage of BGP-LS SAFI

The BGP SPF extensions [I-D.ietf-lsvr-bgp-spf] define a new BGP-LS SAFI for announcement of BGP SPF link-state. The NLRI format and its associated attributes follow the format of BGP-LS for node, link, and prefix announcements. Whether the peering model within a CLOS follows hop-by-hop peering described in [RFC7938] or any controller-based or route-reflector peering, an operator can exchange BGP SPF SAFI routes over the BGP peering by simply configuring BGP SPF SAFI between the necessary BGP speakers.

The BGP-LS SPF SAFI can also co-exist with BGP IP Unicast SAFI which could exchange overlapping IP routes. The routes received by these SAFIs are evaluated, stored, and announced separately according to the rules of [RFC4760]. The tie-breaking of route installation is a matter of the local policies and preferences of the network operator.

Finally, as the BGP SPF peering is done following the procedures described in [RFC4271], all the existing transport security mechanisms including [RFC5925] are available for the BGP-LS SPF SAFI.

6.1.1. Relationship to Other BGP AFI/SAFI Tuples

Normally, the BGP-LS AFI/SAFI is used solely to compute the underlay and is given preference over other AFI/SAFIs. Other BGP SAFIs, e.g., IPv6/IPv6 Unicast VPN would use the BGP-SPF computed routes for next hop resolution. However, if BGP-LS NLRI is also being advertised for controller consumption, there is no need to replicate the Node, Link, and Prefix NLRI in BGP-NLRI. Rather, additional NLRI attributes can be advertised in the BGP-LS SPF AFI/SAFI as required.

6.2. Peering Models

As previously stated, BGP SPF can be deployed using the existing peering model where there is a single hop BGP session on each and every link in the data center fabric [RFC7938]. This provides for both the advertisement of routes and the determination of link and neighboring switch availability. With BGP SPF, the underlay will converge faster due to changes in the decision process which will allow NLRI changes to be advertised faster after detecting a change.

Alternately, BFD [RFC5580] can be used to swiftly determine the availability of links and the BGP peering model can be significantly sparser than the data center fabric. BGP SPF sessions then only be established with enough peers to provide a bi-connected graph. If IEBGP is used, then the BGP routers at tier N-1 will act as route-reflectors for the routers at tier N.

6.2.1. Bi-Connected Graph Heuristic

With this heuristic, discovery of BGP peers is assumed Section 6.3. Additionally, it assumed that the direction of the peering can be ascertained. In the context of a data center fabric, direction is either northbound (toward the spine), southbound (toward the Top-Of-Rack (TOR) switches) or east-west (same level in hierarchy. The determination of the direction is beyond the scope of this document. However, it would be reasonable to assume a technique where the TOR switches can be identified and the number of hops to the TOR is used to determine the direction.

In this heuristic, BGP speakers allow passive session establishment for southbound BGP sessions. For northbound sessions, BGP speakers will attempt to maintain two northbound BGP sessions with different switches (in data center fabrics there is normally a single layer-3 connection anyway). For east-west sessions, passive BGP session establishment is allowed. However, BGP speaker will never actively establish an east-west BGP session unless it can’t establish two northbound BGP sessions.

6.3. BGP Peer Discovery

6.3.1. BGP Peer Discovery Requirements

The most basic requirement is to be able to discover the address of a single-hop peer without pre-configuration. This is being accomplished today with using IPv6 Router Advertisements (RA) [RFC4861] and assuming that a BGP sessions is desired with any discovered peer. Beyond the basic requirement, it is useful to have to following information relating to the BGP session:

o Autonomous System (AS) and BGP Identifier of a potential peer. The latter can be used for debugging and to decrease the likelihood of BGP session establishment collisions.

o Security capabilities supported and for cryptographic authentication, the security capabilities and possibly a key-chain [RFC8177] to be used.

o Session Policy Identifier - A group number or name used to associate common session parameters with the peer. For example, in a data center, BGP sessions with a Top of Rack (ToR) device could have parameters than BGP sessions between leaf and spine.

In a data center fabric, it is often useful to know whether a peer is southbound (towards the servers) or northbound (towards the spine or super-spine) Section 6.2.1. A potential requirement would also be to determine this dynamically. One mechanism, without specifying all the details, might be for the ToRs to be identified when installed and for the others switches in the fabric to determine their level based on the distance from the closest ToR.

If there are multiple links between BGP speakers or the links between BGP speakers are unnumbered, it is also useful to be able to establish multi-hop sessions using the loopback addresses. This will often require the discovery protocol to install route(s) toward the potential peer loopback addresses prior to BGP session establishment.

Finally, a simple BGP discovery protocol could also be used to establish a multi-hop session with one or more controllers by advertising connectivity to one or more controllers. However, once the multi-hop session actually traverses multiple nodes, it is bordering a distance-vector routing protocol and possibly this is not a good requirement for the discovery protocol.

6.3.2. BGP Peer Discovery Alternatives

While BGP peer discovery is not part of [I-D.ietf-lsvr-bgp-spf], there are, at least, three proposals for BGP peer discovery. At least one of these mechanisms will be adopted and will be applicable to deployments other than the data center. It is strongly RECOMMENDED that the accepted mechanism be used in conjunction with BGP SPF in data centers. The BGP discovery mechanism should discovery both peer addresses and endpoints for BFD discovery. Additionally, it would be great if there were a heuristic for determining whether the peer is at a tier above or below the discovering BGP speaker (refer to Section 6.2.1).

The BGP discovery mechanisms under consideration are [I-D.acee-idr-lldp-peer-discovery], [I-D.xu-idr-neighbor-autodiscovery], and [I-D.ymbk-lsvr-lsoe].

6.3.3. Data Center Interconnect (DCI) Applicability

Since BGP SPF is to be used for the routing underlay and DCI gateway boxes typically have direct or very simple connectivity, BGP external sessions would typically not include the BGP SPF SAFI.

6.4. Non-CLOS/FAT Tree Topology Applicability

The BGP SPF extensions [I-D.ietf-lsvr-bgp-spf] can be used in other topologies and avail the inherent convergence improvements. Additionally, sparse peering techniques may be utilized Section 6.2. However, determining whether or to establish a BGP session is more complex and the heuristic described in Section 6.2.1 cannot be used. In such topologies, other techniques such as those described in [I-D.li-dynamic-flooding] may be employed. One potential deployment would be the underlay for a Service Provider (SP) backbone where usage of a single protocol, i.e., BGP, is desired.

7. IANA Considerations

No IANA updates are requested by this document.

8. Security Considerations

This document introduces no new security considerations above and beyond those already specified in the [RFC4271] and [I-D.ietf-lsvr-bgp-spf].

9. Acknowledgements

The authors would like to thank Alvaro Retana and Yan Filyurin for the review and comments.

10. References

10.1. Normative References

[I-D.ietf-lsvr-bgp-spf]
Patel, K., Lindem, A., Zandi, S., and W. Henderickx, "Shortest Path Routing Extensions for BGP Protocol", draft-ietf-lsvr-bgp-spf-01 (work in progress), May 2018.

[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>.

[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>.

10.2. Informative References

[CLOS] "A Study of Non-Blocking Switching Networks", The Bell System Technical Journal, Vol. 32(2), DOI 10.1002/j.1538-7305.1953.tb01433.x, March 1953.

[I-D.acee-idr-lldp-peer-discovery] Lindem, A., Patel, K., Zandi, S., Haas, J., and X. Xu, "BGP Logical Link Discovery Protocol (LLDP) Peer Discovery", draft-acee-idr-lldp-peer-discovery-03 (work in progress), June 2018.

[I-D.li-dynamic-flooding] Li, T. and P. Psenak, "Dynamic Flooding on Dense Graphs", draft-li-dynamic-flooding-05 (work in progress), June 2018.

[I-D.xu-idr-neighbor-autodiscovery] Xu, X., Bi, K., Tantsura, J., Triantafillis, N., and K. Talaulikar, "BGP Neighbor Autodiscovery", draft-xu-idr-neighbor-autodiscovery-08 (work in progress), May 2018.

[I-D.ymbk-lsvr-lsoe] Bush, R. and K. Patel, "Link State Over Ethernet", draft-ymbk-lsvr-lsoe-00 (work in progress), March 2018.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[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>.

[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>.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <https://www.rfc-editor.org/info/rfc4861>.

[RFC4957] Krishnan, S., Ed., Montavont, N., Njedjou, E., Veerepalli, S., and A. Yegin, Ed., "Link-Layer Event Notifications for Detecting Network Attachments", RFC 4957, DOI 10.17487/RFC4957, August 2007, <https://www.rfc-editor.org/info/rfc4957>.

[RFC5580] Tschofenig, H., Ed., Adrangi, F., Jones, M., Lior, A., and B. Aboba, "Carrying Location Objects in RADIUS and Diameter", RFC 5580, DOI 10.17487/RFC5580, August 2009, <https://www.rfc-editor.org/info/rfc5580>.

[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>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[RFC8177] Lindem, A., Ed., Qu, Y., Yeung, D., Chen, I., and J. Zhang, "YANG Data Model for Key Chains", RFC 8177, DOI 10.17487/RFC8177, June 2017, <https://www.rfc-editor.org/info/rfc8177>.

Authors' Addresses

Keyur Patel
Arrcus, Inc.
2077 Gateway Pl
San Jose, CA 95110
USA

Email: keyur@arrcus.com

Acee Lindem
Cisco Systems
301 Midenhall Way
Cary, NC 95110
USA

Email: acee@cisco.com

Shawn Zandi
Linkedin
222 2nd Street
San Francisco, CA 94105
USA

Email: szandi@linkedin.com

Gaurav Dawra
Linkedin
222 2nd Street
San Francisco, CA 94105
USA

Email: gdawra@linkedin.com