Use of BIER Entropy for Data Center CLOS Networks
draft-ietf-bier-entropy-staged-dc-clos-00

Abstract

   Bit Index Explicit Replication (BIER) introduces a new multicast-
   specific BIER Header. BIER can be applied to the Multi Protocol
   Label Switching (MPLS) data plane or Non-MPLS data plane. Entropy is
   a technique used in BIER to support load-balancing. This document
   examines and describes how BIER Entropy is to be applied to Data
   Center CLOS networks for path selection.

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 [RFC2119].

Status of This Memo

Table of Contents

1. Introduction

   Bit Index Explicit Replication (BIER) [RFC8279] is an architecture
   that provides optimal multicast forwarding without requiring
   intermediate routers to maintain any per-flow state by using a
   multicast-specific BIER header. [RFC8296] defines two types of BIER
   encapsulation formats: one is MPLS encapsulation, the other is non-
   MPLS encapsulation. Entropy is a technique used in BIER to support
   load-balancing. This document examines and describes how BIER
   Entropy is to be applied to Data Center CLOS networks for path
   selection.

2. Terminology

Readers of this document are assumed to be familiar with the
terminology and concepts of the documents listed as Normative
References.

3. Problem Statement and Considerations

3.1. Problem Statement

A common choice for a horizontally scalable topology used in Data
Center is a CLOS topology. This topology features an odd number of
stages, for example, a 5-Stage CLOS Topology as a example in
[RFC7938].

ECMP is the fundamental load-sharing mechanism used by a CLOS
topology. Effectively, every lower-tier device will use all of its
directly attached upper-tier devices to load-share traffic destined
to the same IP prefix. The number of ECMP paths between any two Tier
3 devices in CLOS topology is equal to the number of the devices in
the middle stage (Tier 1). For example, Figure 1 illustrates a
topology where Tier 3 device L1 has four paths to reach servers X and
Y, via Tier 2 devices S1 and S2 and then Tier 1 devices S11, S12, S21
and S22 respectively.

```
                                     Tier 1
                                    +-----+
            Cluster                 |SUPER|
   +---------------------------+  +--| S11 |--+
   |                           |  |  +-----+  |
   |                Tier 2     |  |           |       Tier 2
   |               +-----+     |  |  +-----+  |    +-----+
   |   +-------------|SPINE|------+--|SUPER|--+--|SPINE|-------------+
   |   |       +-----| S1  |------+  | S12 |  +--|  S3 |-----+       |
   |   |       |     +-----+     |   +-----+     +-----+     |       |
   |   |       |                 |                           |       |
   |   |       |     +-----+     |   +-----+     +-----+     |       |
   |   | +-----------|SPINE|------+  |SUPER|   +--|SPINE|-----------+ |
   |   | |     | +---| S2  |------+--| S21 |--+--|  S4 |---+ |     | |
   |   | |     | |   +-----+     |   | +-----+ |  +-----+   | |     | |
   |   | |     | |               |   |         |           | |     | |
   |  +-----+ +-----+            |   | +-----+ |      +-----+ +-----+
   |  | LEAF| | LEAF|            |  +--|SUPER|--+      | LEAF| | LEAF|
   |  | L1  | | L2  | Tier 3     |     | S22 |    Tier 3 | L3  | | L4  |
   |  +-----+ +-----+            |     +-----+           +-----+ +-----+
   |   | |     | |               |                        | |     | |
   |   0 0     0 0               |                        X Y     0 0
   |       Servers               |                           Servers
   +---------------------------+
```
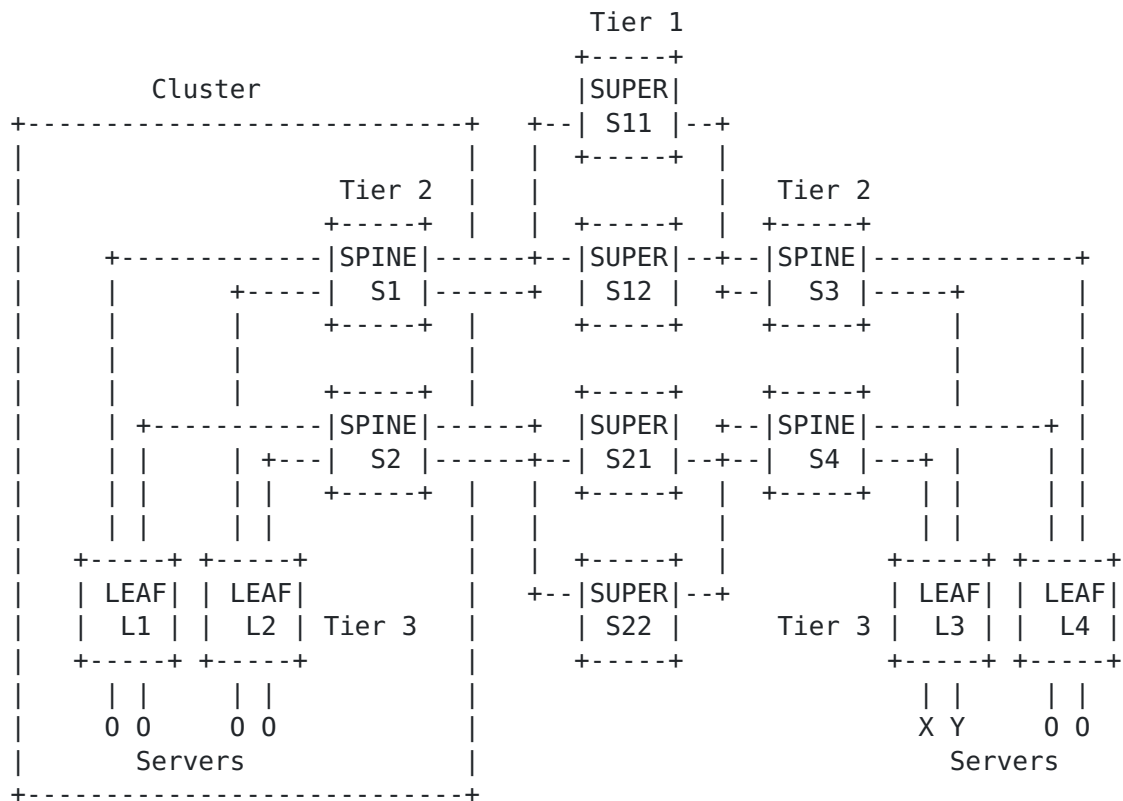
Figure 1: 5-Stage CLOS Topology

When BIER is deployed in a multi-tenant data center network
environment for efficient delivery of Broadcast, Unknown-unicast and
Multicast (BUM) traffic, a network operator may want a deterministic
path for every packet. For example, when L1 needs to send a BUM
packet to L3 and L4, which are in different SIs, L1 has to send the
packet twice, and expects the packet along two deterministic paths of
L1->S1->S11-->L3 and L1->S2->S21-->L4 seperately. Another example of
using a deterministic path in a DC is for per-flow steering of
"elephant" flows defined in [I-D.ietf-spring-segment-routing-msdc].

A deterministic path for a multicast packet, with multiple staged
equal cost paths, is comparable to a traffic-engineering path defined
in [I-D.ietf-mpls-spring-entropy-label] for a unicast path with
multiple hop equal cost paths.

3.2. Considerations

The idea behind entropy is that the ingress router computes a hash
based on several fields from a given packet and places the result in
an additional label, named "entropy label". Then this entropy label
can be used as part of the hash keys used by an transit router. When

entropy label is used, the keys used in the hashing functions are still a local configuration matter. A router may soley use the entropy label or use a combination of multiple fields from the incoming packet. The hashing function is to randomly load balance the mass of flows between the small number of equal cost paths.

If one wants, however, to get a deterministic path from the equal cost paths, one can use part of the 20-bit entropy field. For example, bit 0 to bit 2 of entropy label can represent a value of 0 to 7, and thus can be used to select a deterministic path from 8 equal cost paths. And thus, a 20-bit entropy label can be used by routers in different tiers to select a deterministic path independently by using different parts of the 20-bit entropy label, and form an end-to-end deterministic path.

This is simple and applicable especially for DC CLOS networks, because data delivery in DC CLOS networks for tenants is always multi-staged, with the upstream direction stages having equal cost paths.

[4]. Use of BIER Entropy for DC CLOS Network

[4.1]. Use of BIER Entropy for DC CLOS Network

Take the 5-stage CLOS network in figure 1 as an example.

Tier 2 in every cluster has N nodes, and the Tier 1 has M nodes. M is equal to N multiplied by P.

Tier 3 switches, in upstream direction, act as stage 1 of data delivery and have N equal cost paths to every BFERs in other clusters. Tier 2 switches, in upstream direction, act as stage 2 of data delivery and have P equal cost paths to every BFERs in other clusters.

Example 1: One can configure, on each Tier 3 switch, the use of bit 0 for path selection when N is equal to 2, and configure, on each Tier 2 switch, to use bit 1 for path selection when P is equal to 2.

Example 2: One can configure, on each Tier 3 switch, the use of bit 0 to bit 1 for path selection when N is equal to 4, and configure on each Tier 2 switches the use of bit 2 to bit 7 for path selection when P is equal to 48.

Assume that, each of the Tier 3 and Tier 2 switchs in the example has two parameters, X and Y, configured locally for using part of entropy label to do path selection, then in example 2:

o Each of Tier 3 (Stage 1) switches has a pair of parameters (X1=1, Y1=4)

o Each of Tier 2 (Stage 2) switches has a pair of parameters (X2=X1*Y1=4, Y2=64)

o Each of Tier 3 (Stage 1) switches populates its BIFTs for ECMP, for example, BIFT-0 to BIFT-3.

o Each of Tier 2 (Stage 2) switches populates its BIFTs for ECMP, for example, BIFT-0 to BIFT-47.

For each of Tier 3 (Stage 1) switches, each of the BIFT will have a prefered neighboring BFR. For example, LEAF L1 will have a prefered neighbor S1/S2 for BIFT-0/1 seperately, and when forming the BIFT-0 table through the underlay routing to every BFER, the prefered neighboring BFR will has a highest priority among all the locally available ECMP path.

Then an end-to-end deterministic path for a BIER packet can be had by calculating an entropy label value like this:

o Entropy = (P1-1)*X1 + (P2-1)*X2

Where P1 represents one of the Stage 1 equal cost paths with a value between 1 and N, and P2 represents one of the Stage 2 equal cost paths with a value between 1 and P.

4.2. Steering for elephant flows

One can steer an "elephant" flow to an end-to-end deterministic path, or some divided end-to-end deterministic paths across different SIs.

4.3. Path Division for Tenant flows to different SIs

When the VNEs for a tenant span multiple SIs, then it is useful to divide the BUM packets paths across different SIs.

One can configure a policy to use different paths for BIER SIs when using BIER as the BUM tunnel, on each VNE for each VNI.

4.4. Link Failure and Convergence

As stated above, each of the BIFT on a BFR will have a prefered neighboring BFR. But when the link to the prefered neighbor of some BIFT (say BIFT-X) fail, BIFT-X will converge normally, and the path of this BIFT-X will then probably not being the 'best optimized' path. For example, the link between S1 and L2 fail, then the

prefered neighbor of BIFT-0 of LEAF L1, S1, is no longer the
neighboring BFR for LEAF L2, and the flow using a Entropy using LEAF
L1's BIFT-0 will have to replicate on L1, one packet to S1 for BFER
L3 and L4, and one packet to S2 for BFER L2. If the flow changes to
use a Entropy using LEAF L1's BIFT-1, it will then be the 'best
optimized' path, because the flow doesn't have to replicate on L1,
and it need to forward only one copy to S1 for BFER L2 and L3 and L4.
Such a change to a flow's entropy is the Ingress switch's
responsibility, possibly with the assisstance of a controller.

5. Data-Plane Processing

The use of BIER entropy label to select a path between some equal
cost paths is a local configuration matter. This draft defines a
method to use part of the 20-bit entropy label in each router, and
this needs a data-plane to do some bit operation function. It is
expected to be easier than hashing function.

6. Security Considerations

This document introduces no new security considerations beyond those
already specified in [RFC8279] and [RFC8296].

7. IANA Considerations

This document contains no actions for IANA.

8. Acknowledgements

The authors wish to thank Tony Przygienda, Greg Shepherd, Alia Atlas,
Jeffery Zhang, Andrew Dolganow, and Toerless Eckert for their
reviews, comments and suggestions.

9. References

9.1. Normative References

[I-D.ietf-mpls-spring-entropy-label]
        Kini, S., Kompella, K., Sivabalan, S., Litkowski, S.,
        Shakir, R., and J. Tantsura, "Entropy label for SPRING
        tunnels", draft-ietf-mpls-spring-entropy-label-12 (work in
        progress), July 2018.

[I-D.ietf-spring-segment-routing-msdc]
        Filsfils, C., Previdi, S., Dawra, G., Aries, E., and P.
        Lapukhov, "BGP-Prefix Segment in large-scale data
        centers", draft-ietf-spring-segment-routing-msdc-10 (work
        in progress), October 2018.

   [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>.

   [RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A.,
              Przygienda, T., and S. Aldrin, "Multicast Using Bit Index
              Explicit Replication (BIER)", RFC 8279,
              DOI 10.17487/RFC8279, November 2017,
              <https://www.rfc-editor.org/info/rfc8279>.

   [RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A.,
              Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation
              for Bit Index Explicit Replication (BIER) in MPLS and Non-
              MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January
              2018, <https://www.rfc-editor.org/info/rfc8296>.

   [RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R.,
              Uttaro, J., and W. Henderickx, "A Network Virtualization
              Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365,
              DOI 10.17487/RFC8365, March 2018,
              <https://www.rfc-editor.org/info/rfc8365>.

9.2. Informative 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>.

Authors' Addresses

   Jingrong Xie
   Huawei Technologies

   Email: xiejingrong@huawei.com

   Xiaohu Xu
   Alibaba Inc.

   Email: xiaohu.xxh@alibaba-inc.com

   Gang Yan
   Huawei Technologies

   Email: yangang@huawei.com

Mike McBride
Huawei Technologies

Email: mmcbride7@gmail.com