BESS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2019

S. Mohanty
M. Misra
A. Lindem
A. Sajassi
Cisco Systems, Inc.

March 11, 2019

Weighted HRW and its applications
draft-mohanty-bess-weighted-hrw-00

Abstract

Rendezvous Hashing also known as Highest Random Weight (HRW) has been used in many load balancing applications where the central problem is how to map an object to as server such that the mapping is uniform and also minimally affected by the change in the server set. Recently, it has found use in DF election algorithms in the EVPN context and load balancing using DMZ. This draft deals with the problem of achieving load balancing with minimal disruption when the servers have different weights. It provides an algorithm to do so and also describes a few use-case scenarios where this algorithmic technique can apply.

Status of This Memo

Copyright Notice

Table of Contents

1. 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].

2. Introduction

Given an object O, a set of servers and a set of clients, a fundamental problem is how do the set of clients, independently and unanimously agree in a distributed framework, which server to assign O? This is the distributed hash table problem. The assignment should be "minimally disruptive" which means that there should be a minimal remapping of objects whenever a server is down or a new server comes up or the object set changes. This is a very common problem in practice in the Internet load balancing and web caching as described in the 'Akamai' paper [CHASH], database [DYNAMODB] and networking context.

```
                  +----+     +----+      +----+       +-----+
                  |    |     |    |      |    |       |     |
                  | S0 |     | S1 |      | S2 |       |  Sn |
                  |    |     |    |      |    |       |     |
                  +----+     +----+      +----+       +-----+
                     |          |          |             |
                     |          |          |             |
                     |          |          |             |
                     |          |          |             |
                     |          |          |             |
                     +----------+----------+-------------|


                           O0, O1, O2 ... ON
             Set of Objects need to be assigned to the set of servers.
             All the servers are of same capacities
```

Figure 1 The object to server assignment problem

Figure 1

In the Fig 1, we show a set of servers, S0,..,Sn and object pool O0,..,On and the requirement is to assign Oi to Sj such that the servers are uniformly loaded. In addition, when any server goes down or a new one is introduced, there should be minimal reassignments.

There are two standard techniques to address this problem.

1. Consistent Hashing

2. Rendezvous Hashing

3. HRW Introduction

Highest Random Weight (HRW) as defined in [HRW1999]is originally proposed in the context of Internet Caching and proxy Server load balancing. Given an object name and a set of servers, HRW maps a request to a server using the object-id (Oi) and server-id(Sj) rather than the state of the server states. HRW computes a hash, Hash(Oi, Sj) from the server-id and the object-id; this hash value can be considered as a score, and forms an ordered list of the servers based on the hash value (i.e. score) in decreasing order. The server for which the score is the highest, serves as the primary responsible for that particular object, and the server with the next highest score serves as the backup server. HRW always maps a given object object name to the same server within a given cluster; consequently it can be used at client sites to achieve global consensus on object-server mappings. When that server goes down, the backup server becomes the responsible designate.

Choosing an appropriate hash function that is statistically oblivious to the key distribution and imparts a good uniform distribution of the hash output is an important aspect of the algorithm. The original HRW [HRW1999] provides pseudorandom functions based on Unix utilities rand and srand and easily constructed XOR functions that perform considerably well. Any good uniform hash function like the Jenkins hash for instance will also work. HRW already finds use in multicast and ECMP [RFC2991],[RFC2992].

4. HRW with weights

The issue when the servers are not of the same capacity is also quite a common problem. However this problem has not gained as much attention as it should. In such a case, an obvious approach is to take the normalized weight factor into account, fi=wi/Sum(wi)and multiply the Hash(Oi, Sj) with that value i.e. the value fi*Hash(Oi, Sj). The Cache Array Routing Protocol [CARP] used this method. However there is a problem with this approach, since any change in weight of any of the servers, will result in a change in the normalized weights for everyone. This will necessitate re-computing all the weighted hash values all over again. Therefore this approach does not have the minimal disruption property of the HRW. We address this issue of the weighted HRW with minimal disruption in this draft.

Instead of re-normalizing the weights, or, in other words relatively scaling them, the approach taken by [WHRW] is to adjust the score before weighing them. When a server is added, removed or modified (its weight changes), only the score for that server changes. That server may win or lose some objects. Other servers remain affected. There is no needless transfer of objects between servers whose weight did not change. [WHRW] uses a clever way to accomplish this by defining the score function as:

1. Score(Oi, Sj) = -wi/log(Hash(Oi, Sj)/Hmax); where Hmax is the maximum hash value.

The author provides a mathematical proof as to why this choice of the Score function works with very mild assumptions on the probability distribution of the hash function.

```
                  +----+     +----+      +----+       +----+
                  |    |     |    |      |    |       |    |
                  | S0 |     | S1 |      | S2 |-------| Sn |
                  | w0 |     | w1 |      | w2 |       | wn |
                  +----+     +----+      +----+       +----+
                     |          |          |             |
                     |          |          |             |
                     |          |          |             |
                     |          |          |             |
                     |          |          |             |
                     +----------+----------+-------------|

                           O1, O2 ... ON
             Set of Objects need to be assigned to the set of servers.
             Each server is now associated with a weight
```

Figure 1 The object to server assignment problem

Figure 2

5. HRW and Consistent Hashing

HRW is not the only algorithm that addresses the object to server mapping problem with goals of fair load distribution, redundancy and fast access. There is another family of algorithms that also addresses this problem; these fall under the umbrella of the Consistent Hashing Algorithms [CHASH]. These will not be considered here.

6. Weighted HRW and its application to the EVPN DF Election

The notion and need for the Designated Forwarder is described in [RFC7432]. Consider a CE that is a host or a router that is multi-homed directly to more than one PE in an EVPN instance on a given Ethernet segment. One or more Ethernet Tags may be configured on the Ethernet segment. In this scenario only one of the PEs, referred to as the Designated Forwarder (DF), is responsible for certain actions:

a. Sending multicast and broadcast traffic, on a given Ethernet Tag on a particular Ethernet segment, to the CE.

b. Flooding unknown unicast traffic (i.e. traffic for which an PE does not know the destination MAC address), on a given Ethernet Tag on a particular Ethernet segment to the CE, if the environment requires flooding of unknown unicast traffic.

```
                                +---------------+
                                |   IP/MPLS     |
                                |   CORE        |
                  +----+ ES1 +----+           +----+
                  | CE1|-----|    |-----------|    |____ES2
                  +----+     | PE1|           | PE2|    \
                             |    |--------   +----+     \+----+
                             +----+        |    |         | CE2|
                                |          |  +----+     /+----+
                                |          |__|    |____/   |
                                |             | PE3|    ES2 /
                                |             +----+       /
                                |               |         /
                                +-------------+----+     /
                                              | PE4|____/ES2
                                              |    |
                                              +----+
```

Figure 3 Multi-homing Network of EVPN

Figure 3

Figure 3 illustrates a case where there are two Ethernet Segments, ES1 and ES2. PE1 is attached to CE1 via Ethernet Segment ES1 whereas PE2, PE3 and PE4 are attached to CE2 via ES2 i.e. PE2, PE3 and PE4 form a redundancy group. Since CE2 is multi-homed to different PEs on the same Ethernet Segment, it is necessary for PE2, PE3 and PE4 to agree on a DF to satisfy the above mentioned requirements.

The use of HRW in the EVPN DF Election is described in [I-D.ietf-bess-evpn-df-election-framework]. In that draft it is explained how the HRW DF Election performs better than the modulo DF Election algorithm in [RFC7432]. However, it is implicitly assumed there that all the PEs are of the same capacity (weights equal).

DMZ link bandwidth for load balancing flows across multiple EBGP egress points is described in [I-D.ietf-idr-link-bandwidth]. It has been extended to the case of cumulative DMZ load balancing [I-D.mohanty-bess-ebgp-dmz] in the case of an all EBGP network in the data center. [I-D.ietf-bess-evpn-unequal-lb] describes the use of the DMZ in the EVPN DF Election. The argument is made that ideally one should be able to change the link bandwidth in one or more of the multi-homed PEs rather than have to change in all of the multi-homed PEs simultaneously. The draft describes the bandwidth increments to be taken into consideration and proposes an iterative way to assign the score function. The description in Section 4.3.2 of [I-D.ietf-bess-evpn-unequal-lb] is an non-optimal solution and somewhat empirical. It does not obey the minimal disruption property of the HRW.

In contrast to the procedures for weighted HRW in 4.3.2 of [I-D.ietf-bess-evpn-unequal-lb], we can achieve an optimal solution for weighted HRW in [I-D.ietf-bess-evpn-unequal-lb] using the score function as described in Section 4 above and obviating the need to take bandwidth increments. It is an order of magnitude faster and efficient and minimally disruptive.

7. Weighted HRW and its application to Resilient Hashing

With the exponential increase in the number of physical links used in data centers, there is also the potential for an increase in the number of failed physical links. In systems that employ static hashing for load balancing flows across members of port channels or Equal Cost Multipath (ECMP) groups, each flow is hashed to a link. When a link fails, all flows including those that were previously mapped to the non-failed links are rehashed across the remaining working links. This causes packet reordering of flows that were in fact not mapped to the link that failed. A similar rehashing with packet re-ordering also happens when a link is added to the port channel or Equal Cost Multipath (ECMP) group. With the ever increasing number of physical links used in the data centers there the possibility for increasing number of failed links only increases. Hence the resilient hashing is very important.

However when the links are not of the same speed, Resilient hashing for ECMP does not apply per-se. However, one can use the method explained in Section 4 to achieve resilient hashing even in the Unequal Cost Multipath (UCMP)case or when member links are of different bandwidths.

8. Weighted HRW and its application to Multicast DR Election

[I-D.mankamana-pim-bdr]propose a mechanism to elect backup DR on a shared LAN. A backup DR on LAN would be useful for faster convergence. When the access bandwidth is different for the PIM routers and we want to do a load balancing among the PIM routers for DR/backup DR functionality with regards to the various (S,G) flow, technique similar to Section 4 can be applied. The details of the problem is out of the scope of the current draft and is being worked on separately at this time.

9. Protocol Considerations

A request needs to registered with IANA registry for the weighted HRW EVPN DF Election Algorithm in the DF Alg field in the DF Election Extended Community in draft [I-D.ietf-bess-evpn-df-election-framework].

10. Operational Considerations

TBD.

11. Security Considerations

This document raises no new security issues for EVPN.

12. Acknowledgements

The authors would like to thank Shyam Sethuram and Peter Psenak for useful discussions related to this draft.

13. References

13.1. Normative References

[HRW1999] Thaler, D. and C. Ravishankar, "Using Name-Based Mappings to Increase Hit Rates", IEEE/ACM Transactions in networking Volume 6 Issue 1, February 1998.

[I-D.ietf-bess-evpn-df-election-framework] Rabadan, J., satyamoh@cisco.com, s., Sajassi, A., Drake, J., Nagaraj, K., and S. Sathappan, "Framework for EVPN Designated Forwarder Election Extensibility", draft-ietf-bess-evpn-df-election-framework-09 (work in progress), January 2019.

[I-D.ietf-bess-evpn-unequal-lb] Malhotra, N., Sajassi, A., Rabadan, J., Drake, J., Lingala, A., and S. Thoria, "Weighted Multi-Path Procedures for EVPN All-Active Multi-Homing", draft-ietf-bess-evpn-unequal-lb-00 (work in progress), September 2018.

[I-D.ietf-idr-extcomm-iana] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", draft-ietf-idr-extcomm-iana-02 (work in progress), December 2013.

[I-D.ietf-idr-link-bandwidth]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", draft-ietf-idr-link-bandwidth-07 (work in progress), March 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>.

[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>.

[RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <https://www.rfc-editor.org/info/rfc7432>.

[WHRW] Resch, J., "New hashing Algorithms for Data Storage", Storage Developer Conference 18, November 2015.

13.2. Informative References

[CARP] Valloppillil, V. and K. Ross, "Cache Array Routing Protocol v1.1", IEEE/ACM Transactions in networking Volume 6 Issue 1, February 1998.

[CHASH] Karger, D., Lehman, E., Leighton, T., Panigrahy, R., Levine, M., and D. Lewin, "Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web", ACM Symposium on Theory of Computing ACM Press New York, May 1997.

[CLRS2009]
Cormen, T., Leiserson, C., Rivest, R., and C. Stein, "Introduction to Algorithms (3rd ed.)", MIT Press and McGraw-Hill ISBN 0-262-03384-4., February 2009.

[DYNAMODB]
Decennia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., and W. Vogels, "Dynamo: Amazon’s Highly Available Key-value Store", SOSP 07, October 2007.

[I-D.mankamana-pim-bdr]
mishra, m., "PIM Backup Designated Router Procedure", draft-mankamana-pim-bdr-00 (work in progress), June 2018.

[I-D.mohanty-bess-ebgp-dmz]
satyamoh@cisco.com, s., Millisor, A., and A. Vayner, "Cumulative DMZ Link Bandwidth and load-balancing", draft-mohanty-bess-ebgp-dmz-00 (work in progress), March 2018.

[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, DOI 10.17487/RFC2991, November 2000, <https://www.rfc-editor.org/info/rfc2991>.

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <https://www.rfc-editor.org/info/rfc2992>.

Authors' Addresses

Satya Ranjan Mohanty
Cisco Systems, Inc.
225 West Tasman Drive
San Jose, CA 95134
USA

Email: satyamoh@cisco.com

Mankamana Misra
Cisco Systems, Inc.
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: mankamis@cisco.com

Acee Lindem
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: acee@cisco.com

Ali Sajassi
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: sajassi@cisco.com