Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 14, 2020

X. Xu
Alibaba, Inc
L. Fang
Expedia, Inc
J. Tantsura
Apstra, Inc.
S. Ma
Juniper

October 12, 2019

OSPF Flooding Reduction in Massively Scale Data Centers (MSDCs)
draft-xu-lsr-ospf-flooding-reduction-in-msdc-03

Abstract

OSPF is one of the used underlay routing protocol for MSDC (Massively Scalable Data Center) networks. For a given OSPF router within the CLOS topology, it would receive multiple copies of exactly the same LSA from multiple OSPF neighbors. In addition, two OSPF neighbors may send each other the same LSA simultaneously. The unnecessary link-state information flooding wastes the precious process resource of OSPF routers greatly due to the presence of too many OSPF neighbors for each OSPF router within the CLOS topology. This document proposes extensions to OSPF so as to reduce the OSPF flooding within such MSDC networks. The reduction of the OSPF flooding is much beneficial to improve the scalability of MSDC networks. These modifications are applicable to both OSPFv2 and OSPFv3.

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

time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 14, 2020.

Copyright Notice

Table of Contents

1. Introduction

OSPF is commonly used as an underlay routing protocol for Massively Scalable Data Center (MSDC) networks where CLOS is the most popular topology. MSDCs are also called Large-Scale Data Centers.

For a given OSPF router within the CLOS topology, it would receive multiple copies of exactly the same LSA from multiple OSPF neighbors. In addition, two OSPF neighbors may send each other the same LSA simultaneously. The unnecessary link-state information flooding significantly wastes the precious process resource of OSPF routers and therefore OSPF could not scale very well in MSDC networks.

To simplify the network management task, centralized controllers are becoming fundamental network elements in most MSDCs. One or more controllers are usually connected to all routers within an MSDC network via a Local Area Network (LAN) which is dedicated for network management purpose (called management LAN), as shown in Figure 1.

```
        +----------+                  +----------+
        |Controller|                  |Controller|
        +----+-----+                  +-----+----+
             |DR                            |BDR
             |                              |
             |                              |
---+---------+---+----------+-----------+---+---------+-Management LAN
   |             |          |           |             |
   |Non-DR       |Non-DR    |Non-DR     |Non-DR       |Non-DR
   |             |          |           |             |
   |         +---+--+       |       +---+--+          |
   |         |Router|       |       |Router|          |
   |         *------*-      |      /*---/--*          |
   |        /     \   --    |    //    /    \         |
   |        /     \     --  |  //      /    \         |
   |       /       \      --|//       /      \        |
   |       /        \      /*-       /        \       |
   |      /          \   // | --    /         \       |
   |      /          \ //   |   --  /          \      |
   |     /           /X     |     --           \      |
   |     /         //  \    |     / --          \     |
   |    /        //    \    |     /   --         \    |
   |    /      //       \   |    /      --       \    |
   |   /     //          \  |   /         --      \   |
   |   /   //             \ |  /            --     \  |
   |  /  //               \ |  /              --   \  |
 +-+- //*                +\\+-/-+               +---\-++
 |Router|                |Router|               |Router|
 +------+                +------+               +------+
```

Figure 1

With the assistance of these controllers which are acting as OSPF Designated Router (DR)/Backup Designated Router (BDR) for the management LAN, OSPF routers within the MSDC network don't need to exchange any other types of OSPF packet than the OSPF Hello packet among them. As specified in [RFC2328], these Hello packets are used for the purpose of establishing and maintaining neighbor relationships and ensuring bidirectional communication between OSPF neighbors, and even the DR/BDR election purpose in the case where those OSPF routers are connected to a broadcast network. In order to obtain the full topology information (i.e., the fully synchronized

link-state database) of the MSDC's network, these OSPF routers only need to exchange the link-state information with the controllers being elected as OSPF DR/BDR for the management LAN instead.

To further suppress the flooding of multicast OSPF packets originated from OSPF routers over the management LAN, OSPF routers would not send multicast OSPF Hello packets over the management LAN. Instead, they just wait for OSPF Hello packets originated from the controllers being elected as OSPF DR/BDR initially. Once OSPF DR/BDR for the management LAN have been discovered, they start to send OSPF Hello packets directly (as unicasts) to OSPF DR/BDR periodically. In addition, OSPF routers would send other types of OSPF packets (e.g., Database Descriptor packet, Link State Request packet, Link State Update packet, Link State Acknowledgment packet) to OSPF DR/BDR for the management LAN as unicasts as well. In contrast, the controllers being elected as OSPF DR/BDR would send OSPF packets as specified in [RFC2328]. As a result, OSPF routers within the MSDC would not receive OSPF packets from one another unless these OSPF packets are forwarded as unknown unicasts over the management LAN. Through these modifications to the legacy OSPF router behaviors, the OSPF flooding is greatly reduced, which is much beneficial to improve the overall scalability of MSDC networks. These modifications specified in this document are applicable to both OSPFv2 [RFC2328] and OSPFv3 [RFC5340].

The mechanism for OSPF refresh and flooding reduction in stable topologies as described in [RFC4136] may be considered as well.

2. Terminology

This memo makes use of the terms defined in [RFC2328].

3. Modifications to Legacy OSPF Behaviors

3.1. OSPF Routers as Non-DRs

After the exchange of OSPF Hello packets among OSPF routers, the OSPF neighbor relationship among them would transition to and remain in the 2-WAY state. OSPF routers would originate Router-LSAs and/or Network-LSAs accordingly depending upon the link-types. Note that the neighbors in the 2-WAY state would be advertised in the Router-LSAs and/or Network-LSA. This is slightly different from the legacy OSPF router behavior as specified in [RFC2328] where the neighbors in the TWO-WAY state would not be advertised. However, these self-originated LSAs need not to be exchanged directly among them anymore. Instead, these LSAs only need to be sent solely to the controllers being elected as OSPF DR/BDR for the management LAN.

To further reduce the flood of multicast OSPF packets over the management LAN, OSPF routers SHOULD send OSPF packets as unicasts. More specifically, OSPF routers SHOULD send unicast OSPF Hello packets periodically to the controllers being elected as OSPF DR/BDR. In other words, OSPF routers SHOULD NOT send any OSPF Hello packet over the management LAN until they have found an OSPF DR/BDR for the management LAN. Note that OSPF routers, within the MSDC, SHOULD NOT be elected as OSPF DR/BDR for the management LAN (This is done by setting the Router Priority of those OSPF routers to zero). As a result, OSPF routers would not see each other over the management LAN. Furthermore, OSPF routers SHOULD send all other types of OSPF packets than OSPF Hello packets to the controllers being elected as OSPF DR/BDR as unicasts as well.

To avoid the data traffic from being forwarded across the management LAN, the cost of all OSPF routers' interfaces to the management LAN SHOULD be set to the maximum value.

When a given OSPF router lost its connection to the management LAN, it SHOULD actively establish FULL adjacency with all of its OSPF neighbors within the MSDC network. As such, it could obtain the full LSDB of the MSDC network while flooding its self-originated LSAs to the remaining part of the whole network. That's to say, for a given OSPF router within the MSDC network, it would not actively establish FULL adjacency with its OSPF neighbor in the 2-WAY state by default. However, it SHOULD NOT refuse to establish FULL adjacency with a given OSPF neighbors when receiving Database Description Packets from that OSPF neighbor.

3.2. Controllers as DR/BDR

The controllers being elected as OSPF DR/BDR would send OSPF packets as multicasts or unicasts as per [RFC2328]. In addition, Link State Acknowledgment packets are RECOMMENDED to be sent as unicasts rather than multicasts.

4. Acknowledgements

The authors would like to thank Acee Lindem and Mohamed Boucadair for their valuable comments and suggestions on this document.

5. IANA Considerations

TBD.

6. Security Considerations

TBD.

7. References

7.1. 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>.

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

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

7.2. Informative References

[RFC4136] Pillay-Esnault, P., "OSPF Refresh and Flooding Reduction in Stable Topologies", RFC 4136, DOI 10.17487/RFC4136, July 2005, <https://www.rfc-editor.org/info/rfc4136>.

Authors' Addresses

Xiaohu Xu
Alibaba, Inc

Email: xiaohu.xxh@alibaba-inc.com

Luyuan Fang
Expedia, Inc

Email: luyuanf@gmail.com

Jeff Tantsura
Apstra, Inc.

Email: jefftant.ietf@gmail.com

Shaowen Ma
Juniper

Email: mashao@juniper.net