Workgroup: Network Working Group
Internet-Draft:
draft-xu-lsr-isis-flooding-reduction-in-msdc-05
Published: 31 January 2024
Intended Status: Standards Track
Expires: 3 August 2024
Authors: X. Xu (China Mobile)
L. Fang (eBay)
J. Tantsura (Nvidia)
S. Ma (Google)

IS-IS Flooding Reduction in MSDC

Abstract

IS-IS is a commonly used routing protocol in MSDC (Massively Scalable Data Center) networks where CLOS is the most popular topology. In a CLOS topology, each IS-IS router would receive multiple copies of the same LSP (Link State Packet) from multiple IS-IS neighbors. Moreover, two IS-IS neighbors may send each other the same LSP simultaneously. The unnecessary link-state information flooding results in a large waste of resources for IS-IS routers, as there are too many neighbors for each router. To address this scaling problem, this document introduces some extensions to the IS-IS protocol. These extensions aim to significantly reduce the IS-IS flooding within MSDC networks, which can greatly improve the scalability of such networks.

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any 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 3 August 2024.

Copyright Notice

Table of Contents

1. Introduction

IS-IS is a commonly used routing protocol in MSDC (Massively Scalable Data Center) networks where CLOS is the most popular topology. In a CLOS topology, each IS-IS router would receive multiple copies of the same LSP (Link State Packet) from multiple IS-IS neighbors. Moreover, two IS-IS neighbors may send each other the same LSP simultaneously. The unnecessary link-state information flooding results in a large waste of resources for IS-IS routers, as there are too many neighbors for each router.

As a result, some MSDC operators had to opt for BGP as the routing protocol [RFC7938]. However, with the introduction of high-performance Ethernet networks, which are widely used in AI and high-performance computing (HPC), it has become essential to have visibility of the whole network topology and even the link capacity and load information for global load-balancing. Therefore, for large-scale AI and HPC Ethernet networks, link-state routing protocols like IS-IS should be reconsidered as the routing protocol. However, it is crucial to address the scaling issue associated with link-state routing protocols as mentioned earlier.

This document presents an effective solution to the scaling issue mentioned above. Instead of transmitting link-state information between neighboring IS-IS routers with the MSDC network fabric, link-state information originating from each IS-IS router will be gathered by centralized controllers. These controllers will then distribute

the collected link-state information to all IS-IS routers within the MSDC. As illustrated in Figure 1, all IS-IS routers in an MDSC network fabric will be linked to one or more centralized controllers through a dedicated Local Area Network (LAN). This LAN is specifically intended for link-state information collection and distribution. For redundancy purposes, there should be at least two link-state collection and distribution LANs.

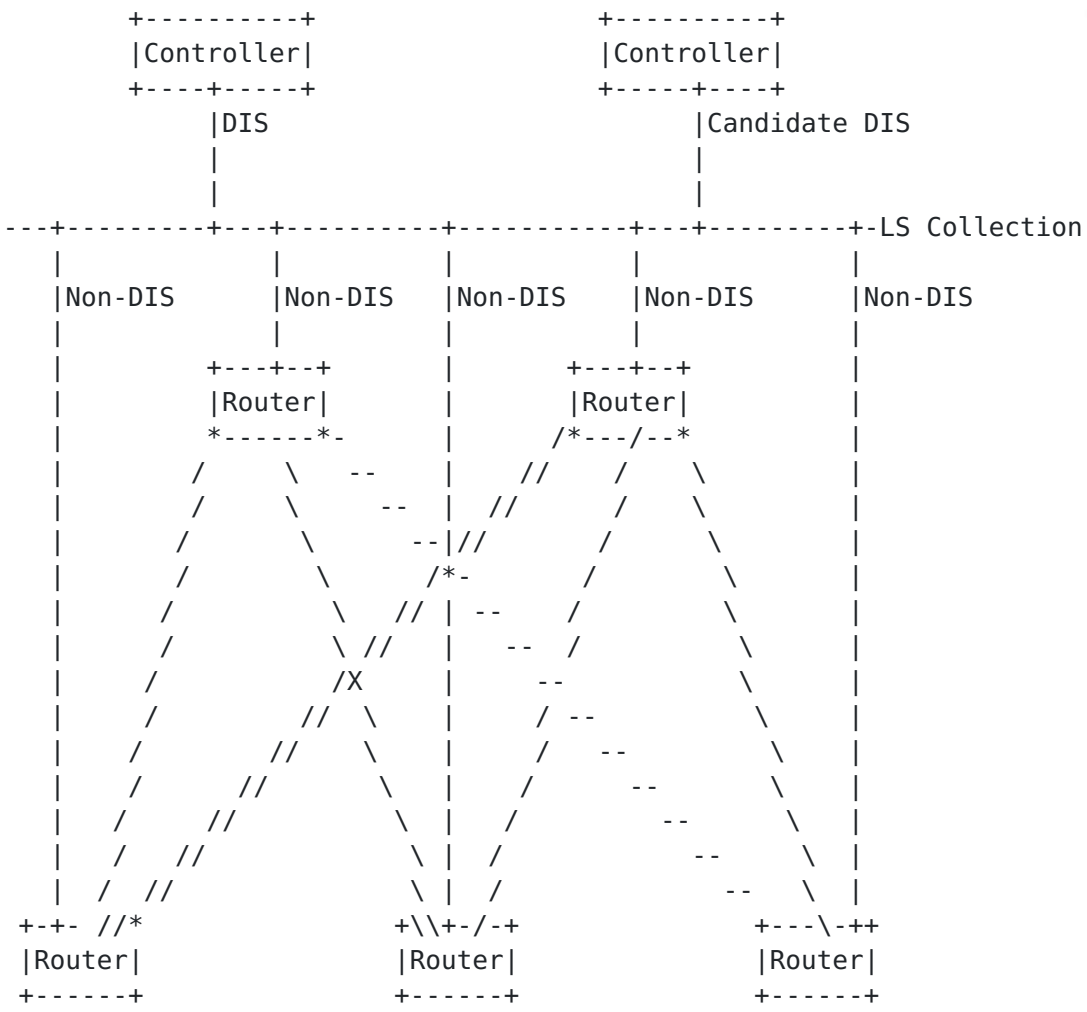

Figure 1

In the MSDC network, the IS-IS routers do not need to exchange any IS-IS Protocol Datagram Units (PDUs) other than Hello packets among them. This is due to the presence of a controller that acts as an IS-IS Designated Intermediate System (DIS) for the link-state collection and distribution LAN. To obtain the complete topology information of the MSDC network, these IS-IS routers exchange the link-state information with the controller, which is elected as IS-IS DIS for the link-state collection and distribution LAN.

To further reduce the flooding of the multicast IS-IS PDUs over the link-state collection and distribution LAN, IS-IS routers will not send multicast IS-IS Hello packets over that LAN. Instead, they will

wait for IS-IS Hello packets from the controller that has been elected as IS-IS DIS initially. Once an IS-IS DIS has been discovered, the routers will start sending IS-IS Hello packets directly to the IS-IS DIS at regular intervals as unicasts. Consequently, IS-IS routers would only form an adjacency with the IS-IS DIS over that LAN. Additionally, IS-IS routers will send IS-IS PDUs to the IS-IS DIS as unicasts. However, the IS-IS DIS will continue to send IS-IS PDUs as before. These changes to the current IS-IS router behaviors will significantly reduce IS-IS flooding and improve the scalability of MSDC networks.

2. Terminology

This memo makes use of the terms defined in [RFC1195].

3. Modifications to Current IS-IS Behaviors

3.1. IS-IS Routers as Non-DIS

IS-IS routers exchange Hello packets bidirectionally. After that, they originate Link State PDUs (LSPs) accordingly. However, these self-originated LSPs don't need to be directly exchanged between the routers. They only need to be sent to the IS-IS DIS for the link-state collection and distribution LAN. It is important to note that IS-IS routers should not be elected as IS-IS DIS for the link-state collection and distribution LAN (this can be done by setting the DIS Priority of those IS-IS routers to zero).

To further minimize the number of multicast IS-IS PDUs transmitted over the link-state collection and distribution LAN, IS-IS routers should send IS-IS PDUs as unicasts. Specifically, IS-IS routers must send unicast IS-IS Hello packets periodically to the controller elected as IS-IS DIS. This means that IS-IS routers will not send any IS-IS Hello packet over the link-state collection and distribution LAN until they have identified an IS-IS DIS for the link-state collection and distribution LAN. As a result, IS-IS routers will not discover each other over the link-state collection and distribution LAN, and will not establish adjacencies with each other. Moreover, IS-IS routers should send all types of IS-IS PDUs to the IS-IS DIS as unicasts as well.

To prevent data traffic from being forwarded across the link-state collection and distribution LAN, the interfaces of all IS-IS routers to the LAN must be set to the maximum cost value.

3.2. Controllers as DIS

When a controller is elected as the IS-IS DIS, it would send IS-IS PDUs as multicasts or unicasts as normal. Additionally, it is required to accept and process those unicast IS-IS PDUs originated from other IS-IS routers. Upon receiving any new LSP from a given IS-IS router, the DIS must flood it immediately to the link-state

collection and distribution LAN. This serves two purposes: 1) to acknowledge the receipt of that LSP implicitly, and 2) to synchronize that LSP to all other IS-IS routers.

To reduce the frequency of advertising the Complete Sequence Number PDU (CSNP) on the DIS for the link-state collection and distribution LAN, it is recommended that IS-IS routers send an explicit acknowledgement with a Partial Sequence Number PDU (PSNP) upon receiving a new LSP from that DIS.

4. Acknowledgements

The authors would like to thank Peter Lothberg and Erik Auerswald for their valuable comments and suggestions on this document.

5. IANA Considerations

TBD.

6. Security Considerations

TBD.

7. References

7.1. Normative References

[RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, DOI 10.17487/RFC1195, December 1990, <https://www.rfc-editor.org/info/rfc1195>.

[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>.

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>.

[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>.

Authors' Addresses

Xiaohu Xu
China Mobile

Email: xuxiaohu_ietf@hotmail.com

Luyuan Fang
eBay

Email: luyuanf@gmail.com

Jeff Tantsura
Nvidia

Email: jefftant.ietf@gmail.com

Shaowen Ma
Google

Email: shaowen@google.com