Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 16, 2021

H. Chen
Futurewei
M. Toy
Verizon
A. Wang
China Telecom
L. Liu
Fujitsu
X. Liu
Volta Networks
March 15, 2021

IGP for Network High Availability
draft-chen-lsr-ctr-availability-02

Abstract

This document describes protocol extensions to OSPF and IS-IS for improving the reliability or availability of a network controlled by a controller cluster.

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 September 16, 2021.

Table of Contents

1. Introduction

More and more networks are controlled by central controllers or controller clusters. A controller cluster is a single controller externally. It normally consists of two or more controllers internally working together to control a network, i.e., every network element (NE) in the network. The reliability or availability of a network is heavily dependent on its controller cluster. The issues or failures in the controller cluster may impact the reliability or availability of the network greatly.

For a controller cluster comprising two or more controllers (i.e., primary controller, secondary controller, and so on), the failures in the cluster may split the cluster into a few of separated controller

groups. These groups do not know each other and may be out of synchronization. Two or more groups may be elected to control the network at the same time, which may cause some issues.

This document proposes some procedures and extensions to OSPF and IS-IS for the separated controllers or controller groups to know each other thus elect one new primary controller or controller group correctly when the cluster is split because of failures in the cluster.

2. Terminologies

The following terminologies are used in this document.

IGP: Interior Gateway Protocol

OSPF: Open Shortest Path First

IS-IS: Intermediate System to Intermediate System

LSA: Link State Advertisement in OSPF

LSP: Link State Protocol PDU in IS-IS

PDU: Protocol Data Unit

LS: Link Sate, which is LSA in OSPF or LSP in IS-IS

NE: Network Element

CE: Customer Edge

PE: Provider Edge

3. IGP for Controller Cluster Reliability

This section briefs the mechanism of controller cluster reliability or availability using IGP, and illustrates some details through a simple example.

3.1. Overview of Mechanism

When a cluster of controllers is split into a few of separated groups because of failures in the cluster, the live controllers are still actually connected to the network (i.e., network elements). Through some of these connections, each group can get the information about the other groups. A new primary controller or controller group is correctly elected to control the network based on the information.

Each controller may comprise an IGP as an information proxy, called IGP information proxy or IGP for short. The IGP has an IGP adjacency relation with each of a given number of NEs (such as one NE) in the network. When one adjacency is broken, a new adjacency is created and maintained if possible. The given number of adjacency relations is retained.

In normal operations, the cluster has all its controllers connected. They are the primary controller controlling the network, the secondary controller, and so on. They have current position 1, 2, and so on respectively. The primary controller advertises the information about the controllers via its IGP adjacencies. The extensions to IGP below is used.

When the cluster is split into a few separated groups, each group elects an intent primary controller, secondary controller and so on from the group, which have intent position 1, 2, and so on respectively. The intent primary controller advertises the information about the controllers in the group.

The information advertised by the (intent) primary controller includes its current (intent) position, its old position, its priority to become a primary controller, the number of controllers, and the IDs of the controllers which are ordered according to their (intent) positions. In addition, a flag C indicating that whether it is Controlling the network (i.e., it is the primary controller or intent primary controller) is included.

3.2. Example

Figure 1 shows a controller cluster comprising two controllers: the primary controller and the secondary controller. Each controller includes an IGP as an information proxy.

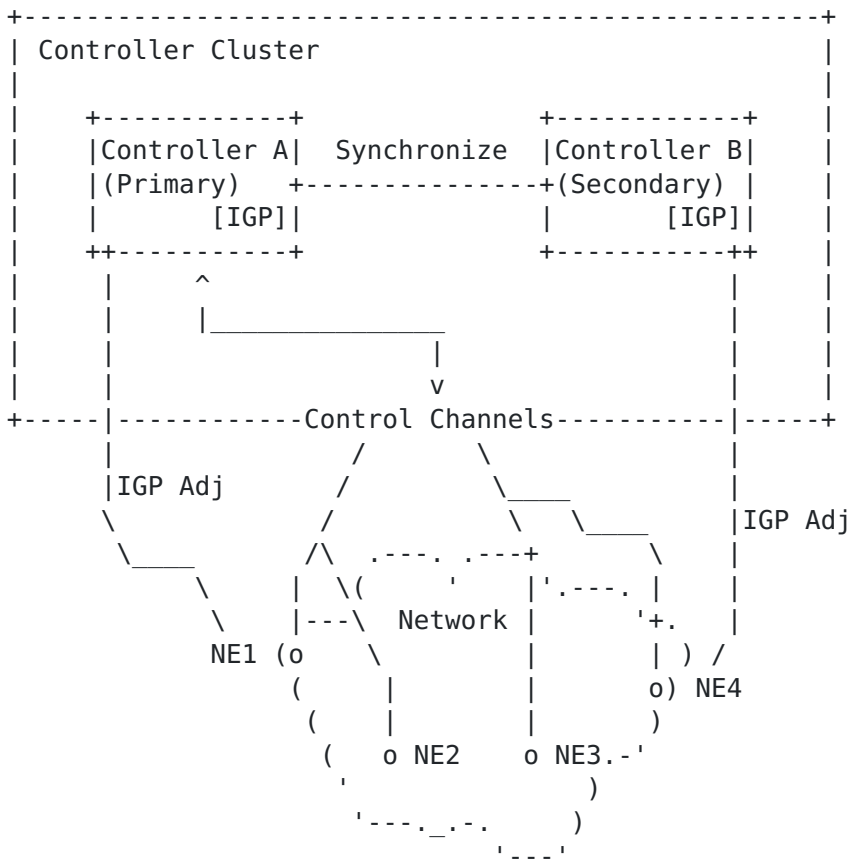

Figure 1: Controller Cluster of 2 Controllers

The IGP in a controller has one IGP adjacency relation with one NE in the network. In Figure 1, the IGP in controller A has IGP adjacency with NE1, the IGP in B has IGP adjacency with NE4.

In normal operations, the IGP of the primary controller originates link state (LS) containing the information about the controllers connected to it. The LS originated by Controller A (Primary) in Figure 1 having the following contents:

C = 1, A's current Position = 1, A's OldPosition = 1, A's Priority, NoControllers = 2, A's ID, B's ID

When failures happen in the cluster, the live controllers act as follows:

For the Secondary Controller (e.g., B) alive, if the primary controller is dead, it promotes itself as the new primary controller; if the primary controller is alive but separated from the secondary

controller, the secondary controller will not promote itself to be a new primary controller.

For the Primary Controller (e.g., A), if it is alive, it continues to be the primary controller.

With the extensions to IGP, the secondary controller can determine the status of the primary controller through using IGP and obtaining the information about the primary controller. The conditions that the primary controller is alive but separated from the secondary controller (i.e., condition a: the connection between the primary controller and the secondary controller in the cluster failed, but condition b: the two controllers are alive) can be determined by the secondary controller as follows:

For condition a, when the heartbeat from the primary stops, the secondary knows that the connection between the primary and secondary controller failed.

For condition b, it checks its link state database (LSDB) in the IGP to see whether the IGP for the primary controller is connected to some network elements and advertises the LS. If so, the primary controller is alive; otherwise, it is dead.

4. Extensions to IGP

This section describes extensions to OSPF and IS-IS.

4.1. Extensions to OSPF

A new TLV, called OSPF Controllers TLV, is defined. When OSPF acts as a proxy of a controller in a cluster, it may advertise the information about the controllers such as the number of controllers connected to it (including itself) in its router information LSA, which contains a Controllers TLV of the following format.

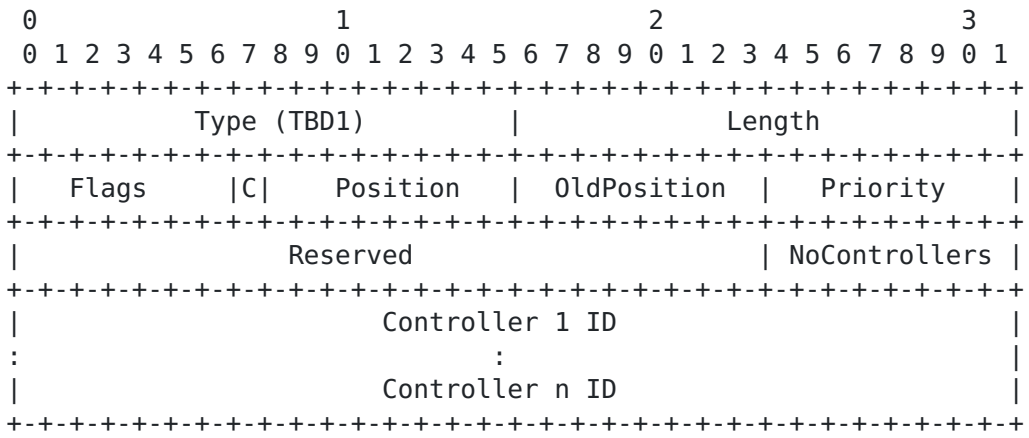

Figure 2: OSPF Controllers TLV

Type: TBD1 is to be assigned by IANA.

Length: It indicates the length of the value portion in octets.

Flag (8 bits): One flag bit, C-bit, is defined. When set, it indicates that the position is the position of the current active primary controller. In this case, C = 1 and Position = 1, which indicate that the controller is the current active primary controller controlling the network.

Position (8 bits): It indicates the current/intent position of the controller in the controller cluster or group. 1: primary (first) controller, 2: secondary controller, 3: third controller, and so on (i.e., Controller Position of value n: n-th controller in the cluster or group).

OldPosition (8 bits): It indicates the old position of the controller in the controller cluster before it is split.

Priority (8 bits): It indicates the priority of the controller to be elected as a primary controller.

Reserved (24 bits): Reserved field, must set to zero for transmission and ignored for reception.

NoControllers (8 bits): It indicates the number of controllers connected to the controller advertising the TLV.

Controller i ID (32 bits): It represents the identifier (ID) of controller i at position i (i = 1, ..., n) in the cluster or group.

When the information about the controllers is changed, OSPF of a primary controller originates an OSPF Router Information Opaque LSA, which includes a OSPF Controllers TLV.

4.2. Extensions to IS-IS

Similar to OSPF, a new TLV, called IS-IS Controllers TLV, is defined. When IS-IS acts as a proxy of a controller in a cluster, it may advertise the information about the cluster such as the number of controllers connected to it (including itself) in its LSP, which contains an IS-IS Controllers TLV of the following format.

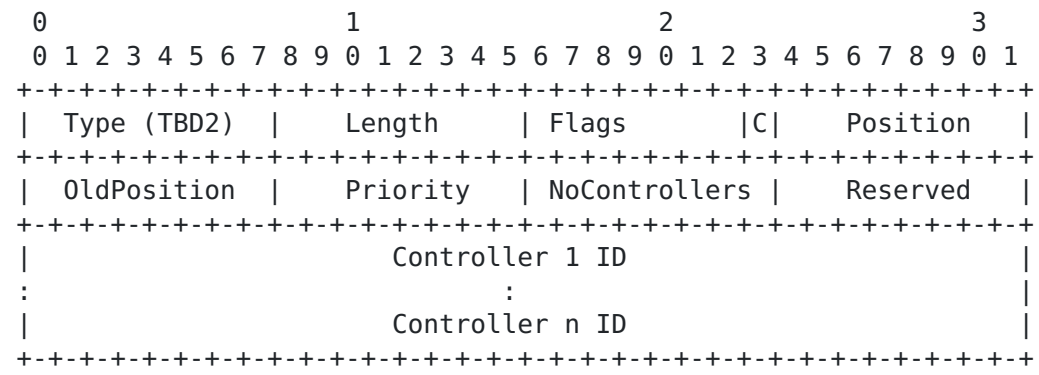

Figure 3: IS-IS Controllers TLV

Type (8 bits): TBD2 is to be assigned by IANA.

Length (8 bits): It indicates the length of the value portion in octets.

All other fields: The meaning of each of the other fields is the same as the one of the corresponding field in the OSPF Controllers TLV defined above.

When the information about the controllers is changed, the IS-IS of a primary controller originates an LSP, which includes an IS-IS Controllers TLV.

5. Recovery Procedure

This section describes the recovery procedure for a controller cluster of n (n > 2) controllers, which are the primary controller A, the secondary controller B, ..., the n-th controller N.

When failures happen in the cluster, it may be split into a few separated groups of controllers. In one policy, the group with the maximum number of controllers is responsible for controlling the

network as the primary group of the cluster, in which the new primary controller, secondary controller, and so on are elected.

For each separated group of controllers, the intent primary controller, secondary controller, and so on are elected. The intent primary controller of the group advertises the information about the group through its IGP. The information includes its intent position, its old position, its priority to become a primary controller, the number of controllers in the group, and identifiers of the controllers in the group. The identifiers of the controllers are ordered according to their positions. The identifier of the intent primary controller, which has position 1, is the first one; The identifier of the intent secondary controller, which has position 2, is the second one; and so on. Thus every separated group has the information about the other groups and can determine which group has the maximum number of controllers.

In the case of tie (i.e., two or more groups have the same maximum number of controllers), the group with the highest priority controller wins in one policy. In another policy, the group with the highest old position controller (e.g., the old primary controller) wins.

Some details of the recovery procedures in the current and intent primary controller in a controller cluster or group are as follows.

In normal operations, it advertises Controllers TLV containing:

C = 1, Position = 1, Old Position = 1, Primary Controller's priority, NoControllers = n, Primary Controller's ID, secondary controller's ID, ..., and n-th Controller's ID.

When failures cause the cluster split, it advertises Controllers TLV containing:

C = 0, Position = 1, Old Position = 1, Intent Primary Controller's priority, NoControllers = m (m is the number of controllers in the group that the primary controller is connected after the failures), Intent Primary Controller's ID, IDs of the other controllers connected.

Then after a given time, it checks if the group is elected as the primary group. If so, it advertises Controllers TLV containing:

C = 1, Position = 1, Old Position = 1, its Priority, NoControllers = m, the IDs of the controllers in the group.

One example is that failures split the cluster into two separated groups: group 1 comprising A and C, group 2 consisting of B and N. Each group elects its intent primary controller, secondary controller, and so on. Suppose that controller A and C are elected as the intent primary and secondary controller respectively in group 1; controller B and N are elected as the intent primary and secondary controller respectively in group 2.

Each of the intent primary controllers A and B advertises the information about the controllers in its group. The information advertised by A includes:

C = 0, Position = 1, OldPosition = 1, A's Priority, NoControllers = 2, A's ID, C's ID.

The information advertised by B includes:

C = 0, Position = 1, OldPosition = 2, B's Priority, NoControllers = 2, B's ID, N's ID.

Group 1 and 2 have the same number of controllers, which is 2. But OldPosition in group 1 is higher than that in group 2. Group 1 is elected as the primary group, and the intent primary controller A in the primary group is determined as the current primary controller. After the determination, the information about the controllers in group 1 (i.e., the primary group) is changed. The updated information advertised by A includes:

C = 1, Position = 1, OldPosition = 1, A's Priority, NoControllers = 2, A's ID, C's ID.

6. IANA Considerations

TBD

7. Security Considerations

TBD

8. Acknowledgements

TBD

9. References

9.1. Normative References

[ISO10589] International Organization for Standardization, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, Nov. 2002.

[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>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, Ed., "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, DOI 10.17487/RFC5329, September 2008, <https://www.rfc-editor.org/info/rfc5329>.

9.2. Informative References

[RFC4970] Lindem, A., Ed., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, DOI 10.17487/RFC4970, July 2007, <https://www.rfc-editor.org/info/rfc4970>.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Mehmet Toy
Verizon
USA

Email: mehmet.toy@verizon.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing, 102209
China

Email: wangaj3@chinatelecom.cn

Lei Liu
Fujitsu

USA

Email: liulei.kddi@gmail.com

Xufeng Liu
Volta Networks

McLean, VA
USA

Email: xufeng.liu.ietf@gmail.com