Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 9, 2020

H. Chen
Futurewei
Y. Fan
Casa Systems
A. Wang
China Telecom
L. Liu
Fujitsu
X. Liu
Volta Networks

March 8, 2020

BGP for Network High Availability
draft-chen-idr-ctr-availability-00

Abstract

This document describes protocol extensions to BGP 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 9, 2020.

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 as a single controller externally 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 as primary groups to control the network at the same time, which may cause some issues.

This document proposes some procedures and extensions to BGP 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.

BGP: Border Gateway Protocol

NE: Network Element

CE: Customer Edge

PE: Provider Edge

3. BGP for Controller Cluster Reliability

This section briefs the mechanism of controller cluster reliability or availability using BGP, 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 has a BGP session with each of a give number of the same NEs in the network and the session is established and maintained over an IP path between the controller and the NE. The session is a session of BGP with extensions.

In one example or configuration, the given number of NEs is one NE with the highest BGP ID. Suppose that node PE2 as NE has the highest BGP ID. The session between the primary controller (e.g., A) and the NE (e.g., PE2) is the session of BGP with extensions. Each of the non-primary controllers (e.g., B, C, ...) creates and maintains a BGP session with this NE (e.g., PE2).

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 BGP sessions to the given number of the same NEs.

For example, it sends the information in a BGP message to the NE (e.g., PE2), which transfers the information to each of the other controllers via the BGP sessions to the other controllers.

When the cluster is split into a few separated groups of controllers, 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 in each group advertises the information about the controllers in its group.

The information advertised by the (intent) primary controller includes its current (intent) position, its old position, its priority to become a primary controller, number of controllers in its group or cluster, and the IDs of the controllers which are ordered in 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 has a BGP session with the same NE, which is NE4.

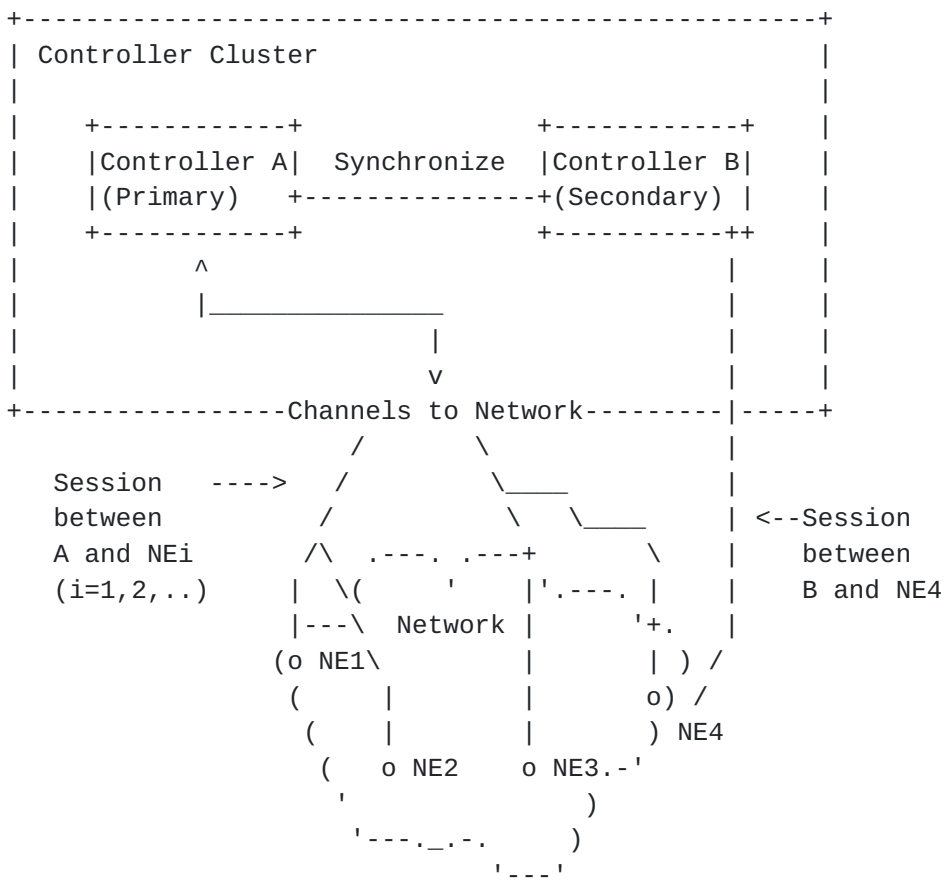

Figure 1: Controller Cluster of 2 Controllers

The primary BGP controller (i.e., A) has a BGP session with each NE in the network, including NE4. The secondary controller (i.e., B) has a BGP session with the same NE4 in the network and the session is established and maintained over an IP path between B and NE4.

In normal operations, controller A (Primary) sends NE4 a BGP message containing the information about the controllers connected to it. NE4 transfers the information to controller B (Secondary). The information includes:

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 BGP, the secondary controller can determine the status of the primary controller based on the information about the primary controller received. 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 whether the information about the primary controller is updated within a given time. If so, the primary controller is alive; otherwise, it is dead.

4. Extensions to BGP

This section describes extensions to BGP.

4.1. Capability

During a BGP session establishment, BGP Speakers advertise their support for BGP extensions for network reliability, especially the High Availability of Controller cluster (HAC). A new Controller HA Support Capability Triple is defined for HAC below. A BGP speaker indicates its support for HAC by including the triple in the Capabilities Optional Parameter in its OPEN message if it supports for HAC.

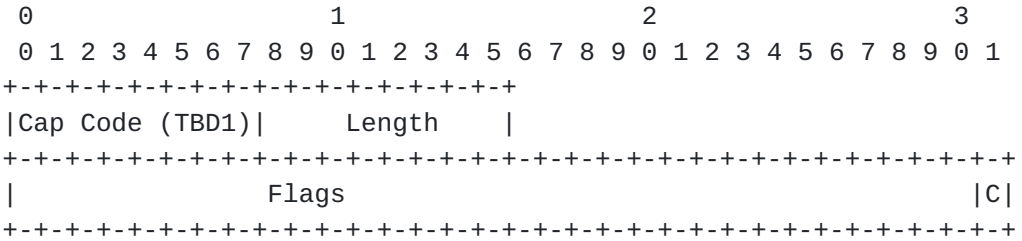

Figure 2: Controller HA Support Capability Triple

Cap Code (8 bits): TBD1 is to be assigned by IANA.

Length: It indicates the length of the Capability value portion in octets, which is 4.

Flag (32 bits): One flag bit, C-bit, is defined. When it is set to one, it indicates that the BGP speaker supports the high availability of controller cluster as a Controller. When it is set to zero, it indicates that the BGP speaker supports the high availability of controller cluster as a network element (NE).

When two BGP speakers establish a BGP session between them, each of the speakers indicates its support for HAC by including a Controller HA Support Capability Triple in the Capabilities Optional Parameter in the OPEN message if it supports for HAC.

For a BGP speaker supporting for HAC, if it receives the Controller HA Support Capability Triple in the OPEN message from the other BGP speaker over the BGP session, it records that the other BGP speaker (i.e., the other/remote end of the session) supports for HAC; otherwise, it records that the other speaker does not. Thus for all its BGP sessions, it knows whether each session's remote end BGP speaker supports for HAC. If the C-bit in the Triple is set to one, the BGP speaker is a controller; otherwise, it is a NE.

A BGP as a controller supporting for HAC acts on the information about the controllers in its cluster or group as follows:

It sends the information in a BGP UPDATE message to each of a given set of NEs that runs BGP with HAC support whenever the information changes. The given set of NEs may be the one NE with the highest BGP ID.

It adjusts the positions of the controllers accordingly whenever there is a change in the information about the controllers received from the NE supporting for HAC.

An NE running BGP with HAC support receives the information about the controllers from the BGP as a controller supporting for HAC, and sends the information to every BGP as a controller supporting for HAC and having a BGP session with the NE except for the one from which the information is received.

4.2. Controller NLRI

A new Address Family Identifier (AFI) and Sub-address Family Identifier (SAFI), called Controllers AFI and SAFI, are defined to carry the information about controllers with Network Layer Reachability Information (NLRI). Under the AFI and SAFI, a new NLRI, called Controllers NLRI, is defined to contain the information. A

controller in a cluster may advertise the information in a BGP UPDATE message containing a Controllers NLRI of the following format.

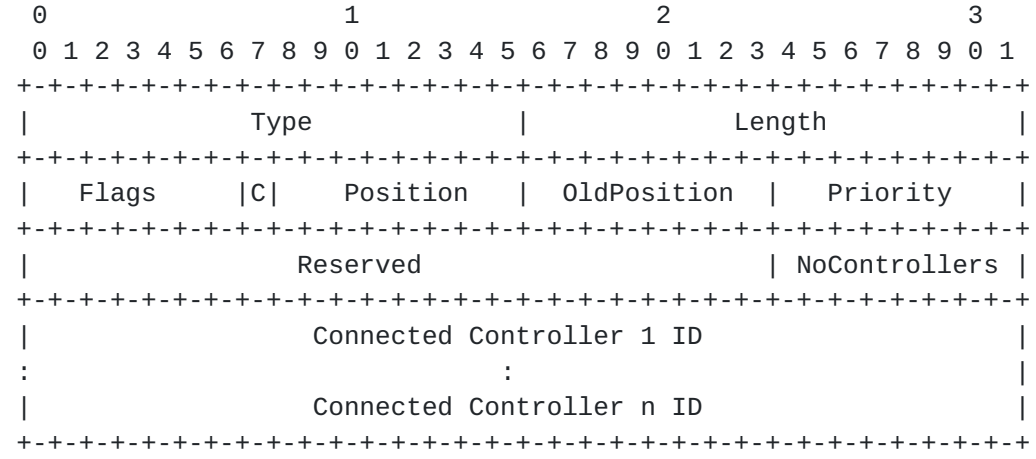

Figure 3: Controllers NLRI

Type (16 bits): TBD2 is to be assigned by IANA.

Length (16 bits): 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.

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 its group. 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 old position controller (e.g., the old primary controller) wins in one policy. In another policy, the group with the highest priority 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 the information about controllers 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 the information about controllers 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 the information about controllers 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

[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>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

9.2. Informative References

[RFC8283] Farrel, A., Ed., Zhao, Q., Ed., Li, Z., and C. Zhou, "An Architecture for Use of PCE and the PCE Communication Protocol (PCEP) in a Network with Central Control", RFC 8283, DOI 10.17487/RFC8283, December 2017, <https://www.rfc-editor.org/info/rfc8283>.

Authors' Addresses

Huaimo Chen
Futurewei
Boston, MA
USA

Email: Huaimo.chen@futurewei.com

Yanhe Fan
Casa Systems
USA

Email: yfan@casa-systems.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