INTERNET-DRAFT
Intended Status: <Standard Track>
Expires: December 30,2017

M. Sun
B.Pithawala
HUAWEI Technologies
F.Gao
Baidu Inc

June 28,2017

<BGP Support for Fast Link Status Notification>

draft-sun-idr-bgp-ls-notification-00

Abstract

This document describes the use of Border Gateway Protocol (BGP) community. This optional transitive community will instruct router to monitor itself ports . With this community, controller only needs to send route update message once and will get the feedback only if link status changes. In particular this community can help controller get the link status changing notification much faster than current method.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html

Copyright and License Notice

Table of Contents

1 Introduction

With the advent of micro services application architecture and the continued advances in massively scaled distributed systems, majority of traffic traversing the data center network is within the data center (east-west). This necessitates the data center network to have deterministic latency (preferably ultra-low), high scalability, high availability and low cost. For those requirements, current large-scale data center network is mostly based on CLOS architecture, [RFC7938] shows a typical 3 layer(5 stages) CLOS architecture(in Figure 1,3 layer means Leaf-Agg-Spine).

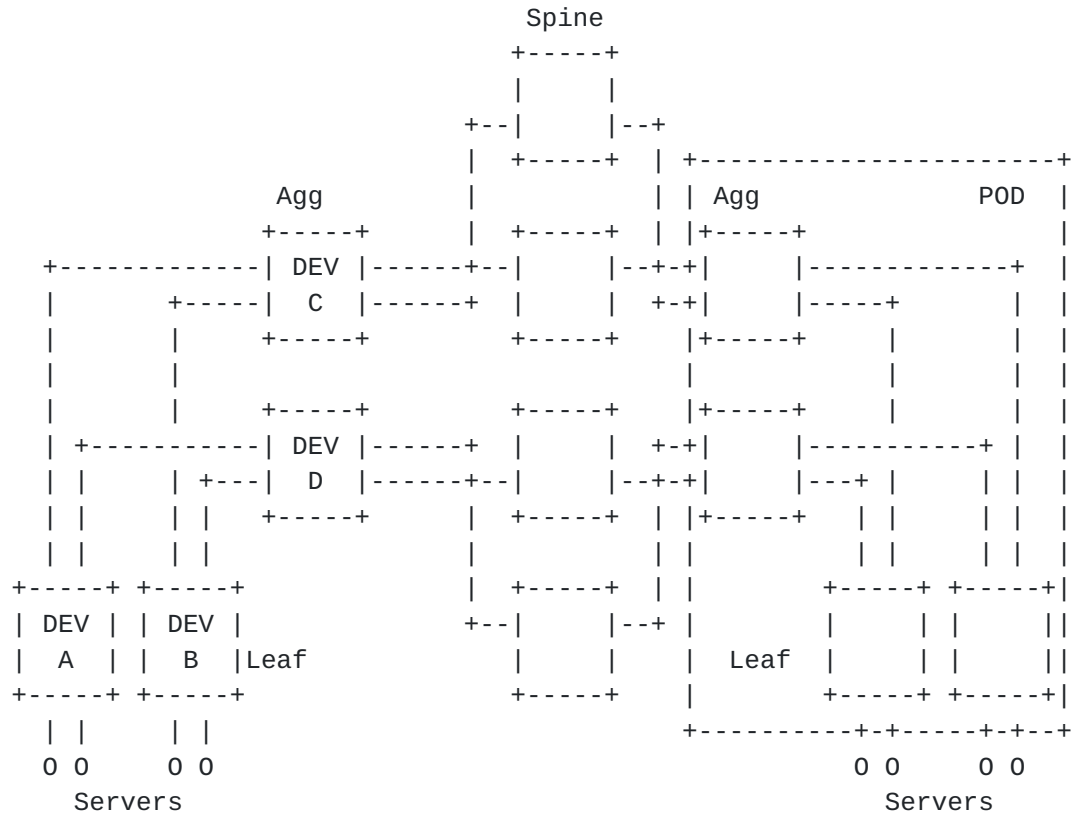

Figure 1 3-Layer Clos Topology

Note: Leaf is switching node that is connected with servers, Agg is exchange node that aggregates Leaf, and Spine is core exchange node.

Nowadays, the scale of this architecture can support 100k servers. The number of links in network is nearly up to 200k links. Managing the large number of switches and links in a data center from a Controller is a difficult scale problem.

1.1 Large-scale DC Routing Solution

[RFC7938] introduces a link detection solution based on BGP.This RFC uses ebgp to connect switches (physical link) and use ibgp to connect switches and controller (logical link). The ebgp connections are made using the local loopback addresses of the Routers/Switches.Since this solution does not have any IGP in the network to convey the local loopback addresses to form the EBGP connection, the solution uses a centralized controller to initiate the messages to convey loopback address of a Router to its neighbor. It uses a combination of ibgp and ebgp connections and messages to achieve the following as Figure 2.

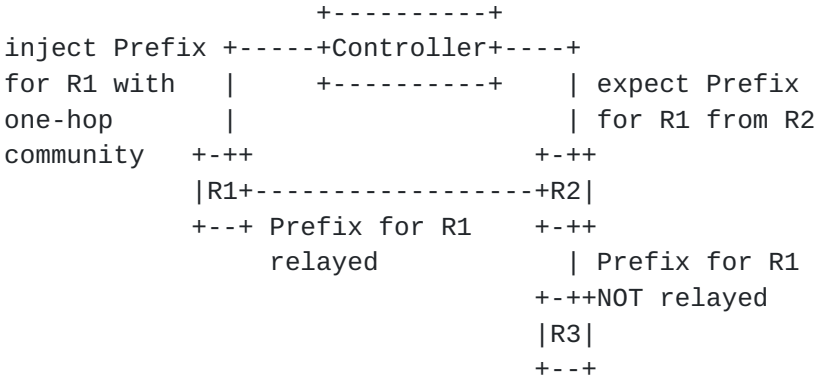

Figure 2 one kind of link detection method

In Figure 2, the controller periodically updates the packets to the source of the link, determines link status (status of link connecting to routers/switches) according to whether controller receives update message from destination link node.The controller sends route message to switch R1 periodically, which only contains one-hop community attribute.R1 publishes this message to its neighbor R2 through ebgp with no_export attribute in it.R2 sends this message to controller through ibgp instead of sending message to R3 because of no_export attribute.If controller receives route message from R2 within specified time, it is assumed that R1->R2 status is normal. Otherwise, R1->R2 status is down.

But when link detection packets sending frequency is high, the controller load is heavy, i.e. controller processing capacity is not enough, and firewall device does not accept this large flow of traffic.On the other hand,when link detection packets sending frequency is low, the convergence speed of network is slow, that will lead to loop or network interruption and other issues. Network reliability is unacceptable.With single controller multi-threaded

exabgp + virtual router vyatta, experimental test data shows that this solution can only support 1k links and 512 servers in non-block network.

1.2 BFD protocol and Hellos Protocol

Existing mainstream distributed link monitoring methods are Protocol Hellos [RFC 2328]and BFD protocol[RFC 5880].

Protocol Hellos: Since a protocol (ebgp) is initiated over the link, the status of the link could be inferred by receiving periodic hellos (or the lack of hellos).Protocol hellos are generally regarded as a slow link detection mechanism. Increasing the frequency of hellos only creates a scale issues at many points in the network without really providing sub-second link detection.

BFD solution configures BFD session at both ends of the link which need to be detected. Each end sends detection BFD messages and link will report failure if the detection message is not received on time.BFD needs plenty of configurations to different devices and different ports. In VRRP track, 100k servers need to configure 200k links and 200k ends. At the same time, 100k servers use BFD need to configure 200k links and 400k ends which may cause some unexpectable errors with high cost.

2. Another Centralized Link Detection Method Based on BGP

2.1 Basic Principle

Considering current large-scale DCN link detection method, there are many problems of periodical detection method. When the frequency of sending and receiving messages is high, the controller load will be too heavy. The controller processing capacity is not enough and firewall devices cannot accept this large flow of traffic. On the other hand, when the frequency is low, the convergence speed of network will decrease. This may cause network interruption and worse network reliability.

Compared with traditional link detection method, this solution propose an efficient optimization method which can monitor links automatically. This method can reduce lots of manual configuration work, avoid various types of errors and high cost. Furthermore, it also eases the collection of link status notifications for the controller.

In Figure 3, if the controller need to detect link status from R1 to R2, the process is as following.

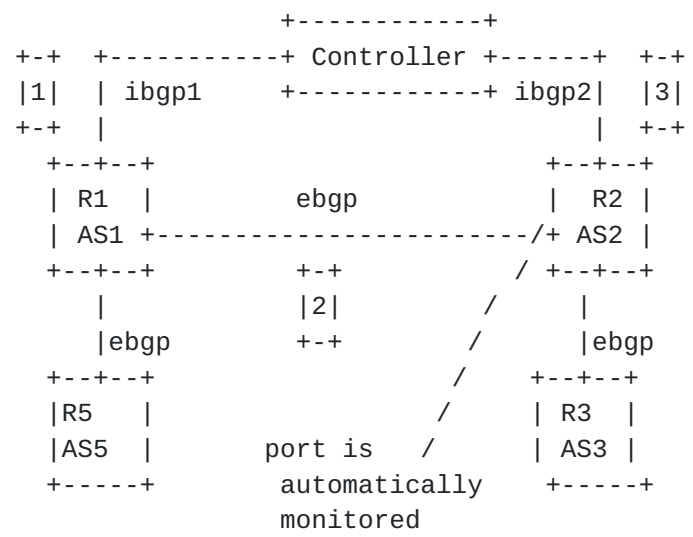

Figure 3 the principle of this solution

Step 1:

a) Controller sends route update message A1 to R1 (nonperiodic, just once) then they can establish a peer. In A1, there's instructions that can enable R1's port (link) status monitoring function.

b) is the same as a>, only the objective is R2.

c) The A1 message only contains one-hop community attribute and its prefix is used to identify device R1.

Step 2:

When R1 receives route update message A1 from controller, it will add a no_export attribute so it can only publish to egbp neighbor R2. R2 will publish this route message to controller through ibgp instead of its ebgp neighbor device R3.

a) R2 finds that message A1 comes from R1 according to the community in A1.

b) Here we need to define a dedicated bit in communities to specify that R2 should start to monitor its link when it receives this indication. Hence, start to monitor all the links from R1 to R2 in this step.

step 3

If it detects ports (links) status has changed in step 2 b), on the

one hand, if the port status switches from normal to fault, R2 will tell controller a withdraw message through ibgp. On the other hand, R2 will tell controller a announce message through ibgp.

step 4

When controller receives route A1 update message from R2:

a) Find corresponding link based on received A1 update message <prefix, srcIP>. Prefix marks network device R1 and srcIP means device R2. The <prefix, srcIP> can tell controller this is the link from R1 to R2.

b) If the message is route announce type, link status is normal, otherwise, the withdraw type means link status is fault.

It is important to notice here that we do not prefer any link detection mechanism and the BGP implementation on a vendor's device is free to activate any link detection mechanism it chooses (some examples are BFD, either auto-sensing feature etc.).

2.2 Advantages and Benefits of this solution

Generally speaking, we need a dedicated bit of communities that can notify R2 to start monitoring the link between R1 and R2. It's quite simple but there are many advantages of this solution.

1. It needs no extra configuration and can monitor corresponding ports (links) automatically. It helps controller know about every link status with existing BGP protocols. It can avoid lots of manual configuration and unnecessary errors and costs caused by manual configuration.

2. It can solve the conflict that network needs fast convergence time but controller capacity constraint. Using this solution, network with single controller can support 100k servers while other method can only support 512 servers.

3. The performance of real-time link failure recovery is better. With experiments, link failure report time reduces from 3s to less than 50ms, link failure recovery time decreases from 1s to less than 50ms.

3 IANA Considerations

The IANA has registered Transitive Extended Community Types in RFC7153. This registry contains values of the high-order octet (the "Type" field) of a Transitive Extended Community.

This method only needs one unassigned type value to notify device monitoring corresponding links(ports).

4 References

4.1 Normative References

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

[RFC7153] E. Rosen, Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, March 2014.

[RFC7938] P. Lapukhov, A. Premji, J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, August 2016.

4.2 Informative References

[RFC3765] Huston, G., "NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control", RFC 3765, April 2004.

[RFC6286] E. Chen, J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, June 2011.

[RFC6608] J. Dong, M. Chen, A. Suryanarayana, "Subcodes for BGP Finite State Machine Error", RFC 6608, May 2012.

[RFC7606] E. Chen, Ed., J. Scudder, Ed., P. Mohapatra, K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, August 2015.

[RFC7705] W. George, S. Amante, "Autonomous System Migration Mechanisms and Their Effects on the BGP AS_PATH Attribute", RFC 7705, November 2015.

[RFC7752] H. Gredler, Ed., J. Medved, S. Previdi, A. Farrel, S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, March 2016.

Authors' Addresses

Marcus Sun
HUAWEI TECHNOLOGIES CO.,LTD
12 E. Mozhou Rd.Nanjing,Jiangsu
China

EMail: marcus.sun@huawei.com

Burjiz Pithawala
HUAWEI TECHNOLOGIES CO.,LTD
2330 Central Expressway, Santa Clara, CA 95050
US

EMail: burjiz.pithawala1@huawei.com

Feng Gao
BAIDU Inc.
10 shangdi shijie Haidian, Beijing

Email:gaofeng04@baidu.com