TEAS Working Group
Internet-Draft
Intended status: Experimental
Expires: April 24, 2019

A. Wang
China Telecom
Q. Zhao
B. Khasanov
H. Chen
Huawei Technologies
R. Mallya
Juniper Networks
October 21, 2018

PCE in Native IP Network
draft-ietf-teas-pce-native-ip-02

Abstract

This document defines the CCDR framework for traffic engineering within native IP network, using Dual/Multi-BGP session strategy and PCE-based central control architecture. The proposed central mode control framework conforms to the concept that defined in [RFC8283]. The scenario and simulation results of CCDR traffic engineering is described in draft [I-D.ietf-teas-native-ip-scenarios].

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 April 24, 2019.

Table of Contents

1. Introduction

Draft [I-D.ietf-teas-native-ip-scenarios] describes the scenario and simulation results for traffic engineering in native IP network. In summary, the requirements for traffic engineering in native IP network are the followings:

- No complex MPLS signaling procedure.
- End to End traffic assurance, determined QoS behavior.
- Identical deployment method for intra- and inter- domain.
- No influence to existing router forward behavior.
- Can utilize the power of centrally control(PCE) and flexibility/ robustness of distributed control protocol.
- Coping with the differentiation requirements for large amount traffic and prefixes.

- Flexible deployment and automation control.

This document defines the framework for traffic engineering within native IP network, using Dual/Multi-BGP session strategy, to meet the above requirements in dynamical and central control mode. The related PCEP protocol extensions to transfer the key parameters between PCE and the underlying network devices(PCC) are provided in draft [I-D.ietf-pce-pcep-extension-native-ip].

2. Conventions used in this document

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] .

3. Dual-BGP Framework for Simple Topology

Dual-BGP framework for simple topology is illustrated in Fig.1, which is comprised by SW1, SW2, R1, R2. There are multiple physical links between R1 and R2. Traffic between IP11 and IP21 is normal traffic, traffic between IP12 and IP22 is priority traffic that should be treated differently.

Only native IGP/BGP protocol is deployed between R1 and R2. The traffic between each address pair may change timely and the corresponding source/destination addresses of the traffic may also change dynamically.

The key ideas of the Dual-BGP framework for this simple topology are the followings:

- Build two BGP sessions between R1 and R2, via the different loopback address lo0, lo1 on these routers.

- Send different prefixes via the two BGP sessions. (For example, IP11/IP21 via the BGP pair 1 and IP12/IP22 via the BGP pair 2).

- Set the explicit peer route on R1 and R2 respectively for BGP next hop of lo0, lo1 to different physical link address between R1 and R2.

The traffic between the IP11 and IP21, and the traffic between IP12 and IP22 will go through different physical links between R1 and R2, each type of traffic occupy different dedicated physical links.

If there is more traffic between IP12 and IP22 that needs to be assured , one can add more physical links between R1 and R2 to reach the loopback address lo1(also the next hop for BGP Peer pair2). In

this cases the prefixes that advertised by two BGP peers need not be changed.

If, for example, there is traffic from another address pair that needs to be assured (for example IP13/IP23), and the total volume of assured traffic does not exceed the capacity of the previous appointed physical links, one need only to advertise the newly added source/destination prefixes via the BGP peer pair2. The traffic between IP13/IP23 will go through the assigned dedicated physical links as the traffic between IP12/IP22.

Such decouple philosophy gives network operator flexible control ability on the network traffic, achieve the determined QoS assurance effect to meet the application's requirement. No complex MPLS signal procedures is introduced, the router need only support native IP protocol.

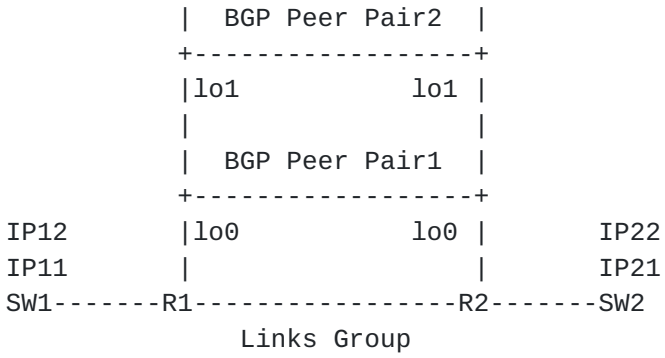

Fig.1 Design Philosophy for Dual-BGP Framework

4. Dual-BGP Framework in Large Scale Topology

When the assured traffic spans across one large scale network, as that illustrated in Fig.2, the dual BGP sessions cannot be established hop by hop especially for the iBGP within one AS.

For such scenario, we should consider to use the Route Reflector (RR) to achieve the similar Dual-BGP effect, select one router which performs the role of RR (for example R3 in Fig.2), every other edge router will establish two BGP peer sessions with the RR, using their different loopback addresses respectively. The other two steps for traffic differentiation are same as that described in the Dual-BGP simple topology usage case.

For the example shown in Fig.2, if we select the R1-R2-R4-R7 as the dedicated path, then we should set the explicit peer routes on these routers respectively, pointing to the BGP next hop (loopback

addresses of R1 and R7, which are used to send the prefix of the assured traffic) to the actual address of the physical link.

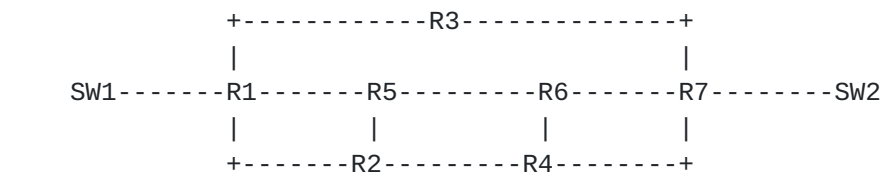

Fig.2 Dual-BGP Framework for Large Scale Network

5. Multi-BGP Strategy for Extended Traffic Differentiation

In general situation, several additional traffic differentiation criteria exist, including:

- Traffic that requires low latency links and is not sensitive to packet loss.
- Traffic that requires low packet loss but can endure higher latency.
- Traffic that requires lowest jitter path.

These different traffic requirements can be summarized in the following table:

Flow No.	Latency	Packet Loss	Jitter
1	Low	Normal	Don't care
2	Normal	Low	Dont't care
3	Normal	Normal	Low

Table 1. Traffic Requirement Criteria

For Flow No.1, we can select the shortest distance path to carry the traffic; for Flow No.2, we can select the idle links to form its end to end path; for Flow No.3, we can let all assured traffic pass one single path, no ECMP distribution on the parallel links is required.

It is almost impossible to provide an end-to-end (E2E) path with latency, jitter, packet loss constraints to meet the above requirements in large scale IP-based network via the distributed routing protocol, but these requirements can be solved using the CCDR framework since the PCE has the overall network view, can collect

real network topology and network performance information about the underlying network, select the appropriate path to meet various network performance requirements of different traffic.

6. CCDR Procedures for Multi-BGP Strategy

The procedures to implement the Multi-BGP strategy are the followings:

- PCE gets topology and link utilization information from the underlying network, calculates the appropriate link path upon application's requirements..

- PCE sends the key parameters to edge/RR routers(R1, R7 and R3 in Fig.3) to build multi-BGP peer relations and advertises different prefixes via them.

- PCE sends the route information to the routers (R1,R2,R4,R7 in Fig.3) on forwarding path via PCEP, to build the path to the BGP next-hop of the advertised prefixes.

- If the assured traffic prefixes were changed but the total volume of assured traffic does not exceed the physical capacity of the previous end-to-end path, then PCE needs only change the related information on edge routers (R1,R7 in Fig.3).

- If the volume of assured traffic exceeds the capacity of previous calculated path, PCE must recalculate the appropriate path to accommodate the exceeding traffic via some new end-to-end physical links. After that PCE needs to update on-path routers to build such path hop by hop.

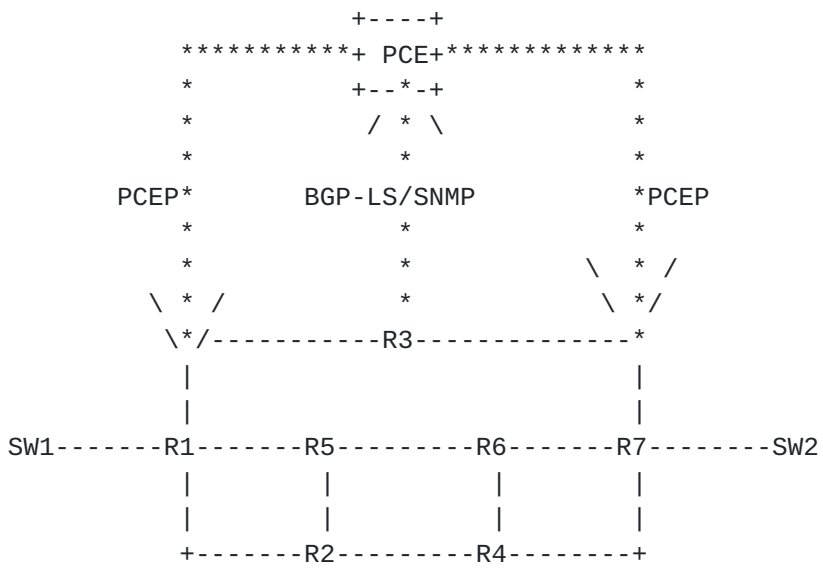

Fig.3 PCE based framework for Multi-BGP deployment

7. PCEP Extension for Key Parameters Delivery

The PCEP protocol needs to be extended to transfer the following key parameters:

o BGP peer address and advertised prefixes.

o Explicit route information to BGP next hop of advertised prefixes.

Once the router receives such information, it should establish the BGP session with the peer appointed in the PCEP message, advertise the prefixes that contained in the corresponding PCEP message, and build the end to end dedicated path hop by hop. Details of communications between PCEP and BGP subsystems in router's control plane are out of scope of this draft and will be described in separate draft [I-D.ietf-pce-pcep-extension-native-ip] .

The reason that we selected PCEP as the southbound protocol instead of OpenFlow, is that PCEP is suitable for the changes in control plane of the network devices, there OpenFlow dramatically changes the forwarding plane. We also think that the level of centralization that requires by OpenFlow is hardly achievable in many today's SP networks so hybrid BGP+PCEP approach looks much more interesting.

8. Deployment Consideration

8.1. Scalability

In CCDR framework, PCE needs only to influence the edge routers for the prefixes differentiation via the multi-BGP deployment. The route information for these prefixes within the on-path routers were distributed via the BGP protocol. Unlike the solution from BGP Flowspec, the on-path router need only keep the specific policy routes to the BGP next-hop of the differentiate prefixes, not the specific routes to the prefixes themselves. This can lessen the burden from the table size of policy based routes for the on-path routers, and has more scalabilities when comparing with the solution from BGP flowspec or Openflow.

8.2. High Availability

CCDR framework is based on the distributed IP protocol. If the PCE failed, the forwarding plane will not be impacted, as the BGP session between all devices will not flap, and the forwarding table will remain the same. If one node on the optimal path is failed, the assurance traffic will fall over to the best-effort forwarding path. One can even design several assurance paths to load balance/hot standby the assurance traffic to meet the path failure situation, as done in MPLS FRR.

For high availability of PCE/SDN-controller, operator should rely on existing HA solutions for SDN controller, such as clustering technology and deployment.

8.3. Incremental deployment

Not every router within the network support will support the PCEP extension that defined in [I-D.ietf-pce-pcep-extension-native-ip] simultaneously. For such situations, router on the edge of domain can be upgraded first, and then the traffic can be assured between different domains. Within each domain, the traffic will be forwarded along the best-effort path. Service provider can selectively upgrade the routers on each domain in sequence.

9. Security Considerations

Solution described in this draft puts more requirements on the function of PCE and its communication with the underlay devices. The PCE should have the capability to calculate the loop-free e2e path upon the status of network condition and the service requirements in real time. The PCE need also to consider the router order during deployment to eliminate the possible transient traffic loop.

This solution does not require the change of forward behavior on the underlay devices, then there will no additional security impact for the devices.

When deploy the solution on network, service provider should also consider more on the protection of SDN controller and their communication with the underlay devices, which is described in document [RFC5440] and [RFC8253]

10. IANA Considerations

This document does not require any IANA actions.

11. Contributors

Penghui Mi and Shaofu Peng contribute the contents of this draft.

12. Acknowledgement

The author would like to thank Deborah Brungard, Adrian Farrel, Huaimo Chen, Vishnu Beeram, Lou Berger, Dhruv Dhody and Jessica Chen for their supports and comments on this draft.

13. Normative References

[I-D.ietf-pce-pcep-extension-native-ip]
Wang, A., Khasanov, B., Cheruathur, S., and C. Zhu, "PCEP Extension for Native IP Network", draft-ietf-pce-pcep-extension-native-ip-01 (work in progress), June 2018.

[I-D.ietf-teas-native-ip-scenarios]
Wang, A., Huang, X., Qou, C., Li, Z., Huang, L., and P. Mi, "CCDR Scenario, Simulation and Suggestion", draft-ietf-teas-native-ip-scenarios-01 (work in progress), June 2018.

[I-D.ietf-teas-pcecc-use-cases]
Zhao, Q., Li, Z., Khasanov, B., Dhody, D., Ke, Z., Fang, L., Zhou, C., Communications, T., Rachitskiy, A., and A. Gulida, "The Use Cases for Path Computation Element (PCE) as a Central Controller (PCECC).", draft-ietf-teas-pcecc-use-cases-02 (work in progress), October 2018.

[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>.

[RFC5440] Vasseur, JP., Ed. and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, DOI 10.17487/RFC5440, March 2009, <https://www.rfc-editor.org/info/rfc5440>.

[RFC8253] Lopez, D., Gonzalez de Dios, O., Wu, Q., and D. Dhody, "PCEPS: Usage of TLS to Provide a Secure Transport for the Path Computation Element Communication Protocol (PCEP)", RFC 8253, DOI 10.17487/RFC8253, October 2017, <https://www.rfc-editor.org/info/rfc8253>.

[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

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

Email: wangaj.bri@chinatelecom.cn

Quintin Zhao
Huawei Technologies
125 Nagog Technology Park
Acton, MA 01719
USA

Email: quintin.zhao@huawei.com

Boris Khasanov
Huawei Technologies
Moskovskiy Prospekt 97A
St.Petersburg 196084
Russia

Email: khasanov.boris@huawei.com

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: huaimo.chen@huawei.com

Raghavendra Mallya
Juniper Networks
1133 Innovation Way
Sunnyvale, California 94089
USA

Email: rmallya@juniper.net