TEAS Working Group
Internet Draft

A.Wang
China Telecom
Quintin Zhao
Boris Khasanov
Huawei Technologies
Kevin Mi
Tencent Company
Raghavendra Mallya
Juniper Networks
Shaofu Peng
ZTE Corporation

Intended status: Standard Track
March 6, 2017
Expires: September 5, 2017

PCE in Native IP Network
draft-wang-teas-pce-native-ip-02.txt

Status of this Memo

Copyright Notice

Abstract

This document defines the scenario and solution for traffic engineering within Native IP network, using Dual/Multi-BGP session strategy and PCE-based central control architecture. The proposed central mode control solution conforms to the concept that defined in draft [I-D.draft-ietf-teas-pce-control-function]. And together with draft [I-D.draft-ietf-teas-pcecc-use-cases], the solution portfolio for traffic engineering in MPLS and Native IP network is almost completed.

Table of Contents

1. Introduction

Currently, PCE based traffic assurance requires the underlying network devices support MPLS and the network must deploy multiple LSPs to assure the end-to-end traffic performance. LDP/RSVP-TE or Segment Routing should be enabled within the network to establish various MPLS paths. Such solution will certainly work but they does not cover the needs in legacy Native IP network, which demands less signaling protocol and less complex traffic steering policy.

Within Native IP network, the solution for traffic engineering is generally hop-by-hop differentiate treatment. To achieve the end2end QoS performance assurance, one can only deploy some dedicated links statically, but such solution is not feasible in the service provider network, because the complexity of underlying network and the variation of application traffic from time to time.

In summary, the requirements for traffic engineering in Native IP network are the following:
1) No complex MPLS signaling procedure.
2) End to End traffic assurance, determined QoS behavior.
3) Flexible deployment and automation control.

This document defines the solution for traffic engineering within Native IP network, using Dual/Multi-BGP session strategy and PCE-based central control architecture, to meet the above requirements in dynamical and central control mode. Future PCEP protocol extensions to transfer the key parameters between PCE and the underlying network devices(PCC) are provided in draft [draft-wang-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 solution for simple topology.

This section introduces the dual-BGP solution for simple topology that illustrated in Fig.1, which is comprised by SW1, SW2, R1, R2. There are multiple physical links between R1 and R2. Let's assume 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 idea of the Dual-BGP solution for this simple topology is the following:

1) Build two BGP sessions between R1 and R2, via the different loopback address lo0, lo1 on these routers.
2) 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).
3) 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.

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

If there is more traffic between IP12 and IP13 that needs to be assured , one can add more physical links on 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 peer need not be changed.

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

Such decouple philosophy gives the network operator more flexible control ability on the network traffic, get 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.

```
                          |  BGP Peer Pair2  |
                          +------------------+
                          |lo1           lo1 |
                          |                  |
                          |  BGP Peer Pair1  |
                          +------------------+
               IP12       |lo0           lo0 |       IP22
               IP11       |                  |       IP21
               SW1-------R1-----------------R2-------SW2
                              Links Group
```

Fig.1 Design Philosophy for Dual-BGP Solution

4. Dual-BGP 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 neighbor by neighbor 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, that is to say, select one router which performs the role of RR (for example R3 in Fig.2 - Dual-BGP Solution using Route Reflector for large scale network), every other router will establish two BGP sessions with the RR, using their different loopback addresses respectively. The other two steps for traffic differentiation are same as one 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

```
                      +------------R3--------------+
                      |                            |
          SW1-------R1-------R5---------R6-------R7--------SW2
                      |        |          |        |
                      +-------R2---------R4--------+
```

Fig.2 Dual-BGP solution for large scale network

5. Multi-BGP for Extended Traffic Differentiation

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

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

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 the traffic pass one single path, no ECMP distribution on the parallel links is required.

It is difficult and almost impossible to provide an end-to-end (E2E) path with latency, latency variation, packet loss, and bandwidth utilization constraints to meet the above requirements in large scale IP-based network via the traditional distributed routing protocol, but these requirements can be solved using the PCE-based architecture 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 the various network performance requirements of different traffic type.

6. PCE based solution for Multi-BGP strategy deployment.

With the advent of SDN concepts towards pure IP networks, it is possible to deploy the PCE related technology into the underlying native IP network, to accomplish the central and dynamic control of network traffic according to the application's various requirements.

The procedure to implement the dynamic deployment of Multi-BGP strategy is the following:

1) PCE gets topology and link utilization information from the underlying network, calculate the appropriate link path upon application's requirements.
2) PCE sends the key parameters to edge/RR routers(R1, R7 and R3 in Fig.3) to build multi-BGP peer relations and advertise different prefixes via them.
3) 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.
4) 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).
5) If volume of the 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 link. After that PCE needs to update on-path routers to build such path hop by hop.

```
                               +----+
                     ***********+PCE +*************
                     *         +--*-+            *
                     *           / * \            *
                     *             *              *
                 PCEP*             *BGP-LS/SNMP   *PCEP
                     *             *              *
                     *             *           \  * /
                   \ * /           *            \ */
                    \*/-----------R3--------------*
                     |                            |
                     |                            |
          SW1-------R1-------R5---------R6-------R7--------SW2
                     |        |          |        |
                     |        |          |        |
                     +-------R2---------R4--------+
```

Fig.3 PCE based solution for Multi-BGP deployment

7. PCEP extension for key parameters delivery.

We need to extend the PCEP protocol to transfer the following key parameters:
1) BGP peer address and advertised prefixes.
2) 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.[draft-wang-pce-extension for native IP]

The reason why we selected PCEP as the southbound protocol instead of OpenFlow, is that PCEP is very 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

This solution requires the parallel work of 2 subsystems in router's control plane: PCE (PCEP) and BGP as well as coordination between them, so it might require additional planning work before deployment.

8.1 Scalability

In current solution, only the edge router of the end2end path needs to keep the detail prefixes of the assured traffic, other on-path routers need only keep explicit peer routes to the edge routers.

The key scalability factor is the number of BGP sessions as on ingress/egress routers as on RRs. Similarly with L3VPN solution, it has very high scalability to deploy in real network.

8.2 High Availability

Current solution is based on the traditional distributed IP protocol, then if the central control PCE 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. From PCE/SDN-controller HA side we will rely on existing HA solutions of SDN controllers such as clustering.

8.3 Incremental deployment

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

8.4 Deployment within Pure IGP network

For some small underlying networks that the routers support only the pure IGP protocol, we can use similar procedures that described within this draft to differentiate the forwarding paths for different applications:

1) Define different loopback addresses on the IGP edge router(ASBR).

2) Redistribute external prefixes into IGP at ASBR, use route tag to label these prefixes at the ASBR.

3) Use route policy to set the explicit peer routes for the tagged prefixes on every on-path routers to the different loopback addresses on ASBR.

The detail of deployment scenario and the corresponding PCEP extension will be exploited further later.

9. Security Considerations

TBD

10. IANA Considerations

TBD

11. Conclusions

TBD

12. References

12.1. Normative References

[RFC4655] Farrel, A., Vasseur, J.-P., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006,<http://www.rfc-editor.org/info/rfc4655>.

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

12.2. Informative References

[I-D.draft-ietf-teas-pce-control-function]

A.Farrel, Q.Zhao et al. "An Architecture for use of PCE and PCEP in a Network with Central Control"

https://datatracker.ietf.org/doc/draft-ietf-teas-pce-central-control/ September, 2016

[I-D. draft-ietf-teas-pcecc-use-cases]

Quintin Zhao, Robin Li, Boris Khasanov et al. "The Use Cases for Using PCE as the Central Controller(PCECC) of LSPs

https://tools.ietf.org/html/draft-ietf-teas-pcecc-use-cases-00

March,2017

[draft-wang-pcep-extension for native IP]

Aijun Wang, Boris Khasanov et al. "PCEP Extension for Native IP Network" https://datatracker.ietf.org/doc/draft-wang-pce-extension-native-ip/

13. Acknowledgments

The authors would like to thank George Swallow, Xia Chen, Jeff Tantsura, Daniele Ceccarelli and Dhruv Dhody for their valuable comments and suggestions.

The authors would also like to thank Lou Berger, Adrian Farrel, King Daniel for their suggestions to put forward this draft.

Authors' Addresses

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

Email: wangaj.bri@chinatelecom.cn

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

EMail: quintin.zhao@huawei.com

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

EMail: khasanov.boris@huawei.com

Kevin Mi
Tencent Company
Tencent Building, Kejizhongyi Avenue,
Hi-techPark,Nanshan District,Shenzhen

Email kevinmi@tencent.com

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

Email: rmallya@juniper.net

Shaofu Peng
ZTE Corporation
No.68 Zijinghua Road,Yuhuatai District
Nanjing 210012
China

Email: peng.shaofu@zte.com.cn