

TEAS Working Group
Internet Draft

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

Intended status: Standard Track
Expires: April 23, 2017

October 24 2016

PCE in Native IP Network
draft-wang-teas-pce-native-ip-01.txt

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#). This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

This Internet-Draft is submitted in full conformance with the provisions of [BCP 78](#) and [BCP 79](#). This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

it for publication as an RFC or to translate it into languages other than English.

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/ietf/1id-abstracts.txt>

<A.Wang>

Expires April 23, 2017

[Page 1]

This Internet-Draft will expire on April 24, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

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-zhao-teas-pcecc-use-cases](#)], the solution portfolio for traffic engineering in MPLS and Native IP network is almost completed.

Table of Contents

| | |
|--|--------------------|
| 1. Introduction | 3 |
| 2. Conventions used in this document | 3 |
| 3. Dual-BGP solution for simple topology.. | 3 |
| 4. Dual-BGP in large Scale Topology | 5 |
| 5. Multi-BGP for Extended Traffic Differentiation..... | 6 |
| 6. SDN/PCE based solution for Multi-BGP strategy deployment..... | 7 |
| 7. PCEP extension for key parameter transformation. | 8 |
| 8. Deployment Consideration..... | 9 |
| 9. Security Considerations | 10 |
| 10. IANA Considerations..... | 10 |
| 11. Conclusions | 10 |
| 12. References | 10 |
| 12.1. Normative References | 10 |
| 12.2. Informative References..... | 11 |
| 13. Acknowledgments | 11 |

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 always hop-by-hop differentiate service. To achieve the end2end QoS performance assurance, one can only deploy dedicated links statically to meet such requirements. Such solution is not feasible in the service provider network, because the volume and path of application traffic will be vary from time to time and the network is very complex.

In summary, there are scenarios that can't be deployed within current PCE-based MPLS network, because of the following requirements:

- 1) Native IP environment, No complex MPLS signaling procedure.
- 2) End to End traffic assurance, Determined QoS behavior.
- 3) Flexible deployment with central control.

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, to meet the above requirements in dynamical and central control mode. Future PCEP protocol extension to transfer the key parameters between PCE and the underlying network devices(PCC)is provided in draft [[draft-wang-pcep-extension](#) for 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 first 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. Traffic between IP11 and IP21 is normal traffic, traffic between IP12 and IP22 is priority traffic that should be treated differently.

There is only Native IP protocol being deployed between R1 and R2. The traffic between each address pair will be changed timely and the corresponding source/destination addresses of the traffic may also be changed 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 static 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 case 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), 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.

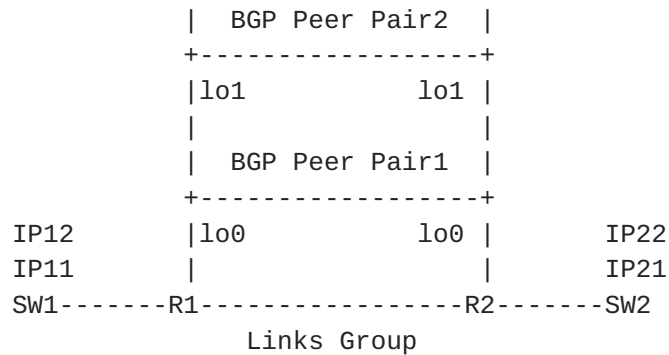


Fig.1 Design Philosophy for Dual-BGP Solution

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 static 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

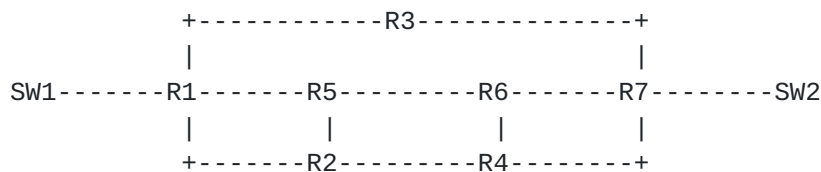


Fig.2 Dual-BGP solution for large scale network

5. Multi-BGP for Extended Traffic Differentiation

The following requirement was discussed in the document so far, the ability to classify traffic into two classes: Assured traffic (high priority) or best effort (normal) traffic. Dual-BGP solution (simple topology or large scale topology) can meet above requirements. In general, 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 | Don'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 SDN/PCE-based architecture since the SDN Controller/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

6. SDN/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 underlying topology information via the BGP-LS protocol from one of BGP routers in the network, such as the route reflector R3 in Fig.3
- 2) PCE also collects the link utilization information via SNMP or NetFlow protocols.
- 3) PCE will calculate the appropriate link path depending on application's requirement (for example bi-direction traffic assurance between SW1/SW2), that path can be assigned to such traffic flow in dedicated mode, other regular traffic will not pass through such physical links.
- 4) After that PCE will send via PCEP extensions the key parameters to R1 and R7 respectively, to let R1 and R7 build another i/eBGP neighbor relations with R3 and advertise prefixes that are owned by SW1/SW2.
- 5) If the calculated dedicated path goes via some physical links that belong to R1-R2-R4-R7, then PCE also build the PCEP connections with these on-path routers and send similar key parameters to them via PCEP to build the path to the BGP next-hop via address of physical links between R1/R2, R2/R4, R4/R7.
- 6) If the assured traffic prefixes were changed but the total volume of assured traffic was not exceed the physical capacity of the previous end-to-end path, then PCE needs only change the related information on R1 and R7.
- 7) 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.

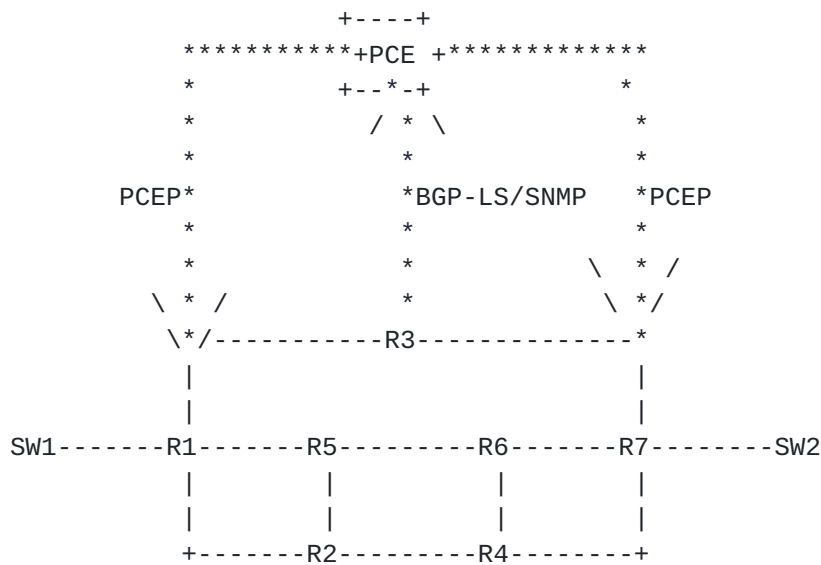


Fig.3 PCE based solution for Multi-BGP deployment

7. PCEP extension for key parameters delivery.

In order to inform underlying routers about Multi-BGP deployment scenario and keep the overall implementation as simple as possible, we want to extend the PCEP protocol to transfer the following key parameters:

- 1) BGP peer address and assured prefixes that will be advertised via this BGP session
- 2) Static route information to BGP next hop of these 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

Internet-Draft PCE in Native IP Network October 24, 2016

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 head/end or edge router of the end2end path needs to keep the detail prefixes of the assured traffic, other on-path routers need only keep very few static routes to the edge routers.

The key scalability factor is the number of BGP sessions as on ingress/egress routers as on RRs. Possible scalability restrictions of this topic require additional research and will be added in later versions of this draft.

Overall, 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](#) for native IP]simulatineously. For such situations, firstly router on the edge of sub domain can be upgraded, then the traffic can be assured between different sub domains. Within each sub domain, the traffic

8.4 Deployment within Pure underlying OSPF network

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

- 1) Put different loopback addresses on the edge router within different OSPF area.
- 2) Redistribute the external prefixes into different OSPF areas, which are identified by different loop addresses.
- 3) OSPF will use these loop addresses as the "forward address" the external prefix.
- 4) Modify the routes to these "forward addresses" on each on-path OSPF routers according to the calculation path of centrally controlled PCE.

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

Internet-Draft PCE in Native IP Network October 24, 2016
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/](https://datatracker.ietf.org/doc/draft-ietf-teas-pce-central-control/) September, 2016

[I-D. [draft-zhao-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-zhao-teas-pcecc-use-cases-01>
July, 2016

[[draft-wang-pcep-extension](#) for native IP]

Aijun Wang, Boris Khasanov et al. "PCEP Extension for Native IP
Network"

13. Acknowledgments

TBD

Authors' Addresses

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

Email: wangaj@ctbri.com.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