TEAS Working Group
Internet-Draft
Intended status: Experimental
Expires: October 18, 2019

A. Wang
China Telecom
Q. Zhao
B. Khasanov
H. Chen
Huawei Technologies
R. Mallya
Juniper Networks
April 16, 2019

# PCE in Native IP Network
# draft-ietf-teas-pce-native-ip-03

## Abstract

This document defines the framework for traffic engineering within native IP network, using Dual/Multi-BGP sessions 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 traffic engineering in Native IP network 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 October 18, 2019.

Table of Contents



## 1. Introduction

Draft [I-D.ietf-teas-native-ip-scenarios] describes the scenarios, simulation results and suggestions for traffic engineering in native IP network. To meet the requirements of various scenarios, the solution for traffic engineering in native IP network should have the followings criteria:

- No complex MPLS signaling procedures.
- End to End traffic assurance, determined QoS behavior.
- Identical deployment method for intra-domain and inter-domain.
- No influence to forwarding behavior of the router.
- Can exploit 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 centrally control mode(Centrally Control Dynamic Routing, abbreviated as CCDR ). 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. CCDR Framework in Simple Topology

Fig.1 illustrates the CCDR framework for traffic engineering in simple topology. The topology is comprised by four devices which are SW1, SW2, R1, R2. There are multiple physical links between R1 and R2. Traffic between IP11(on SW1) and IP21(on SW2) is normal traffic, traffic between IP12(on SW1) and IP22(on SW2) 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 in real time and the corresponding source/destination addresses of the traffic may also change dynamically.

The key ideas of the CCDR 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 established 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.

After the above actions, 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 set of traffic occupies 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 the 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 needs only support native IP protocol.

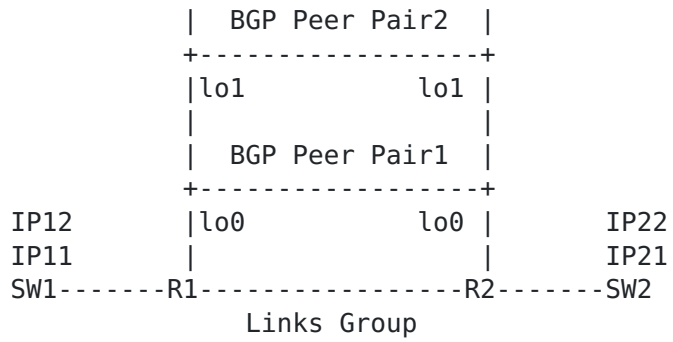


Fig.1 CCDR framework in simple topology

## 4. CCDR Framework in Large Scale Topology

When the assured traffic spans across the 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 effect. Every edge router will establish two BGP peer sessions with the RR via different loopback addresses respectively. The other steps for traffic differentiation are same as that described in the CCDR framework for simple topology.

As shown in Fig.2, if we select R3 as the RR, every edge router(R1 and R7 in this example) will build two BGP session with the RR. If the PCE calculates select the dedicated path as R1-R2-R4-R7, then the

operator 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 selected forwarding address.

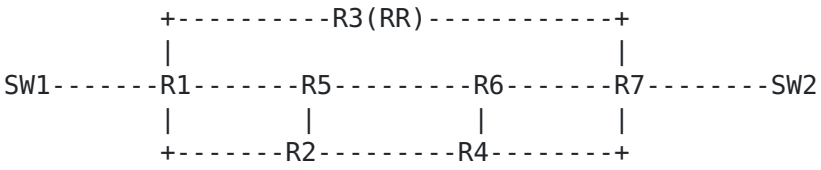


Fig.2 CCDR framework in large scale network

## 5. CCDR Multi-BGP Strategy

In general situation, different applications may require different QoS criteria, which may include:

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

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 path that is comprised by underloading links from end to end; for Flow No.3, we can let all assured traffic pass the determined single path, no ECMP distribution on the parallel links is desired.

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 with the assistance of PCE controller, because 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 traffics.

## 6. CCDR Framework for Multi-BGP Strategy

The framework to implement the CCDR Multi-BGP strategy are the followings:

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

- PCE sends the key parameters to edge/RR routers(R1, R7 and R3 in Fig.3) to establish multi-BGP peer sessions 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, PCE needs only change the prefixed advertised via the edge routers (R1,R7 in Fig.3).

- If the volume of assured traffic exceeds the capacity of previous calculated path, PCE can recalculate the appropriate paths to accommodate the exceeding traffic. After that, PCE needs to update on-path routers to build the forwarding path hop by hop.

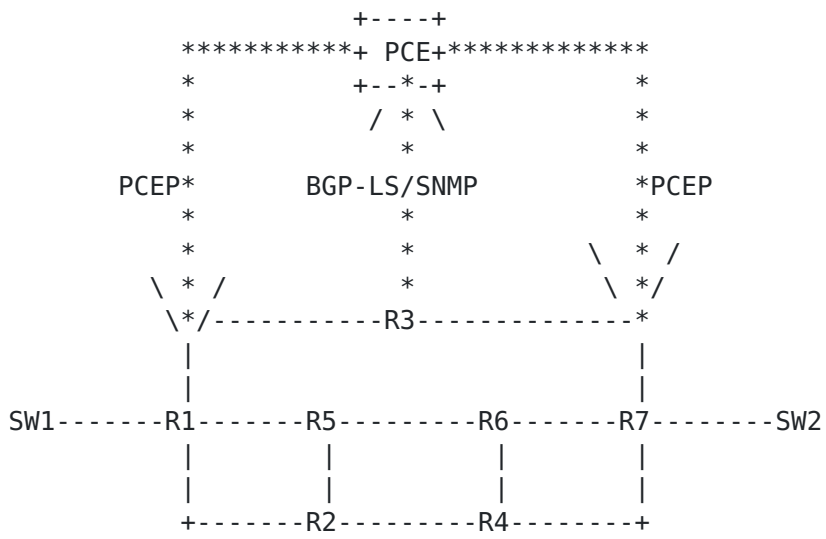


Fig.3 CCDR framework for Multi-BGP deployment

## 7. PCEP Extension for Key Parameters Delivery

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

- Peer addresses pair that is used to build the BGP session
- Advertised prefixes and their associated BGP session.
- 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, while 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 influence the edge routers for the prefixes advertisement 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 expandability when comparing with the solution from BGP flowspec or Openflow.

### 8.2. High Availability

The 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 unchaned.

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 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

The PCE should have the capability to calculate the loop-free end to end path upon the status of network condition and the service requirements in real time.

The PCE need consider the explicit route deployment order (for example, from tail router to head router) to eliminate the possible transient traffic loop.

CCDR framework described in this draft puts more requirements on the function of PCE and its communication with the underlay devices. Service provider should consider more on the protection of SDN controller and their communication with the underlay devices, which is described in document [RFC5440] and

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

## 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., Zhu, C., and S. Fang, "PCEP Extension for Native IP Network", draft-ietf-pce-pcep-extension-native-ip-03 (work in progress), March 2019.

[I-D.ietf-teas-native-ip-scenarios]
Wang, A., Huang, X., Qou, C., Li, Z., and P. Mi, "Scenario, Simulation and Suggestion of PCE in Native IP Network", draft-ietf-teas-native-ip-scenarios-02 (work in progress), October 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-03 (work in progress), March 2019.

[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