TEAS Working Group
Internet-Draft
Intended status: Experimental
Expires: May 1, 2021

A. Wang
China Telecom
B. Khasanov
Huawei Technologies
Q. Zhao
Etheric Networks
H. Chen
Futurewei

October 28, 2020

PCE in Native IP Network
draft-ietf-teas-pce-native-ip-12

Abstract

This document defines the architecture for traffic engineering within native IP network, using multiple BGP sessions strategy and PCE -based central control mechanism. It uses the Central Control Dynamic Routing (CCDR) procedures described in this document, and the Path Computation Element Communication Protocol (PCEP) extension specified in draft ietf-pce-pcep-extension-native-ip.

Table of Contents

1. Introduction

[RFC8735] describes the scenarios and simulation results for traffic engineering in the native IP network to provide End-to-End (E2E) performance assurance and QoS using PCE based centralized control, referred to as Centralized Control Dynamic Routing (CCDR). Based on the various scenarios and analysis as per [RFC8735], the solution for traffic engineering in native IP network should meet the following criteria:

- Same solution for native IPv4 and IPv6 traffic.
- Support for intra-domain and inter-domain scenarios.
- Achieve End to End traffic assurance, with determined QoS behavior.
- No changes in routers forwarding behavior.
- Capability to use the power of centrally control and the flexibility/robustness of distributed network control plane.

- Different network requirements such as large traffic amount and prefix scale.

- Adjusting the optimal path dynamically upon the change of network status. No need for physical links resources reservation in advance.

Stateful PCE [RFC8231] specifies a set of extensions to PCEP to enable stateful control of paths such as MPLS-TE Label Switched Paths(LSP)s between and across PCEP sessions in compliance with [RFC4657]. It includes mechanisms to achieve state synchronization between Path Computation Clients(PCCs) and PCEs, delegation of control of LSPs to PCEs, and PCE control of timing and sequence of path computations within and across PCEP sessions. Furthermore, [RFC8281] specifies a mechanism to dynamically instantiate LSPs on a PCC based on the requests from a stateful PCE or a controller using stateful PCE. [RFC8283] introduces the architecture for PCE as a central controller as an extension of the architecture described in [RFC4655] and assumes the continued use of PCEP as the protocol used between PCE and PCC.[RFC8283] further examines the motivations and applicability for PCEP as a Southbound Interface (SBI), and introduces the implications for the protocol.

This document defines the architecture for traffic engineering within native IP network, using multiple BGP session strategy, to meet the above criteria in dynamical and centrally control mode. The architecture is referred as CCDR architecture. It depends on the central control (PCE) element to compute the optimal path for selected traffic, and utilizes the dynamic routing behavior of traditional IGP/BGP protocols to forward such traffic.

The control messages between PCE and underlying network node are transmitted via Path Computation Element Communications Protocol (PCEP) protocol. The related PCEP extensions are provided in draft [I-D.ietf-pce-pcep-extension-native-ip].

2. Terminology

This document uses the following terms defined in [RFC5440]:

- PCE

- PCEP

- PCC

Other terms are defined in this document:

- CCDR: Central Control Dynamic Routing
- E2E: End to End
- ECMP: Equal-Cost Multipath
- RR: Route Reflector
- SDN: Software Defined Network

3. CCDR Architecture in Simple Topology

Figure 1 illustrates the CCDR architecture 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 prefix PF11(on SW1) and prefix PF21(on SW2) is normal traffic, traffic between prefix PF12(on SW1) and prefix PF22(on SW2) is priority traffic that should be treated accordingly.

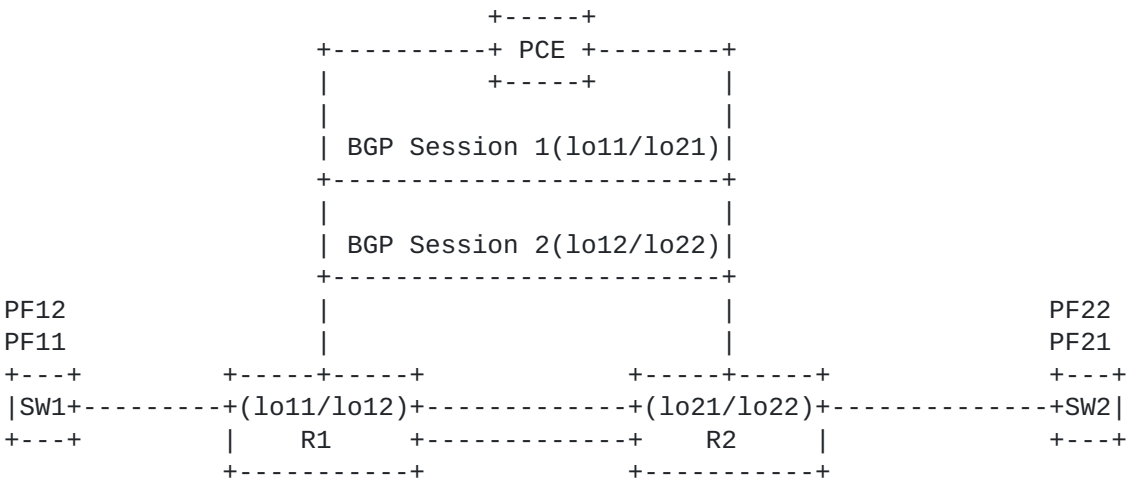

Figure 1: CCDR architecture in simple topology

In Intra-AS scenario, IGP and BGP are deployed between R1 and R2. In inter-AS scenario, only native BGP protocol is deployed. 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 architecture for this simple topology are the followings:

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

- Set the explicit peer route on R1 and R2 respectively for BGP next hop to different physical link addresses between R1 and R2. Such explicit peer route can be set in the format of static route to BGP peer address, which is different from the route learned from the IGP protocol.

- Send different prefixes via the established BGP sessions. For example, PF11/PF21 via the BGP session 1 and PF12/PF22 via the BGP session 2.

After the above actions, the bi-direction traffic between the PF11 and PF21, and the bi-direction traffic between PF12 and PF22 will go through different physical links between R1 and R2.

If there is more traffic between PF12 and PF22 that needs to be assured , one can add more physical links between R1 and R2 to reach the the next hop for BGP session 2. In this cases the prefixes that advertised by the BGP peers need not be changed.

If, for example, there is bi-directional traffic from another address pair that needs to be assured (for example prefix PF13/PF23), and the total volume of assured traffic does not exceed the capacity of the previously provisioned physical links, one need only to advertise the newly added source/destination prefixes via the BGP session 2. The bi-direction traffic between PF13/PF23 will go through the assigned dedicated physical links as the traffic between PF12/PF22.

Such decouple philosophy achieves the flexible control capability for the network traffic, to achieve the determined QoS assurance effect to meet the application's requirement. The router needs only support native IP and multiple BGP sessions setup via different loopback addresses.

4. CCDR Architecture in Large Scale Topology

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

For such scenario, we should consider using the Route Reflector (RR) [RFC4456] to achieve the similar effect. Every edge router will establish two BGP sessions with the RR via different loopback addresses respectively. The other steps for traffic differentiation are same as that described in the CCDR architecture for simple topology.

As shown in Figure 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 selects the dedicated path as R1-R2-R4-R7, then the operator should set the explicit peer routes via PCEP protocol 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.

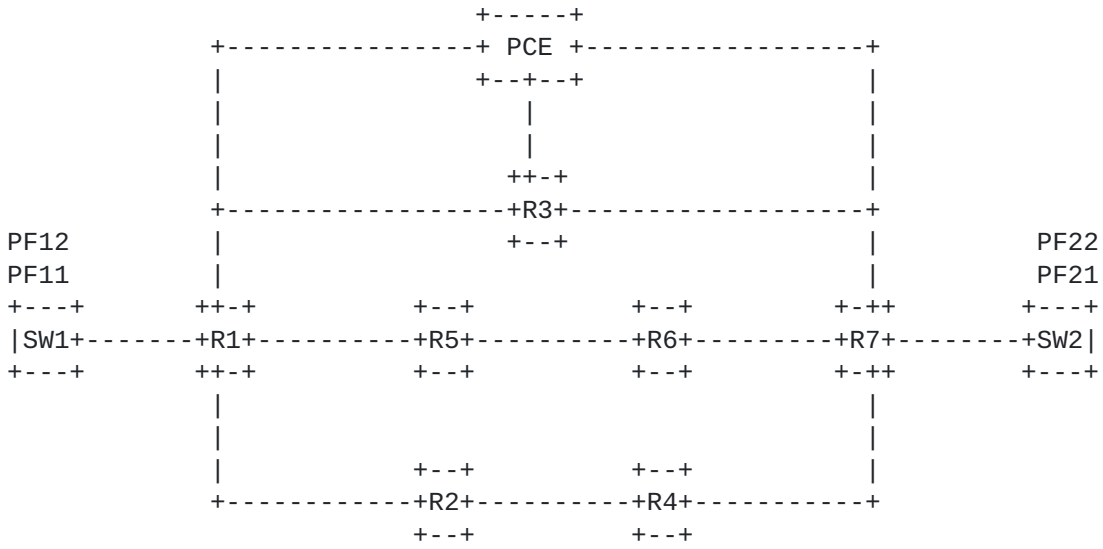

Figure 2: CCDR architecture in large scale network

5. CCDR Multiple BGP Sessions 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:

Prefix Set 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 Prefix Set No.1, we can select the shortest distance path to carry the traffic; for Prefix Set No.2, we can select the path that has end to end under-loading links; For Prefix Set No.3, we can let all assured traffic pass the determined single path, no Equal Cost Multipath (ECMP) distribution on the parallel links is desired.

It is almost impossible to provide an End-to-End (E2E) path efficiently 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, as that described in [RFC4655] and [RFC8283] 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.

The architecture to implement the CCDR Multiple BGP sessions strategy is the followings:

Here PCE is the main component of the Software Definition Network (SDN) controller and is responsible for optimal path computation for priority traffic.

- SDN controller gets topology via BGP-LS [RFC7752] and link utilization information via existing Network Monitor System (NMS) from the underlying network.

- PCE calculates the appropriate path upon application's requirements, sends the key parameters to edge/RR routers(R1, R7 and R3 in Fig.3) to establish multiple BGP sessions. The loopback addresses used for BGP sessions should be planned in advance and distributed in the domain.

- PCE sends the route information to the routers (R1,R2,R4,R7 in Fig.3) on forwarding path via PCEP [I-D.ietf-pce-pcep-extension-native-ip] , to build the path to the BGP next-hop of the advertised prefixes.

- PCE send the prefixes information to the PCC to let them advertises different prefixes via the specified BGP session.

- If the assured traffic prefixes were changed but the total volume of assured traffic does not exceed the physical capacity of the previous E2E 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 and add 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.

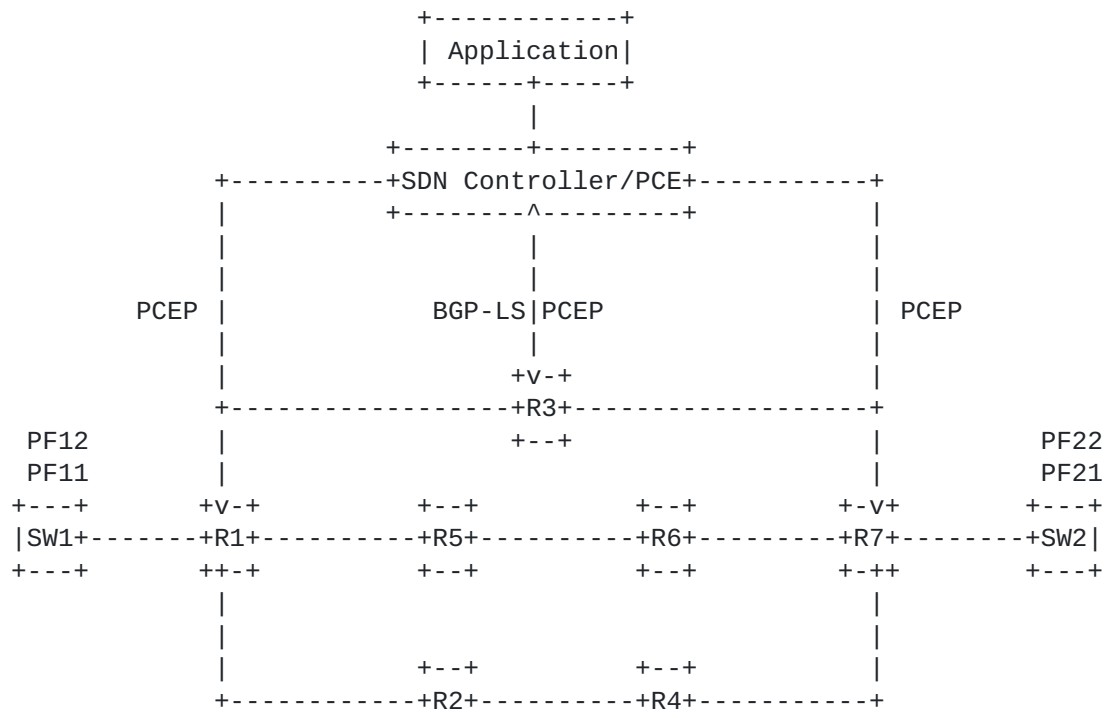

Figure 3: CCDR architecture for Multi-BGP deployment

6. PCEP Extension for Key Parameters Delivery

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

- Peer information that is used to build the BGP session

- Explicit route information to BGP next hop of advertised prefixes

- Advertised prefixes and their associated BGP session.

Once the router receives such information, it should establish the BGP session with the peer appointed in the PCEP message, build the end to end dedicated path hop by hop and advertise the prefixes that contained in the corresponding PCEP message.

The dedicated path is preferred by making sure that the explicit route created by PCE has the higher priority (lower route preference) than the route information created by other dynamic protocols.

All above dynamically created states (BGP sessions, Explicit route, Prefix advertised prefix,) will be cleared on the expiration of state timeout interval which is based on the existing Stateful PCE [RFC8231] and PCECC [RFC8283] mechanism.

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] .

7. Deployment Consideration

7.1. Scalability

In CCDR architecture, PCE needs only influence the edge routers for the prefixes advertisement via the multiple BGP sessions deployment. The route information for these prefixes within the on-path routers were distributed via the BGP protocol.

For multiple domains deployment, the PCE or the pool of PCEs that responsible for these domains need only control the edge router to build multiple EBGP sessions, all other procedures are the same that in one domain.

Unlike the solution from BGP Flowspec, the on-path router need only keep the specific policy routes for 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 BGP flowspec or Openflow solution. For example, if we want to differentiate 1000 prefixes from the normal traffic, CCDR needs only one explicit peer route in every on-path router, but the BGP flowspec or Openflow needs 1000 policy routes on them.

7.2. High Availability

The CCDR architecture 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 unchanged.

If one node on the optimal path is failed, the priority traffic will fall over to the best-effort forwarding path. One can even design several assurance paths to load balance/hot-standby the priority traffic to meet the path failure situation.

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

7.3. 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.

8. Security Considerations

A PCE needs to assure calculation of E2E path based on the status of network and the service requirements in real-time.

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

The setup of BGP session, prefix advertisement and explicit peer route establishment are all controlled by the PCE. To prevent the bogus PCE to send harmful messages to the network nodes, the network devices should authenticate the validity of PCE and keep secures communication channel between them. Mechanism described in [RFC8253] should be used to avoid such situation.

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

9. IANA Considerations

This document does not require any IANA actions.

10. Acknowledgement

The author would like to thank Deborah Brungard, Adrian Farrel, Vishnu Beeram, Lou Berger, Dhruv Dhody, Raghavendra Mallya , Mike Koldychev, Haomian Zheng, Penghui Mi, Shaofu Peng and Jessica Chen for their supports and comments on this draft.

11. References

11.1. Normative References

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, DOI 10.17487/RFC4655, August 2006, <https://www.rfc-editor.org/info/rfc4655>.

[RFC4657] Ash, J., Ed. and J. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol Generic Requirements", RFC 4657, DOI 10.17487/RFC4657, September 2006, <https://www.rfc-editor.org/info/rfc4657>.

[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>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC8231] Crabbe, E., Minei, I., Medved, J., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for Stateful PCE", RFC 8231, DOI 10.17487/RFC8231, September 2017, <https://www.rfc-editor.org/info/rfc8231>.

[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>.

[RFC8281] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "Path Computation Element Communication Protocol (PCEP) Extensions for PCE-Initiated LSP Setup in a Stateful PCE Model", RFC 8281, DOI 10.17487/RFC8281, December 2017, <https://www.rfc-editor.org/info/rfc8281>.

[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>.

[RFC8735] Wang, A., Huang, X., Kou, C., Li, Z., and P. Mi, "Scenarios and Simulation Results of PCE in a Native IP Network", RFC 8735, DOI 10.17487/RFC8735, February 2020, <https://www.rfc-editor.org/info/rfc8735>.

11.2. Informative References

[I-D.ietf-pce-pcep-extension-native-ip] Wang, A., Khasanov, B., Fang, S., Tan, R., and C. Zhu, "PCEP Extension for Native IP Network", draft-ietf-pce-pcep-extension-native-ip-09 (work in progress), October 2020.

Authors' Addresses

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

Email: wangaj3@chinatelecom.cn

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

Email: bhassanov@yahoo.com

Quintin Zhao
Etheric Networks
1009 S CLAREMONT ST
SAN MATEO, CA 94402
USA

Email: qzhao@ethericnetworks.com

Huaimo Chen
Futurewei
Boston, MA
USA

Email: huaimo.chen@futurewei.com