ï»¿

Internet Engineering Task Force (IETF)
Request for Comments: 8821
Category: Informational
ISSN: 2070-1721

A. Wang
China Telecom
B. Khasanov
Yandex LLC
Q. Zhao
Etheric Networks
H. Chen
Futurewei

April 2021

PCE-Based Traffic Engineering (TE) in Native IP Networks

Abstract

This document defines an architecture for providing traffic engineering in a native IP network using multiple BGP sessions and a Path Computation Element (PCE)-based central control mechanism. It defines the Centralized Control Dynamic Routing (CCDR) procedures and identifies needed extensions for the Path Computation Element Communication Protocol (PCEP).

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are candidates for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc8821.

Copyright Notice

Copyright (c) 2021 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 (https://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. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

[RFC8283], based on an extension of the PCE architecture described in [RFC4655], introduced a broader use applicability for a PCE as a central controller. PCEP continues to be used as the protocol between the PCE and the Path Computation Client (PCC). Building on that work, this document describes a solution of using a PCE for centralized control in a native IP network to provide end-to-end (E2E) performance assurance and QoS for traffic. The solution combines the use of distributed routing protocols and a centralized controller, referred to as Centralized Control Dynamic Routing (CCDR).

[RFC8735] describes the scenarios and simulation results for traffic engineering in a native IP network based on use of a CCDR architecture. Per [RFC8735], the architecture for traffic engineering in a 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 E2E traffic assurance, with determined QoS behavior, for traffic requiring a service assurance (prioritized traffic).

* No changes in a router's forwarding behavior.

* Based on centralized control through a distributed network control plane.

* Support different network requirements such as high traffic volume and prefix scaling.

* Ability to adjust the optimal path dynamically upon the changes of network status. No need for reserving resources for physical links in advance.

Building on the above documents, this document defines an architecture meeting these requirements by using a strategy of multiple BGP sessions and a PCE as the centralized controller. The architecture depends on the central control element (PCE) to compute the optimal path and utilizes the dynamic routing behavior of IGP and BGP for forwarding the traffic.

2. Terminology

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

PCE: Path Computation Element

PCEP: PCE Protocol

PCC: Path Computation Client

Other terms are used in this document:

CCDR: Centralized Control Dynamic Routing

E2E: End to End

ECMP: Equal-Cost Multipath

RR: Route Reflector

SDN: Software-Defined Network

3. CCDR Architecture in a Simple Topology

Figure 1 illustrates the CCDR architecture for traffic engineering in a simple topology. The topology is composed of four devices, which are SW1, SW2, R1, and 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.

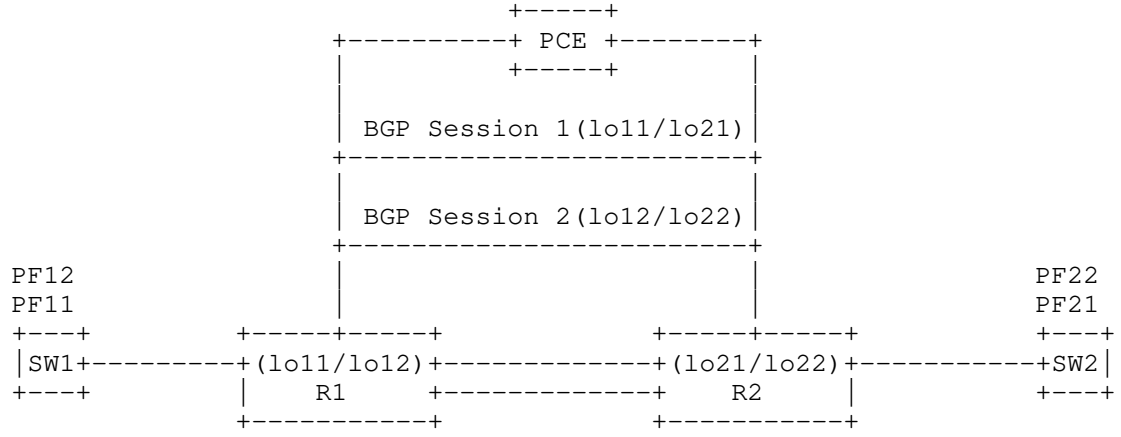


Figure 1: CCDR Architecture in a Simple Topology

In the intra-domain scenario, IGP and BGP combined with a PCE are deployed between R1 and R2. In the inter-domain scenario, only native BGP 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 following:

* Build two BGP sessions between R1 and R2 via the different loopback addresses on these routers (lo11 and lo12 are the loopback addresses of R1, and lo21 and lo22 are the loopback addresses of R2).

* Using the PCE, set the explicit peer route on R1 and R2 for BGP next hop to different physical link addresses between R1 and R2. The explicit peer route can be set in the format of a static route, which is different from the route learned from IGP.

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

After the above actions, the bidirectional traffic between the PF11 and PF21, and the bidirectional 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 assured transport, one can add more physical links between R1 and R2 to reach the next hop for BGP session 2. In this case, the prefixes that are advertised by the BGP peers need not be changed.

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

Such a decoupling philosophy of the IGP/BGP traffic link and the physical link achieves a flexible control capability for the network traffic, satisfying the needed QoS assurance to meet the application's requirement. The router needs only to support native IP and multiple BGP sessions set up via different loopback addresses.

4. CCDR Architecture in a Large-Scale Topology

When the priority traffic spans a large-scale network, such as that illustrated in Figure 2, the multiple BGP sessions cannot be established hop by hop within one autonomous system. For such a scenario, we propose using a Route Reflector (RR) [RFC4456] to achieve a 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 the same as that described in the CCDR architecture for the 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 sessions 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 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 priority traffic) to the selected forwarding address.

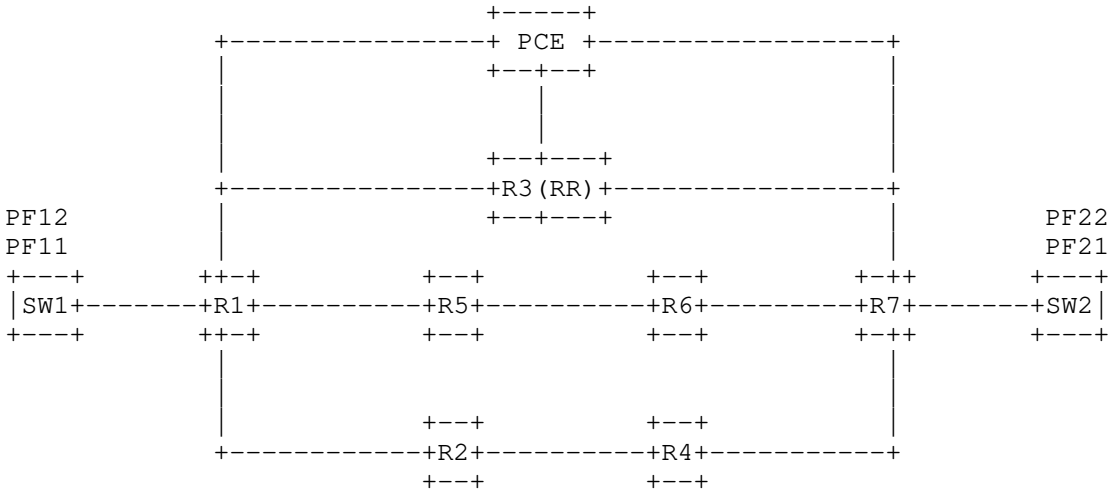


Figure 2: CCDR Architecture in a Large-Scale Network

5. CCDR Multiple BGP Sessions Strategy

Generally, 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 are summarized in Table 1.

Prefix Set 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 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 E2E under-loaded links; for Prefix Set No.3, we can let traffic pass over a determined single path, as no ECMP distribution on the parallel links is desired.

It is almost impossible to provide an E2E path efficiently with latency, jitter, and packet loss constraints to meet the above requirements in a large-scale, IP-based network only using a distributed routing protocol, but these requirements can be met with the assistance of PCE, as described in [RFC4655] and [RFC8283]. The PCE will have the overall network view, ability to collect the real-time network topology, and the network performance information about the underlying network. The PCE can select the appropriate path to meet the various network performance requirements for different traffic.

The architecture to implement the CCDR multiple BGP sessions strategy is as follows:

The PCE will be responsible for the optimal path computation for the different priority classes of traffic:

* PCE collects topology information via BGP-LS [RFC7752] and link utilization information via the existing Network Monitoring System (NMS) from the underlying network.

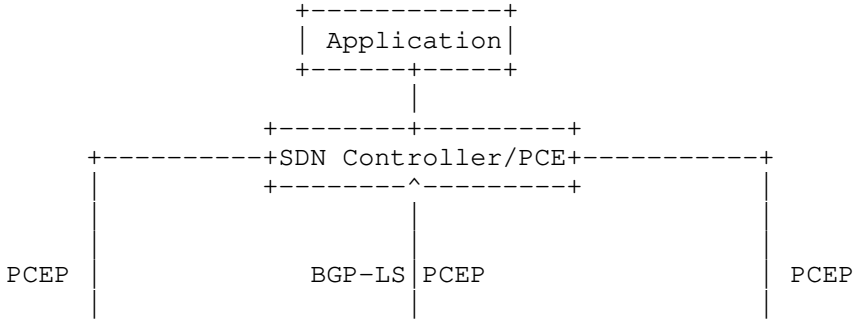
* PCE calculates the appropriate path based upon the application's requirements and sends the key parameters to edge/RR routers (R1, R7, and R3 in Figure 3) to establish multiple BGP sessions. The loopback addresses used for the BGP sessions should be planned in advance and distributed in the domain.

* PCE sends the route information to the routers (R1, R2, R4, and R7 in Figure 3) on the forwarding path via PCEP to build the path to the BGP next hop of the advertised prefixes. The path to these BGP next hops will also be learned via IGP, but the route from the PCEP has the higher preference. Such a design can assure the IGP path to the BGP next hop can be used to protect the path assigned by PCE.

* PCE sends the prefix information to the PCC (edge routers that have established BGP sessions) for advertising different prefixes via the specified BGP session.

* The priority traffic may share some links or nodes if the path the shared links or nodes can meet the requirement of application. When the priority traffic prefixes are changed, but the total volume of priority traffic does not exceed the physical capacity of the previous E2E path, the PCE needs only change the prefixes advertised via the edge routers (R1 and R7 in Figure 3).

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



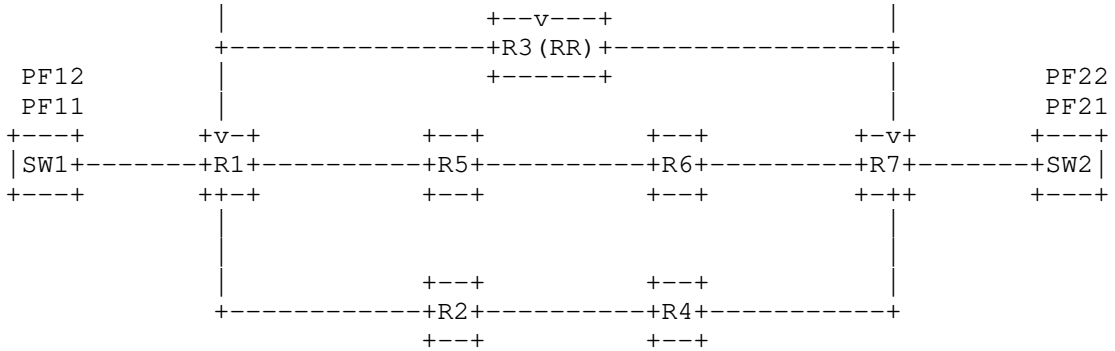


Figure 3: CCDR Architecture for Multi-BGP Sessions Deployment

6. PCEP Extension for Critical Parameters Delivery

PCEP needs to be extended to transfer the following critical parameters:

* Peer information that is used to build the BGP session.

* Explicit route information for 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 E2E dedicated path hop by hop, and advertise the prefixes that are 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 of the above dynamically created states (BGP sessions, explicit routes, and advertised prefixes) will be cleared on the expiration of the state timeout interval, which is based on the existing stateful PCE [RFC8231] and PCE as a Central Controller (PCECC) [RFC8283] mechanism.

Regarding the BGP session, it is not different from that configured manually or via Network Configuration Protocol (NETCONF) and YANG. Different BGP sessions are used mainly for the clarification of the network prefixes, which can be differentiated via the different BGP next hop. Based on this strategy, if we manipulate the path to the BGP next hop, then the path to the prefixes that were advertised with the BGP sessions will be changed accordingly. Details of communications between PCEP and BGP subsystems in the router's control plane are out of scope of this document.

7. Deployment Considerations

7.1. Scalability

In the CCDR architecture, only the edge routers that connect with the PCE are responsible for the prefix advertisement via the multiple BGP sessions deployment. The route information for these prefixes within the on-path routers is distributed via BGP.

For multiple domain deployment, the PCE, or the pool of PCEs responsible for these domains, needs only to control the edge router to build the multiple External BGP (EBGP) sessions; all other procedures are the same as within one domain.

The on-path router needs only to keep the specific policy routes for the BGP next hop of the differentiated prefixes, not the specific routes to the prefixes themselves. This lessens the burden of the table size of policy-based routes for the on-path routers; and has more expandability compared with BGP Flowspec or OpenFlow solutions.

For example, if we want to differentiate 1,000 prefixes from the normal traffic, CCDR needs only one explicit peer route in every on-path router, whereas the BGP Flowspec or OpenFlow solutions need 1,000 policy routes on them.

7.2. High Availability

The CCDR architecture is based on the use of native IP. If the PCE fails, the forwarding plane will not be impacted, as the BGP sessions between all the devices will not flap, and the forwarding table remains unchanged.

If one node on the optimal path fails, the priority traffic will fall over to the best-effort forwarding path. One can even design several paths to load balance or to create a hot standby of the priority traffic to meet a path failure situation.

For ensuring high availability of a PCE/SDN-controllers architecture, an operator should rely on existing high availability solutions for SDN controllers, such as clustering technology and deployment.

7.3. Incremental Deployment

Not every router within the network needs to support the necessary PCEP extension. For such situations, routers on the edge of a domain can be upgraded first, and then the traffic can be prioritized between different domains. Within each domain, the traffic will be forwarded along the best-effort path. A service provider can selectively upgrade the routers on each domain in sequence.

7.4. Loop Avoidance

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

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

7.5. E2E Path Performance Monitoring

It is necessary to deploy the corresponding E2E path performance monitoring mechanism to assure that the delay, jitter, or packet loss index meets the original path performance aim. The performance monitoring results should provide feedback to the PCE in order for it to accomplish the re-optimization process and send the update control message to the related PCC if necessary. Traditional OAM methods (ping, trace) can be used.

8. Security Considerations

The setup of BGP sessions, prefix advertisement, and explicit peer route establishment are all controlled by the PCE. See [RFC4271] and [RFC4272] for BGP security considerations. The Security Considerations found in Section 10 of [RFC5440] and Section 10 of [RFC8231] should be considered. To prevent a bogus PCE sending harmful messages to the network nodes, the network devices should authenticate the validity of the PCE and ensure a secure communication channel between them. Mechanisms described in [RFC8253] should be used.

The CCDR architecture does not require changes to the forwarding behavior of the underlay devices. There are no additional security impacts on these devices.

9. IANA Considerations

This document has no IANA actions.

10. References

10.1. Normative References

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[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>.

[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>.

[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>.

10.2. Informative References

[RFC4655] Farrel, A., Vasseur, J.-P., 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>.

[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>.

Acknowledgments

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, Donald Eastlake, Alvaro Retana, Martin Duke, Magnus Westerlund, Benjamin Kaduk, Roman Danyliw, Ã\211ric Vyncke, Murray Kucherawy, Erik Kline, and Jessica Chen for their supports and comments on this document.

Authors' Addresses

Aijun Wang
China Telecom
Changping District

Beiqijia Town
Beijing
102209
China

Email: wangaj3@chinatelecom.cn

Boris Khasanov
Yandex LLC
Ulitsa Lva Tolstogo 16
Moscow
Russian Federation

Email: bhassanov@yahoo.com

Quintin Zhao
Etheric Networks
1009 S Claremont St
San Mateo, CA 94402
United States of America

Email: qzhao@ethericnetworks.com

Huaimo Chen
Futurewei
Boston, MA
United States of America

Email: huaimo.chen@futurewei.com