TEAS Working Group A. Wang

Internet-Draft China Telecom

Intended status: Informational B. Khasanov

Expires: May 29, 2021 Yandex LLC

Q. Zhao

Etheric Networks

H. Chen

Futurewei

November 25, 2020

Path Computation Element (PCE) Traffic Engineering (TE) in Native IP Networks

draft-ietf-teas-pce-native-ip-14

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 Central Control Dynamic Routing (CCDR) procedures and

identifies needed extensions for the Path Computation Element

Communication Protocol (PCEP).

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 May 29, 2021.

Copyright Notice

Copyright (c) 2020 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

Wang, et al. Expires May 29, 2021 [Page 1]

Internet-Draft PCE in Native IP Network November 2020

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 . . . . . . . . . . . . . . . . . . . . . . . . 2

2. Terminology . . . . . . . . . . . . . . . . . . . . . . . . . 3

3. CCDR Architecture in Simple Topology . . . . . . . . . . . . 4

4. CCDR Architecture in Large Scale Topology . . . . . . . . . . 5

5. CCDR Multiple BGP Sessions Strategy . . . . . . . . . . . . . 6

6. PCEP Extension for Key Parameters Delivery . . . . . . . . . 8

7. Deployment Consideration . . . . . . . . . . . . . . . . . . 9

7.1. Scalability . . . . . . . . . . . . . . . . . . . . . . . 9

7.2. High Availability . . . . . . . . . . . . . . . . . . . . 9

7.3. Incremental deployment . . . . . . . . . . . . . . . . . 10

7.4. Loop Avoidance . . . . . . . . . . . . . . . . . . . . . 10

8. Security Considerations . . . . . . . . . . . . . . . . . . . 10

9. IANA Considerations . . . . . . . . . . . . . . . . . . . . . 10

10. Acknowledgement . . . . . . . . . . . . . . . . . . . . . . . 11

11. References . . . . . . . . . . . . . . . . . . . . . . . . . 11

11.1. Normative References . . . . . . . . . . . . . . . . . . 11

11.2. Informative References . . . . . . . . . . . . . . . . . 12

Authors' Addresses . . . . . . . . . . . . . . . . . . . . . . . 12

1. Introduction

[RFC8283], based on an extension of the PCE (Path Computation Element) architecture described in

[RFC4655], introduced a broader use applicability for a PCE as a

central controller. PCEP (PCE Protocol) continues to be used as the protocol

between PCE and PCC (Path Computation Client). Building on that work, this document describes

a solution 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:

o Same solution for native IPv4 and IPv6 traffic.

Wang, et al. Expires May 29, 2021 [Page 2]

Internet-Draft PCE in Native IP Network November 2020

o Support for intra-domain and inter-domain scenarios.

o Achieve End to End traffic assurance, with determined QoS

behavior, for traffic requiring a service assurance (prioritized

traffic).

o No changes in router's forwarding behavior.

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

o Support different network requirements such as high traffic

volume and prefix scaling.

o Ability to adjust the optimal path dynamically upon the changes of

network status. No need for physical links resources reservations

to be done in advance.

Building on the above documents, this document defines an

architecture meeting these requirements by using a multiple BGP

session strategy and a PCE as the centralized controller. The

architecture depends on the central control (PCE) element to compute

the optimal path, and utilizes the dynamic routing behavior of IGP/

BGP protocols for forwarding the traffic.

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]:

o PCE – Path Computation Element

o PCEP – PCE Protocol

o PCC – Path Computation Client

Other terms are defined in this document:

o CCDR: Central Control Dynamic Routing

o E2E: End to End

o ECMP: Equal-Cost Multipath

o RR: Route Reflector

Wang, et al. Expires May 29, 2021 [Page 3]

Internet-Draft PCE in Native IP Network November 2020

o 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 comprises 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.

+-----+

+----------+ PCE +--------+

| +-----+ |

| |

| BGP Session 1(lo11/lo21)|

+-------------------------+

| |

| BGP Session 2(lo12/lo22)|

+-------------------------+

PF12 | | PF22

PF11 | | PF21

+---+ +-----+-----+ +-----+-----+ +---+

|SW1+---------+(lo11/lo12)+-------------+(lo21/lo22)+--------------+SW2|

+---+ | R1 +-------------+ R2 | +---+

+-----------+ +-----------+

Figure 1: CCDR architecture in simple topology

In the Intra-AS scenario, IGP and BGP combined with a PCE are

deployed between R1 and R2. In the inter-AS scenario, only the

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 following:

o Build two BGP sessions between R1 and R2, via the different

loopback addresses on these routers.

o 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 the IGP

protocol.

Wang, et al. Expires May 29, 2021 [Page 4]

Internet-Draft PCE in Native IP Network November 2020

o 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-directional traffic between the PF11

and PF21, and the bi-directional 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 bi-directional 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

bi-directional 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, achieving the needed QoS assurance 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 priority traffic spans a large scale network, such as that

illustrated in Figure 2, the multiple BGP sessions cannot be

established hop by hop, for example, the iBGP within one AS.

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 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

priority traffic) to the selected forwarding address.

Wang, et al. Expires May 29, 2021 [Page 5]

Internet-Draft PCE in Native IP Network November 2020

+-----+

+----------------+ PCE +------------------+

| +--+--+ |

| | |

| | |

| ++-+ |

+------------------+R3+-------------------+

PF12 | +--+ | PF22

PF11 | | PF21

+---+ ++-+ +--+ +--+ +-++ +---+

|SW1+-------+R1+----------+R5+----------+R6+---------+R7+--------+SW2|

+---+ ++-+ +--+ +--+ +-++ +---+

| |

| |

| +--+ +--+ |

+------------+R2+----------+R4+-----------+

+--+ +--+

Figure 2: CCDR architecture in large scale network

5. CCDR Multiple BGP Sessions Strategy

Generally, different applications may require different QoS criteria,

which may include:

o Traffic that requires low latency and is not sensitive to packet

loss.

o Traffic that requires low packet loss and can endure higher

latency.

o 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 | 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

Wang, et al. Expires May 29, 2021 [Page 6]

Internet-Draft PCE in Native IP Network November 2020

has end to end under-loaded links; for Prefix Set No.3, we can let

traffic pass over a determined single path, as 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, 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 that 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 the follows:

The PCE will be responsible for the optimal path computation for the

different priority classes of traffic:

o PCE collects topology information via BGP-LS [RFC7752] and link

utilization information via the existing Network Monitoring System

(NMS) from the underlying network.

o 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.

o PCE sends the route information to the routers (R1,R2,R4,R7 in

Figure 3) on the 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.

o PCE sends the prefixes information to the PCC for advertising

different prefixes via the specified BGP session.

o If the priority traffic prefixes were 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 prefixed

advertised via the edge routers (R1,R7 in Figure 3).

o 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

Wang, et al. Expires May 29, 2021 [Page 7]

Internet-Draft PCE in Native IP Network November 2020

that, the PCE needs to update the on-path routers to build the

forwarding path hop by hop.

+------------+

| Application|

+------+-----+

|

+--------+---------+

+----------+SDN Controller/PCE+-----------+

| +--------^---------+ |

| | |

| | |

PCEP | BGP-LS|PCEP | PCEP

| | |

| +v-+ |

+------------------+R3+-------------------+

PF12 | +--+ | PF22

PF11 | | PF21

+---+ +v-+ +--+ +--+ +-v+ +---+

|SW1+-------+R1+----------+R5+----------+R6+---------+R7+--------+SW2|

+---+ ++-+ +--+ +--+ +-++ +---+

| |

| |

| +--+ +--+ |

+------------+R2+----------+R4+-----------+

+--+ +--+

Figure 3: CCDR architecture for Multi-BGP sessions deployment

6. PCEP Extension for Critical Parameters Delivery

The PCEP protocol needs to be extended to transfer the following critical

parameters:

o Peer information that is used to build the BGP session

o Explicit route information for BGP next hop of advertised prefixes

o 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

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.

Wang, et al. Expires May 29, 2021 [Page 8]

Internet-Draft PCE in Native IP Network November 2020

All above dynamically created states (BGP sessions, Explicit route, and

Prefix advertised prefix) will be cleared on the expiration of the

state timeout interval which is based on the existing Stateful PCE

[RFC8231] and PCECC [RFC8283] mechanism.

Regarding the BGP session, it is not different from that configured

manually or via NETCONF/YANG. Different BGP sessions are used

mainly for the clarification of the network prefixes, which can be

differentiated via the different BGP nexthop. Based on this

strategy, if we manipulate the path to the BGP nexthop, 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

draft and will be described in a separate document

[I-D.ietf-pce-pcep-extension-native-ip].

7. Deployment Consideration

7.1. Scalability

In the CCDR architecture, only the edge routers that connect with the

PCE are responsible for the prefixes advertisement via the multiple

BGP sessions deployment. The route information for these prefixes

within the on-path routers is distributed via the BGP protocol.

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 EBGP sessions; all other procedures are the

same as within one domain.

Unlike the solution from BGP Flowspec [RFC5575bis], the on-path router needs only

to keep the specific policy routes for the BGP next-hop of the

differentiate 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 1000 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 1000 policy routes on

them.

7.2. High Availability

The CCDR architecture is based on the use of the native IP protocol.

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.

Wang, et al. Expires May 29, 2021 [Page 9]

Internet-Draft PCE in Native IP Network November 2020

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/hot-standby 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 PCEP extension

defined in [I-D.ietf-pce-pcep-extension-native-ip] simultaneously.

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.

8. Security Considerations

The setup of BGP sessions, prefix advertisement, and explicit peer

route establishment are all controlled by the PCE. See [RFC7454] for BGP Security Considerations. 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 will no additional security

impacts on these devices.

9. IANA Considerations

This document does not require any IANA actions.

Wang, et al. Expires May 29, 2021 [Page 10]

Internet-Draft PCE in Native IP Network November 2020

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, Donald Eastlake, 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>.

[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>.

[RFC7454] Durand, J., Pepelnjak, I., and G. Doering, "BGP Operations and Security", BCP 194, RFC 7454, DOI 10.17487/RFC7454, February 2015, <https://www.rfc-editor.org/info/rfc7454>.

[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>.

Wang, et al. Expires May 29, 2021 [Page 11]

Internet-Draft PCE in Native IP Network November 2020

[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

[RFC5575bis] C. Loibl, S. Hares, R. Raszuk, D. McPherson, M. Bacher, "Dissemination of Flow Specification Rules", draft-ietf-idr-rfc5575bis-27, October 2020, work in progress.

[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

Yandex LLC

Ulitsa Lva Tolstogo 16

Moscow

Russia

Email: bhassanov@yahoo.com

Quintin Zhao

Etheric Networks

1009 S CLAREMONT ST

SAN MATEO, CA 94402

USA

Email: qzhao@ethericnetworks.com

Wang, et al. Expires May 29, 2021 [Page 12]

Internet-Draft PCE in Native IP Network November 2020

Huaimo Chen

Futurewei

Boston, MA

USA

Email: huaimo.chen@futurewei.com

Wang, et al. Expires May 29, 2021 [Page 13]