Benchmarking Methodology for SDN Controller Performance
draft-bhuvan-bmwg-sdn-controller-benchmark-meth-01

Abstract

This document defines the methodologies for benchmarking performance of SDN controllers. Terminology related to benchmarking SDN controllers is described in the companion terminology document. SDN controllers have been implemented with many varying designs in order to achieve their intended network functionality. Hence, the authors have taken the approach of considering an SDN controller as a black box, defining the methodology in a manner that is agnostic to protocols and network services supported by controllers. The intent of this document is to provide a standard mechanism to measure the performance of all controller implementations.

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 http://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 January 18, 2016.

Copyright Notice

Copyright (c) 2015 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. 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 ............................................. 3
2. Scope .................................................. 3
3. Test Setup .............................................. 3
3.1 Test setup - Controller working in Standalone Mode .... 4
3.2 Test setup - Controller working in Cluster Mode ....... 5
4. Test Considerations ...................................... 6
4.1 Network Topology ....................................... 6
4.2 Test Traffic ........................................... 6
4.3 Connection Setup ....................................... 6
4.4 Measurement Point Specification and Recommendation . 7
4.5 Connectivity Recommendation ........................... 7
4.6 Test Repeatability ...................................... 7
5. Test Reporting .......................................... 7
6. Benchmarking Tests ..................................... 8
6.1 Performance ........................................... 8
6.1.1 Network Topology Discovery Time .................... 8
6.1.2 Asynchronous Message Processing Time .............. 9
6.1.3 Asynchronous Message Processing Rate ............... 11
6.1.4 Reactive Path Provisioning Time ..................... 12
6.1.5 Proactive Path Provisioning Time .................... 13
6.1.6 Reactive Path Provisioning Rate ...................... 14
6.1.7 Proactive Path Provisioning Rate .................... 15
6.1.8 Network Topology Change Detection Time ............ 16
6.2 Scalability ............................................ 17
6.2.1 Control Sessions Capacity ......................... 17
6.2.2 Network Discovery Size ............................ 18
6.2.3 Forwarding Table Capacity .......................... 19
6.3 Security ............................................... 20
6.3.1 Exception Handling .................................. 20
6.3.2 Denial of Service Handling ......................... 21
6.4 Reliability ............................................ 22
6.4.1 Controller Failover Time ............................ 22
6.4.2 Network Re-Provisioning Time ....................... 23
1. Introduction

This document provides generic methodologies for benchmarking SDN controller performance. An SDN controller may support many northbound and southbound protocols, implement a wide range of applications, and work solely, or as a group to achieve the desired functionality. This document considers an SDN controller as a black box, regardless of design and implementation. The tests defined in the document can be used to benchmark SDN controller for performance, scalability, reliability and security independent of northbound and southbound protocols. These tests can be performed on an SDN controller running as a virtual machine (VM) instance or on a bare metal server. This document is intended for those who want to measure the SDN controller performance as well as compare various SDN controllers performance.

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.

2. Scope

This document defines methodology to measure the networking metrics of SDN controllers. The tests defined in this document enable benchmarking of SDN Controllers in two ways; as a standalone controller and as a cluster of homogeneous controllers. These tests are recommended for execution in lab environments rather than in live network deployments. Performance benchmarking of a federation of controllers is beyond the scope of this document.

3. Test Setup

The tests defined in this document enable measurement of an SDN controllers performance in standalone mode and cluster mode. This section defines common reference topologies that are later referred to in individual tests.
3.1 Test setup - Controller working in Standalone Mode

![Diagram of test setup]
3.2 Test setup - Controller working in Cluster Mode

Figure 2
4. Test Considerations

4.1 Network Topology

The test cases SHOULD use Leaf-Spine topology with at least one SDN node in the topology for benchmarking. The test traffic generators TP1 and TP2 SHOULD be connected to the first and the last SDN leaf node. If a test case uses a test topology with one SDN node, the test traffic generators TP1 and TP2 SHOULD be connected to the same node. However, to achieve a complete performance characterization of the SDN controller, it is recommended that the controller be benchmarked for many network topologies and varying numbers of SDN nodes. This document includes a few sample test topologies, defined in Section 10 - Appendix A for reference. Further, care should be taken to make sure that a loop prevention mechanism is enabled either in the SDN controller, or in the network when the topology contains redundant network paths.

4.2 Test Traffic

Test traffic are used to notify the controller about the arrival of new flows. The test cases SHOULD use multiple frame sizes as recommended in RFC 2544 for benchmarking.

4.3 Connection Setup

There may be controller implementations that support unencrypted and encrypted network connections with SDN nodes. Further, the controller may have backward compatibility with SDN nodes running older versions of southbound protocols. It is recommended that the controller performance be measured with one or more applicable connection setup methods defined below.

1. Unencrypted connection with SDN nodes, running same protocol version.
2. Unencrypted connection with SDN nodes, running different protocol versions.
   Example:
   1. Controller running current protocol version and switch running older protocol version
   2. Controller running older protocol version and switch running current protocol version
3. Encrypted connection with SDN nodes, running same protocol version
4. Encrypted connection with SDN nodes, running different protocol versions.
   Example:
   1. Controller running current protocol version and switch running older protocol version
   2. Controller running older protocol version and switch running current protocol version

4.4 Measurement Point Specification and Recommendation

The measurement accuracy depends on several factors including the point of observation where the indications are captured. For example, the notification can be observed at the controller or test emulator. The test operator SHOULD make the observations/measurements at the interfaces of test emulator unless it is explicitly mentioned otherwise in the individual test.

4.5 Connectivity Recommendation

The SDN controller in the test setup SHOULD be connected directly with the forwarding and the management plane test emulators to avoid any delays or failure introduced by the intermediate devices during benchmarking tests.

4.6 Test Repeatability

To increase the confidence in measured result, it is recommended that this test SHOULD be performed at least 10 times with same number of nodes using same topology.

5. Test Reporting

Each test has a reporting format which is specific to individual tests. In addition, the following test configuration parameters and controller settings parameters MUST be reflected in the test report.

Test Configuration Parameters:
1. Controller name and version
2. Northbound protocols and versions
3. Southbound protocols and versions
4. Controller redundancy mode (Standalone or Cluster Mode)
5. Connection setup (Unencrypted or Encrypted)
6. Network Topology (Mesh or Tree or Linear)
7. SDN Node Type (Physical or Virtual or Emulated)
8. Number of Nodes
9. Number of Links
10. Test Traffic Type
11. Controller System Configuration (e.g., CPU, Memory, Operating System, Interface Speed etc.,)
12. Reference Test Setup (e.g., Section 3.1 etc.,)

Controller Settings Parameters:
1. Topology re-discovery timeout
2. Controller redundancy mode (e.g., active-standby etc.,)

6. Benchmarking Tests

6.1 Performance

6.1.1 Network Topology Discovery Time

Objective:
Measure the time taken by the SDN controller to discover the network topology (nodes and links), expressed in milliseconds.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller MUST support network discovery.
2. Tester should be able to retrieve the discovered topology information either through the controller’s management interface or northbound interface to determine if the discovery was successful and complete.
3. Ensure that the controller’s topology re-discovery timeout has been set to the maximum value to avoid initiation of re-discovery process in the middle of the test.

Procedure:
1. Ensure that the controller is operational, its network applications, northbound and southbound interfaces are up and running.
2. Establish the network connections between controller and SDN nodes.
3. Record the time for the first discovery message (Tm1) received from the controller at forwarding plane test emulator interface I1.
4. Query the controller every 3 seconds to obtain the discovered network topology information through the northbound interface or the management interface and compare it with the deployed network topology information.

5. Stop the test when the discovered topology information is matching with the deployed network topology or the discovered topology information for 3 consecutive queries return the same details.

6. Record the time last discovery message (Tmn) sent to controller from the forwarding plane test emulator interface (I1) when the test completed successfully. (e.g., the topology matches).

Measurement:
Topology Discovery Time $T_{r1} = T_{mn} - T_{m1}$

$$\frac{T_{r1} + T_{r2} + T_{r3} \ldots T_{rn}}{\text{Total Test Iterations}}$$

Reporting Format:
The Topology Discovery Time results MUST be reported in the format of a table, with a row for each successful iteration. The last row of the table indicates the average Topology Discovery Time.

If this test is repeated with varying number of nodes over the same topology, the results SHOULD be reported in the form of a graph. The X coordinate SHOULD be the Number of nodes (N), the Y coordinate SHOULD be the average Topology Discovery Time.

If this test is repeated with same number of nodes over different topologies, the results SHOULD be reported in the form of a graph. The X coordinate SHOULD be the Topology Type, the Y coordinate SHOULD be the average Topology Discovery Time.

6.1.2 Asynchronous Message Processing Time

Objective:
Measure the time taken by the SDN controller to process an asynchronous message, expressed in milliseconds.

Reference Test Setup:
This test SHOULD use one of the test setup described in section 3.1 or section 3.2 of this document.
Prerequisite:
The controller MUST have completed the network topology
discovery for the connected SDN nodes.

Procedure:
1. Generate asynchronous messages from every connected SDN node,
to the SDN controller, one at a time in series from the
forwarding plane test emulator for the test duration.
2. Record every request transmit (T1) timestamp and the
   corresponding response (R1) received timestamp at the
   forwarding plane test emulator interface (I1) for every
   successful message exchange.

Measurement:
\[
\text{Asynchronous Message Processing Time } Tr_1 = \frac{(R_1-T_1) + (R_2-T_2) + \ldots (R_n-T_n)}{N_{rx}}
\]
Where \( N_{rx} \) is the total number of successful messages exchanged

\[
\text{Average Asynchronous Message Processing Time} = \frac{Tr_1 + Tr_2 + Tr_3 + \ldots Tr_n}{\text{Total Test Iterations}}
\]

Reporting Format:
The Asynchronous Message Processing Time results MUST be
reported in the format of a table with a row for each iteration.
The last row of the table indicates the average Asynchronous
Message Processing Time.

The report should capture the following information in addition
to the configuration parameters captured in section 5.
- Successful messages exchanged (\( N_{rx} \))

If this test is repeated with varying number of nodes with same
topology, the results SHOULD be reported in the form of a graph.
The X coordinate SHOULD be the Number of nodes (N), the
Y coordinate SHOULD be the average Asynchronous Message Processing
Time.

If this test is repeated with same number of nodes using
different topologies, the results SHOULD be reported in the form
of a graph. The X coordinate SHOULD be the Topology Type, the
Y coordinate SHOULD be the average Asynchronous Message Processing
Time.
6.1.3 Asynchronous Message Processing Rate

Objective:
To measure the maximum number of asynchronous messages (session aliveness check message, new flow arrival notification message etc.) a controller can process within the test duration, expressed in messages processed per second.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller MUST have completed the network topology discovery for the connected SDN nodes.

Procedure:
1. Generate asynchronous messages continuously at the maximum possible rate on the established connections from all the connected SDN nodes in the forwarding plane test emulator for the Test Duration (Td).
2. Record total number of responses received from the controller (Nrx) as well as the number of messages sent (Ntx) to the controller within the test duration (Td) at the forwarding plane test emulator interface (I1).

Measurement:

\[
\text{Asynchronous Message Processing Rate Tr1} = \frac{Nrx}{Td} \quad \text{Tr1 + Tr2 + Tr3...Trn}
\]

Average Asynchronous Message Processing Rate = \( \frac{\text{Asynchronous Message Processing Rate Tr1}}{\text{Total Test Iterations}} \)

Loss Ratio = \( \frac{(Ntx-Nrx)}{100} \).

Reporting Format:
The Asynchronous Message Processing Rate results MUST be reported in the format of a table with a row for each iteration. The last row of the table indicates the average Asynchronous Message Processing Rate.

The report should capture the following information in addition to the configuration parameters captured in section 5.
- Offered rate (Ntx)
- Loss Ratio
If this test is repeated with varying number of nodes over same topology, the results SHOULD be reported in the form of a graph. The X coordinate SHOULD be the Number of nodes (N), the Y coordinate SHOULD be the average Asynchronous Message Processing Rate.

If this test is repeated with same number of nodes over different topologies, the results SHOULD be reported in the form of a graph. The X coordinate SHOULD be the Topology Type, the Y coordinate SHOULD be the average Asynchronous Message Processing Rate.

6.1.4 Reactive Path Provisioning Time

Objective:
To measure the time taken by the controller to setup a path reactively between source and destination node, expressed in milliseconds.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller MUST contain the network topology information for the deployed network topology.
2. The controller should have the knowledge about the location of destination endpoint for which the path has to be provisioned. This can be achieved through dynamic learning or static provisioning.
3. Ensure that the default action for flow miss in SDN node is 'send to controller'.
4. Ensure that each SDN node in a path requires the controller to make the forwarding decision while paving the entire path.

Procedure:
1. Send a single traffic stream from test traffic generator TP1 to test traffic generator TP2.
2. Record the time of the first flow provisioning request message sent to the controller(Tsf1) from the SDN node at the forwarding plane test emulator interface (I1).
3. Wait for the arrival of first traffic frame at the Traffic Endpoint TP2 or the expiry of test duration (Td).
4. Record the time of the last flow provisioning response message received from the controller(Tdf1) to the SDN node at the forwarding plane test emulator interface (I1).
Measurement:
Reactive Path Provisioning Time \( T_{rl} = T_{df1} - T_{sf1} \).

\[ \frac{Tr_1 + Tr_2 + Tr_3 \ldots Tr_n}{\text{Total Test Iterations}} \]

Reporting Format:
The Reactive Path Provisioning Time results MUST be reported in the format of a table with a row for each iteration. The last row of the table indicates the Average Reactive Path Provisioning Time.

The report should capture the following information in addition to the configuration parameters captured in section 5.
- Number of SDN nodes in the path

6.1.5 Proactive Path Provisioning Time

Objective:
To measure the time taken by the controller to setup a path proactively between source and destination node, expressed in milliseconds.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller MUST contain the network topology information for the deployed network topology.
2. The controller should have the knowledge about the location of destination endpoint for which the path has to be provisioned. This can be achieved through dynamic learning or static provisioning.
3. Ensure that the default action for flow miss in SDN node is ‘drop’.

Procedure:
1. Send single traffic stream from test traffic generator TP1 to TP2.
2. Install the flow entries to reach from test traffic generator TP1 to the test traffic generator TP2 through controller’s northbound or management interface.
3. Wait for the arrival of first traffic frame at the test traffic generator TP2 or the expiry of test duration (\( T_d \)).
4. Record the time when proactive flow is provisioned in the Controller (\( T_{sf1} \)) at the management plane test emulator interface I2.
5. Record the time of the last flow provisioning message received from the controller (Tdf1) at the forwarding plane test emulator interface I1.

Measurement:
Proactive Flow Provisioning Time Tr1 = Tdf1 - Tsf1.

\[
\text{Average Proactive Path Provisioning Time} = \frac{Tr1 + Tr2 + Tr3 \ldots Trn}{\text{Total Test Iterations}}
\]

Reporting Format:
The Proactive Path Provisioning Time results MUST be reported in the format of a table with a row for each iteration. The last row of the table indicates the Average Proactive Path Provisioning Time.

The report shall capture the following information in addition to the configuration parameters captured in section 5.
- Number of SDN nodes in the path

6.1.6 Reactive Path Provisioning Rate

Objective:
Measure the maximum number of independent paths a controller can concurrently establish between source and destination nodes reactively within the test duration, expressed in paths per second.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller MUST contain the network topology information for the deployed network topology.
2. The controller should have the knowledge about the location of destination addresses for which the paths have to be provisioned. This can be achieved through dynamic learning or static provisioning.
3. Ensure that the default action for flow miss in SDN node is ‘send to controller’.
4. Ensure that each SDN node in a path requires the controller to make the forwarding decision while paving the entire path.
Procedure:
1. Send traffic with unique source and destination addresses from test traffic generator TP1.
2. Record total number of unique traffic frames (Ndf) received at the test traffic generator TP2 within the test duration (Td).

Measurement:

\[
\text{Reactive Path Provisioning Rate Tr}_1 = \frac{Ndf}{T_d}
\]

\[
\text{Average Reactive Path Provisioning Rate} = \frac{Tr_1 + Tr_2 + Tr_3 \ldots + Tr_n}{\text{Total Test Iterations}}
\]

Reporting Format:
The Reactive Path Provisioning Rate results MUST be reported in the format of a table with a row for each iteration. The last row of the table indicates the Average Reactive Path Provisioning Rate.

The report should capture the following information in addition to the configuration parameters captured in section 5.
- Number of SDN nodes in the path
- Offered rate

6.1.7 Proactive Path Provisioning Rate

Objective:
Measure the maximum number of independent paths a controller can concurrently establish between source and destination nodes proactively within the test duration, expressed in paths per second.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller MUST contain the network topology information for the deployed network topology.
2. The controller should have the knowledge about the location of destination addresses for which the paths have to be provisioned. This can be achieved through dynamic learning or static provisioning.
3. Ensure that the default action for flow miss in SDN node is ‘drop’.
Procedure:
1. Send traffic continuously with unique source and destination addresses from test traffic generator TP1.
2. Install corresponding flow entries to reach from simulated sources at the test traffic generator TP1 to the simulated destinations at test traffic generator TP2 through controller’s northbound or management interface.
3. Record total number of unique traffic frames received \( N_{df} \) at the test traffic generator TP2 within the test duration \( T_d \).

Measurement:

\[
\text{Proactive Path Provisioning Rate } Tr_1 = \frac{N_{df}}{T_d}
\]

\[
\text{Average Proactive Path Provisioning Rate} = \frac{Tr_1 + Tr_2 + Tr_3 \ldots Tr_n}{\text{Total Test Iterations}}
\]

Reporting Format:
The Proactive Path Provisioning Rate results MUST be reported in the format of a table with a row for each iteration. The last row of the table indicates the Average Proactive Path Provisioning Rate.

The report should capture the following information in addition to the configuration parameters captured in section 5.
- Number of SDN nodes in the path
- Offered rate

### 6.1.8 Network Topology Change Detection Time

Objective:
Measure the time taken by the controller to detect any changes in the network topology, expressed in milliseconds.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller MUST have discovered the network topology information for the deployed network topology.
2. The periodic network discovery operation should be configured to twice the Test duration \( T_d \) value.
Procedure:
1. Trigger a topology change event by bringing down an active SDN node in the topology.
2. Record the time when the first topology change notification is sent to the controller (Tcn) at the forwarding plane test emulator interface (I1).
3. Stop the test when the controller sends the first topology re-discovery message to the SDN node or the expiry of test interval (Td).
4. Record the time when the first topology re-discovery message is received from the controller (Tcd) at the forwarding plane test emulator interface (I1).

Measurement:
Network Topology Change Detection Time Tr1 = Tcd-Tcn. 
\[ \frac{Tr1 + Tr2 + Tr3 \ldots Trn}{\text{Total Test Iterations}} \]
Average Network Topology Change Detection Time

Reporting Format:
The Network Topology Change Detection Time results MUST be reported in the format of a table with a row for each iteration. The last row of the table indicates the average Network Topology Change Time.

6.2 Scalability

6.2.1 Control Session Capacity

Objective:
Measure the maximum number of control sessions that the controller can maintain.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Procedure:
1. Establish control connection with controller from every SDN nodes emulated in the forwarding plane test emulator.
2. Stop the test when the controller starts dropping the control connection.
3. Record the number of successful connections established with the controller (CCn) at the forwarding plane test emulator.
Measurement:
Control Sessions Capacity = CCn.

Reporting Format:
The Control Session Capacity results MUST be reported in addition to the configuration parameters captured in section 5.

6.2.2 Network Discovery Size

Objective:
Measure the network size (number of nodes, links, and hosts) that a controller can discover.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller MUST support automatic network discovery.
2. Tester should be able to retrieve the discovered topology information either through controller’s management interface or northbound interface.

Procedure:
1. Establish the network connections between controller and network nodes.
2. Query the controller for the discovered network topology information and compare it with the deployed network topology information.
3a. Increase the number of nodes by 1 when the comparison is successful and repeat the test.
3b. Decrease the number of nodes by 1 when the comparison fails and repeat the test.
4. Continue the test until the comparison of step 3b is successful.
5. Record the number of nodes for the last iteration (Ns) where the topology comparison was successful.

Measurement:
Network Discovery Size = Ns.

Reporting Format:
The Network Discovery Size results MUST be reported in addition to the configuration parameters captured in section 5.
6.2.3 Forwarding Table Capacity

Objective:
Measure the maximum number of flow entries a controller can manage in its Forwarding table.

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. The controller Forwarding table should be empty.
2. Flow Idle time MUST be set to higher or infinite value.
3. The controller MUST have completed network topology discovery.
4. Tester should be able to retrieve the forwarding table information either through controller’s management interface or northbound interface.

Procedure:
Reactive Flow Provisioning Mode:
1. Send bi-directional traffic continuously with unique source and/or destination addresses from test traffic generators TP1 and TP2 at the asynchronous message processing rate of controller.
2. Query the controller at a regular interval (e.g., 5 seconds) for the number of learnt flow entries from its northbound interface.
3. Stop the test when the retrieved value is constant for three consecutive iterations and record the value received from the last query (Nrp).

Proactive Flow Provisioning Mode:
1. Install unique flows continuously through controller’s northbound or management interface until a failure response is received from the controller.
2. Record the total number of successful responses (Nrp).

Note:
Some controller designs for proactive flow provisioning mode may require the switch to send flow setup requests in order to generate flow setup responses. In such cases, it is recommended to generate bi-directional traffic for the provisioned flows.
Measurement:
Proactive Flow Provisioning Mode:

Max Flow Entries = Total number of flows provisioned (Nrp)

Reactive Flow Provisioning Mode:

Max Flow Entries = Total number of learnt flow entries (Nrp)
Forwarding Table Capacity = Max Flow Entries.

Reporting Format:
The Forwarding Table Capacity results MUST be tabulated with the following information in addition to the configuration parameters captured in section 5.
- Provisioning Type (Proactive/Reactive)

6.3 Security

6.3.1 Exception Handling

Objective:
Determine the effect of handling error packets and notifications on performance tests. The impact MUST be measured for the following performance tests
a. Path Provisioning Rate
b. Path Provisioning Time
c. Network Topology Change Detection Time

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. This test MUST be performed after obtaining the baseline measurement results for the above performance tests.
2. Ensure that the invalid messages are not dropped by the intermediate devices connecting the controller and SDN nodes.

Procedure:
1. Perform the above listed performance tests and send 1% of messages from the Asynchronous Message Processing Rate as invalid messages from the connected SDN nodes emulated at the forwarding plane test emulator.
2. Perform the above listed performance tests and send 2% of messages from the Asynchronous Message Processing Rate as invalid messages from the connected SDN nodes emulated at the forwarding plane test emulator.

Note:
1. Invalid messages can be frames with incorrect protocol fields or any form of failure notifications sent towards controller.

Measurement:
Measurement MUST be done as per the equation defined in the corresponding performance test measurement section.

Reporting Format:
The Exception Handling results MUST be reported in the format of table with a column for each of the below parameters and row for each of the listed performance tests.
- Without Exceptions
- With 1% Exceptions
- With 2% Exceptions

6.3.2 Denial of Service Handling

Objective:
Determine the effect of handling DoS attacks on performance and scalability tests the impact MUST be measured for the following tests:
- Path Provisioning Rate
- Path Provisioning Time
- Network Topology Change Detection Time
- Network Discovery Size

Reference Test Setup:
The test SHOULD use one of the test setups described in section 3.1 or section 3.2 of this document.

Prerequisite:
This test MUST be performed after obtaining the baseline measurement results for the above tests.
Procedure:
1. Perform the listed tests and launch a DoS attack towards controller while the test is running.

Note:
DoS attacks can be launched on one of the following interfaces.
   a. Northbound (e.g., Sending a huge number of requests on northbound interface)
   b. Management (e.g., Ping requests to controller’s management interface)
   c. Southbound (e.g., TCP SYNC messages on southbound interface)

Measurement:
Measurement MUST be done as per the equation defined in the corresponding test’s measurement section.

Reporting Format:
The DoS Attacks Handling results MUST be reported in the format of table with a column for each of the below parameters and row for each of the listed tests.
- Without any attacks
- With attacks

The report should also specify the nature of attack and the interface.

6.4 Reliability

6.4.1 Controller Failover Time

Objective:
Measure the time taken to switch from an active controller to the backup controller, when the controllers work in redundancy mode and the active controller fails.

Reference Test Setup:
The test SHOULD use the test setup described in section 3.2 of this document.

Prerequisite:
1. Master controller election MUST be completed.
2. Nodes are connected to the controller cluster as per the Redundancy Mode (RM).
3. The controller cluster should have completed the network topology discovery.
4. The SDN Node MUST send all new flows to the controller when it receives from the test traffic generator.
5. Controller should have learnt the location of destination (D1) at test traffic generator TP2.
Procedure:
1. Send uni-directional traffic continuously with incremental sequence number and source addresses from test traffic generator TP1 at the rate that the controller processes without any drops.
2. Ensure that there are no packet drops observed at the test traffic generator TP2.
3. Bring down the active controller.
4. Stop the test when a first frame received on TP2 after failover operation.
5. Record the time at which the last valid frame received (T1) at test traffic generator TP2 before sequence error and the first valid frame received (T2) after the sequence error at TP2.

Measurement:
Controller Failover Time = (T2 - T1)
Packet Loss = Number of missing packet sequences.

Reporting Format:
The Controller Failover Time results MUST be tabulated with the following information.
- Number of cluster nodes
- Redundancy mode
- Controller Failover
- Time Packet Loss
- Cluster keep-alive interval

6.4.2 Network Re-Provisioning Time

Objective:
Compute the time taken to re-route the traffic by the controller when there is a failure in existing traffic paths.

Reference Test Setup:
This test SHOULD use one of the test setup described in section 3.1 or section 3.2 of this document.

Prerequisite:
1. Network with the given number of nodes and redundant paths MUST be deployed.
2. Ensure that the controller MUST have knowledge about the location of test traffic generators TP1 and TP2.
3. Ensure that the controller does not pre-provision the alternate path in the emulated SDN nodes at the forwarding plane test emulator.
Procedure:
1. Send bi-directional traffic continuously with unique sequence number from TP1 and TP2.
2. Bring down a link or switch in the traffic path.
3. Stop the test after receiving first frame after network re-convergence.
4. Record the time of last received frame prior to the frame loss at TP2 (TP2-Tlfr) and the time of first frame received after the frame loss at TP2 (TP2-Tffr).
5. Record the time of last received frame prior to the frame loss at TP1 (TP1-Tlfr) and the time of first frame received after the frame loss at TP1 (TP1-Tffr).

Measurement:
Forward Direction Path Re-Provisioning Time (FDRT) = (TP2-Tffr - TP2-Tlfr)
Reverse Direction Path Re-Provisioning Time (RDRT) = (TP1-Tffr - TP1-Tlfr)
Network Re-Provisioning Time = (FDRT+RDRT)/2
Forward Direction Packet Loss = Number of missing sequence frames at TP1
Reverse Direction Packet Loss = Number of missing sequence frames at TP2

Reporting Format:
The Network Re-Provisioning Time results MUST be tabulated with the following information.
- Number of nodes in the primary path
- Number of nodes in the alternate path
- Network Re-Provisioning Time
- Forward Direction Packet Loss
- Reverse Direction Packet Loss

7. References
7.1 Normative References


7.2 Informative References


8. IANA Considerations

This document does not have any IANA requests.

9. Security Considerations

Benchmarking tests described in this document are limited to the performance characterization of controller in lab environment with isolated network.

10. Appendix A - Example Test Topologies

10.1. Leaf-Spine Topology - Three Tier Network Architecture

---

10.2. Leaf-Spine Topology - Two Tier Network Architecture

11. Appendix A - Benchmarking Methodology using OpenFlow(OF) Controllers

This section gives an overview of OpenFlow protocol and provides test methodology to benchmark SDN controllers supporting OpenFlow southbound protocol.

11.1. Protocol Overview

OpenFlow is an open standard protocol defined by Open Networking Foundation (ONF), used for programming the forwarding plane of network switches or routers via a centralized controller.
11.2. Messages Overview

OpenFlow protocol supports three messages types namely controller-to-switch, asynchronous and symmetric.

Controller-to-switch messages are initiated by the controller and used to directly manage or inspect the state of the switch. These messages allow controllers to query/configure the switch (Features, Configuration messages), collect information from switch (Read-State message), send packets on specified port of switch (Packet-out message), and modify switch forwarding plane and state (Modify-State, Role-Request messages etc.).

Asynchronous messages are generated by the switch without a controller soliciting them. These messages allow switches to update controllers to denote an arrival of new flow (Packet-in), switch state change (Flow-Removed, Port-status) and error (Error).

Symmetric messages are generated in either direction without solicitation. These messages allow switches and controllers to set up connection (Hello), verify for liveness (Echo) and offer additional functionalities (Experimenter).

11.3. Connection Overview

OpenFlow channel is used to exchange OpenFlow message between an OpenFlow switch and an OpenFlow controller. The OpenFlow channel connection can be setup using plain TCP or TLS. By default, a switch establishes single parallel connection with SDN controller. A switch may establish multiple parallel connections to single controller (auxiliary connection) or multiple controllers to handle controller failures and load balancing.
11.4.1 Network Topology Discovery Time

Procedure:

<table>
<thead>
<tr>
<th>SDN Nodes</th>
<th>OpenFlow Controller</th>
<th>SDN Controller</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Deploy network with given no. of OF switches&gt;</td>
<td></td>
<td>&lt;Initialize controller app., NB and SB interfaces&gt;</td>
<td></td>
</tr>
<tr>
<td>OFPT_HELLO Exchange</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PACKET_OUT with LLDP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to all switches (Tm1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PACKET_IN with LLDP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rcvd from switch-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------</td>
<td>----------------</td>
<td>-------------</td>
</tr>
<tr>
<td>PACKET_IN with LLDP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rcvd from switch-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PACKET_IN with LLDP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rcvd from switch-n</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Tmn)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<Wait for the expiry of Test Duration (Td)>

Query the controller for discovered n/w topo. (Di)

<Compare the discovered & offered n/w topology>
Legend:
NB: Northbound
SB: Southbound
OF: OpenFlow
Tm1: Time of reception of first LLDP message from controller
Tmn: Time of last LLDP message sent to controller

Discussion:
The Network Topology Discovery Time can be obtained by calculating the time difference between the first PACKET_OUT with LLDP message received from the controller (Tm1) and the last PACKET_IN with LLDP message sent to the controller (Tmn) when the comparison is successful.

11.4.2 Asynchronous Message Processing Time

Procedure:

<table>
<thead>
<tr>
<th>SDN Nodes</th>
<th>OpenFlow Controller</th>
<th>SDN Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>PACKET_IN with single OFP match header</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(T0) -------------------------->

| PACKET_OUT with single OFP action header |

(R0) <---------------------------|

| . | . |

| PACKET_IN with single OFP match header |

(Tn) -------------------------->

| PACKET_OUT with single OFP action header |

(Rn) <---------------------------|

<Wait for the expiry of Test Duration>

<Record the number of PACKET_INs/PACKET_OUTs Exchanged (Nrx)>
Legend:
T0,T1, ..Tn are PACKET_IN messages transmit timestamps.
R0,R1, ..Rn are PACKET_OUT messages receive timestamps.
Nrx : Number of successful PACKET_IN/PACKET_OUT message exchanges

Discussion:
The Asynchronous Message Processing Time will be obtained by
sum of \([(R0-T0),(R1-T1)..(Rn - Tn)]/ Nrx.

11.4.3 Asynchronous Message Processing Rate

Procedure:

<table>
<thead>
<tr>
<th>SDN Nodes</th>
<th>OpenFlow Controller</th>
<th>SDN Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>PACKET_IN with multiple OFP match headers</td>
<td>-----------------------------&gt;</td>
<td></td>
</tr>
<tr>
<td>PACKET_OUT with multiple OFP action headers</td>
<td>&lt;-----------------------------</td>
<td></td>
</tr>
<tr>
<td>PACKET_IN with multiple OFP match headers</td>
<td>-----------------------------&gt;</td>
<td></td>
</tr>
<tr>
<td>PACKET_OUT with multiple OFP action headers</td>
<td>&lt;-----------------------------</td>
<td></td>
</tr>
<tr>
<td>PACKET_IN with multiple OFP match headers</td>
<td>-----------------------------&gt;</td>
<td></td>
</tr>
<tr>
<td>PACKET_OUT with multiple OFP action headers</td>
<td>&lt;-----------------------------</td>
<td></td>
</tr>
<tr>
<td>&lt;Wait for the expiry of Test Duration&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;Record the number of OFP (Nrx) action headers rcvd&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion:
The Asynchronous Message Processing Rate will be obtained by calculating the number of OFP action headers received in all PACKET_OUT messages during the test duration.

11.4.4 Reactive Path Provisioning Time

Procedure:

<table>
<thead>
<tr>
<th>Test Traffic</th>
<th>Test Traffic</th>
<th>SDN Nodes</th>
<th>OpenFlow Controller</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generator TP1</td>
<td>G-ARP (D1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Tsf1)</td>
<td>----------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic (S1,D1)</td>
<td>PACKET_IN(S1,D1)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| (Tdf1) | | | | }

Legend:
G-ARP: Gratuitous ARP message.
Tsf1: Time of first frame sent from TP1
Tdf1: Time of first frame received from TP2

Discussion:
The Reactive Path Provisioning Time can be obtained by finding the time difference between the transmit and receive time of the traffic (Tsf1-Tdf1).
11.4.5 Proactive Path Provisioning Time

Procedure:

<table>
<thead>
<tr>
<th>Test Traffic Generator TP1</th>
<th>Test Traffic Generator TP2</th>
<th>SDN Nodes</th>
<th>OpenFlow Controller</th>
<th>SDN Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-ARP (D1)</td>
<td>PACKET_IN(D1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic (S1,D1)</td>
<td>&lt;Install flow for S1,D1&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Tsf1)</td>
<td>FLOW_MOD(D1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Traffic (S1,D1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Tdf1)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
G-ARP: Gratuitous ARP message.
Tsf1: Time of first frame sent from TP1
Tdf1: Time of first frame received from TP2

Discussion:
The Proactive Path Provisioning Time can be obtained by finding the time difference between the transmit and receive time of the traffic (Tsf1-Tdf1).
11.4.6 Reactive Path Provisioning Rate

Procedure:

<table>
<thead>
<tr>
<th>Test Traffic Generator TP1</th>
<th>Test Traffic Generator TP2</th>
<th>SDN Nodes</th>
<th>OpenFlow Controller</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>G-ARP (D1..Dn)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PACKET_IN(D1..Dn)</td>
<td>--------------------</td>
</tr>
</tbody>
</table>
| Traffic (S1..Sn,D1..Dn)   |                           | PACKET_IN(S1.Sn,D1.Dn) | --
|                            |                           | FLOW_MOD(S1) | <---------------------|
|                            |                           | FLOW_MOD(D1) | <---------------------|
|                            |                           | FLOW_MOD(S2) | <---------------------|
|                            |                           | FLOW_MOD(S2) | <---------------------|
|                            |                           | FLOW_MOD(D2) | <---------------------|
|                            |                           | .            | <---------------------|
|                            |                           | .            | <---------------------|
|                            |                           | FLOW_MOD(Sn) | <---------------------|
|                            |                           | FLOW_MOD(Dn) | <---------------------|
|                            |                           | Traffic (S1..Sn,D1..Dn) | <-------------------|
|                            |                           |               |                     |
Legend:
G-ARP: Gratuitous ARP
D1..Dn: Destination Endpoint 1, Destination Endpoint 2 .... Destination Endpoint n
S1..Sn: Source Endpoint 1, Source Endpoint 2 .., Source Endpoint n

Discussion:
The Reactive Path Provisioning Rate can be obtained by finding the total number of frames received at TP2 after the test duration.

11.4.7 Proactive Path Provisioning Rate

Procedure:
| Traffic (S1.Sn, D1.Dn) (Tdf1) | <--------------------- |

Legend:
- G-ARP: Gratuitous ARP
- D1..Dn: Destination Endpoint 1, Destination Endpoint 2 .... Destination Endpoint n
- S1..Sn: Source Endpoint 1, Source Endpoint 2 .., Source Endpoint n

Discussion:
The Proactive Path Provisioning Rate can be obtained by finding the total number of frames received at TP2 after the test duration.

11.4.8 Network Topology Change Detection Time

Procedure:

<table>
<thead>
<tr>
<th>SDN Nodes</th>
<th>OpenFlow Controller</th>
<th>SDN Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0 PORT_STATUS with link down from S1</td>
<td>&lt;Bring down a link in switch S1&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>--------------------------&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>First PACKET_OUT with LLDP to OF Switch</td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>&lt;---------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;Record time of 1st PACKET_OUT with LLDP T1&gt;</td>
<td></td>
</tr>
</tbody>
</table>

Discussion:
The Network Topology Change Detection Time can be obtained by finding the difference between the time the OpenFlow switch S1 sends the PORT_STATUS message (T0) and the time that the OpenFlow controller sends the first topology re-discovery message (T1) to OpenFlow switches.
11.5 Scalability

11.5.1 Control Sessions Capacity

Procedure:

SDN Nodes                              OpenFlow Controller
|                                       |
|    OFPT_HELLO Exchange for Switch 1   |
|<------------------------------------->|
|                                       |
|    OFPT_HELLO Exchange for Switch 2   |
|<------------------------------------->|
|                                       |
|                  .                    |
|                  .                    |
|                  .                    |
|    OFPT_HELLO Exchange for Switch n   |
|X<----------------------------------->X|

Discussion:
The value of Switch n-1 will provide Control Sessions Capacity.

11.5.2 Network Discovery Size

Procedure:

SDN Nodes                     OpenFlow Controller                     SDN Application
|                            |                           |<Deploy network with given no. of OF switches N>|
|                            |                           |
|    OFPT_HELLO Exchange     |                           |
|<-------------------------->|                           |
|                            |                           |
|    PACKET_OUT with LLDP    |                           |
|      to all switches       |                           |
|<---------------------------|                           |
|                            |                           |
|         PACKET_IN with LLDP|                           |
|          rcvd from switch-1|                           |
|--------------------------->|                           |
PACKET_IN with LLDP
rcvd from switch-2
-------------------------->
.                          
.                          
PACKET_IN with LLDP
rcvd from switch-n
-------------------------->

<Wait for the expiry
of Test Duration (Td)>

Query the controller for
discovered n/w topo.(N1)
--------------------------

<If N1==N repeat Step 1
with N+1 nodes until N1<N >

<If N1<N repeat Step 1
with N=N1 nodes once and
exit>

Legend:
 n/w topo: Network Topology
OF: OpenFlow

Discussion:
The value of N1 provides the Network Discovery Size value. The
test duration can be set to the stipulated time within which the
user expects the controller to complete the discovery process.

11.5.3 Forwarding Table Capacity

Procedure:

<table>
<thead>
<tr>
<th>Test Traffic</th>
<th>SDN Nodes</th>
<th>OpenFlow</th>
<th>SDN Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generator TP1</td>
<td>Controller</td>
<td>Application</td>
<td></td>
</tr>
</tbody>
</table>

Step 1

G-ARP (H1..Hn)
-------------------------->

PACKET_IN(D1..Dn)
-------------------------->
| Step 2 | <Wait for 5 secs> |<Query for FWD entry> (F1) |
|        | <Wait for 5 secs> |<Query for FWD entry> (F2) |
|        | <Query for FWD entry> (F3) |
|        | <Repeat Step 2 until F1==F2==F3> |

Legend:
- G-ARP: Gratuitous ARP
- H1..Hn: Host 1 .. Host n
- FWD: Forwarding Table

Discussion:
Query the controller forwarding table entries for multiple times until the three consecutive queries return the same value. The last value retrieved from the controller will provide the Forwarding Table Capacity value. The query interval is user configurable. The 5 seconds shown in this example is for representational purpose.
11.6 Security

11.6.1 Exception Handling

Procedure:

<table>
<thead>
<tr>
<th>Test Traffic Generator TP1</th>
<th>Test Traffic Generator TP2</th>
<th>SDN Nodes</th>
<th>OpenFlow SDN Controller</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-ARP (D1..Dn)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>-------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>PACKET_IN(D1..Dn)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1 Traffic (S1..Sn,D1..Dn)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>-------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>PACKET_IN(S1..Sa, D1..Da)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PACKET_IN(Sa+1..Sn,Da+1..Dn)</td>
<td></td>
<td></td>
<td>1% incorrect OFP Match header</td>
<td></td>
</tr>
<tr>
<td>FLOW_MOD(D1..Dn)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLOW_MOD(S1..Sa)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic (S1..Sa, D1..Da)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<Wait for Test Duration>

<Record Rx frames at TP2 (Rn1)>
Legend:
G-ARP: Gratuitous ARP
PACKET_IN(Sa+1..Sn,Da+1..Dn): OpenFlow PACKET_IN with wrong version number
Rn1: Total number of frames received at Test Port 2 with 1% incorrect frames
Rn2: Total number of frames received at Test Port 2 with 2% incorrect frames

Discussion:
The traffic rate sent towards OpenFlow switch from Test Port 1 should be 1% higher than the Path Programming Rate. Rn1 will provide the Path Provisioning Rate of controller at 1% of incorrect frames handling and Rn2 will provide the Path Provisioning Rate of controller at 2% of incorrect frames handling.

The procedure defined above provides test steps to determine the effect of handling error packets on Path Programming Rate. Same procedure can be adopted to determine the effects on other performance tests listed in this benchmarking tests.

11.6.2 Denial of Service Handling

Procedure:

Test Traffic Test Traffic SDN Nodes OpenFlow SDN
Generator TP1 Generator TP2 Controller Application

<table>
<thead>
<tr>
<th></th>
<th>G-ARP (D1..Dn)</th>
<th>PACKET_IN(D1..Dn)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-----------------------------------</td>
<td></td>
</tr>
</tbody>
</table>

Traffic (S1..Sn,D1..Dn)  
----------------------------->  
Packet_IN(S1..Sn, D1..Dn) 
---------------------------->

TCP SYN Attack from a switch
---------------------------->

FLOW_MOD(D1..Dn)
-------------------

FLOW_MOD(S1..Sn)
OFP headers <----------------

Traffic (S1..Sn, D1..Dn) 
<------------------------

<Wait for Test Duration>

<Record Rx frames at TP2 (Rn1)>

Legend:
G-ARP: Gratuitous ARP

Discussion:
TCP SYN attack should be launched from one of the emulated/simulated OpenFlow Switch. Rn1 provides the Path Programming Rate of controller upon handling denial of service attack.

The procedure defined above provides test steps to determine the effect of handling denial of service on Path Programming Rate. Same procedure can be adopted to determine the effects on other performance tests listed in this benchmarking tests.
11.7 Reliability
11.7.1 Controller Failover Time

Procedure:

<table>
<thead>
<tr>
<th>Test Traffic</th>
<th>Test Traffic</th>
<th>SDN Nodes</th>
<th>OpenFlow</th>
<th>SDN Controller</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generator TP1</td>
<td>Generator TP2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Traffic (S1..Sn,D1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G-ARP (D1)</td>
</tr>
<tr>
<td></td>
<td>-----------</td>
</tr>
<tr>
<td></td>
<td>PACKET_IN(D1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Traffic (S1,D1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;------------</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Traffic (S2,D1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;------------</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Traffic (Sn-1,D1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;------------</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Traffic (Sn,D1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;------------</td>
</tr>
</tbody>
</table>

<Bring down the active controller>

FLOW_MOD(Sn-1) <-X----------
Legend:
G-ARP: Gratuitous ARP.

Discussion:
The time difference between the last valid frame received before the traffic loss and the first frame received after the traffic loss will provide the controller failover time.

If there is no frame loss during controller failover time, the controller failover time can be deemed negligible.

11.7.2 Network Re-Provisioning Time

Procedure:

<table>
<thead>
<tr>
<th>Test Traffic Generator TP1</th>
<th>Test Traffic Generator TP2</th>
<th>SDN Nodes</th>
<th>OpenFlow Controller</th>
<th>SDN Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-ARP (D1)</td>
<td></td>
<td></td>
<td>PACKET_IN(D1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>G-ARP (S1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PACKET_IN(S1)</td>
<td></td>
</tr>
<tr>
<td>Traffic (S1,D1,Seq.no (1..n))</td>
<td></td>
<td></td>
<td>PACKET_IN(S1,D1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Traffic (D1,S1, Seq.no (1..n))</td>
<td></td>
</tr>
<tr>
<td>Traffic (S1,D1, Seq.no(1))</td>
<td>&lt;------------------</td>
<td></td>
<td>PORT_STATUS(Sa)</td>
<td>&lt;------------------</td>
</tr>
<tr>
<td>Traffic (S1,D1, Seq.no(2))</td>
<td>&lt;------------------</td>
<td></td>
<td>PORT_STATUS(Sa)</td>
<td>&lt;------------------</td>
</tr>
<tr>
<td>Traffic (D1,S1,Seq.no(x))</td>
<td>&lt;------------------</td>
<td></td>
<td>PORT_STATUS(Sa)</td>
<td>&lt;------------------</td>
</tr>
<tr>
<td>Traffic (S1,D1, Seq.no(x))</td>
<td>&lt;------------------</td>
<td></td>
<td>PORT_STATUS(Sa)</td>
<td>&lt;------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Legend:
G-ARP: Gratuitous ARP message.
Seq.no: Sequence number.
Sa: Neighbour switch of the switch that was brought down.

Discussion:
The time difference between the last valid frame received before the traffic loss (Packet number with sequence number x) and the first frame received after the traffic loss (packet with sequence number n) will provide the network path re-provisioning time.

Note that the test is valid only when the controller provisions the alternate path upon network failure.

12. Acknowledgements

The authors would like to thank the following individuals for providing their valuable comments to the earlier versions of this document: Al Morton (AT&T), Sandeep Gangadharan (HP), M. Georgescu (NAIST), Andrew McGregor (Google), Scott Bradner (Harvard University), Jay Karthik (Cisco), Ramakrishnan (Brocade).
13. Authors’ Addresses

Bhuvaneswaran Vengainathan
Veryx Technologies Inc.
1 International Plaza, Suite 550
Philadelphia
PA 19113
Email: bhuvaneswaran.vengainathan@veryxtech.com

Anton Basil
Veryx Technologies Inc.
1 International Plaza, Suite 550
Philadelphia
PA 19113
Email: anton.basil@veryxtech.com

Mark Tassinari
Hewlett-Packard,
8000 Foothills Blvd,
Roseville, CA 95747
Email: mark.tassinari@hp.com

Vishwas Manral
Ionos Corp,
4100 Moorpark Ave,
San Jose, CA
Email: vishwas@ionosnetworks.com

Sarah Banks
VSS Monitoring
930 De Guigne Drive,
Sunnyvale, CA
Email: sbanks@encrypted.net
Abstract

This document defines terminology for benchmarking an SDN Controller’s performance. The terms provided in this document help to benchmark SDN controller’s performance independent of the controller’s supported protocols and/or network services. A mechanism for benchmarking the performance of SDN controllers is defined in the companion methodology document. These two documents provide a standard mechanism to measure and evaluate the performance of various controller implementations.

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 http://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 March 22, 2016.
Copyright Notice

Copyright (c) 2015 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. 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 ................................................ 3
2. Term Definitions ............................................ 4
  2.1. SDN Terms ............................................... 4
    2.1.1. SDN Node ......................................... 4
    2.1.2. SDN Application ................................. 4
    2.1.3. Flow ............................................. 4
    2.1.4. Northbound Interface ......................... 4
    2.1.5. Southbound Interface ......................... 5
    2.1.6. Controller Forwarding Table .................. 5
    2.1.7. Proactive Flow Provisioning Mode .......... 6
    2.1.8. Reactive Flow Provisioning Mode ........... 6
    2.1.9. Path ........................................... 7
    2.1.10. Standalone Mode .................................. 7
    2.1.11. Cluster/Redundancy Mode ................... 7
    2.1.12. Asynchronous Message ......................... 8
    2.1.13. Test Traffic Generator ....................... 8
  2.2. Test Configuration/Setup Terms ....................... 9
    2.2.1. Number of SDN Nodes ......................... 9
    2.2.2. Test Iterations ............................... 9
    2.2.3. Test Duration .................................. 9
    2.2.4. Number of Cluster nodes .................... 10
  2.3. Benchmarking Terms ................................ 10
    2.3.1. Performance .................................... 10
      2.3.1.1. Network Topology Discovery Time ........ 10
      2.3.1.2. Asynchronous Message Processing Time .... 11
      2.3.1.3. Asynchronous Message Processing Rate .... 11
      2.3.1.4. Reactive Path Provisioning Time ......... 12
      2.3.1.5. Proactive Path Provisioning Time ...... 12
1. Introduction

Software Defined Networking (SDN) is a networking architecture in which network control is decoupled from the underlying forwarding function and is placed in a centralized location called the SDN controller. The SDN controller abstracts the underlying network and offers a global view of the overall network to applications and business logic. Thus, an SDN controller provides the flexibility to program, control, and manage network behaviour dynamically through standard interfaces. Since the network controls are logically centralized, the need to benchmark the SDN controller performance becomes significant. This document defines terms to benchmark various controller designs for performance, scalability, reliability and security, independent of northbound and southbound protocols.

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.
2. Term Definitions

2.1. SDN Terms

2.1.1. SDN Node

Definition: An SDN node is an emulated/simulated entity that forwards data in a software defined environment.

Discussion: An SDN node can be an emulated/simulated switch, router, gateway, or any network service appliance that supports standardized or proprietary programmable interface.

Measurement Units: N/A

See Also: None

2.1.2. SDN Application

Definition: Any business logic that alter the network behaviour dynamically through controller’s northbound interface.

Discussion: SDN application can be any business application, cloud orchestration system, network services orchestration etc.,

Measurement Units: N/A

See Also: None

2.1.3. Flow

Definition: A flow is a uni-directional sequence of packets having common properties derived from the data contained in the packet.

Discussion: A flow can be set of packets having same source address, destination address, source port and destination port, or any of these combinations.
2.1.4. Northbound Interface

Definition:
The northbound interface is the application programming interface provided by the SDN controller for the SDN services and applications to interact with the SDN controller.

Discussion:
The northbound interface allows SDN applications and orchestration systems to program and retrieve the network information through the SDN controller.

2.1.5. Southbound Interface

Definition:
The southbound interface is the application programming interface provided by the SDN controller to interact with the SDN nodes.

Discussion:
Southbound interface enables controller to interact with the SDN nodes in the infrastructure for dynamically defining the traffic forwarding behaviour.

2.1.6. Controller Forwarding Table

Definition:
A controller forwarding table contains flow entries learned in one of two ways: first, entries could be learned from traffic received through the data plane, or, second, these entries could be...
statically provisioned on the controller, and distributed to devices via the southbound interface.

Discussion:
The controller forwarding table has an aging mechanism which will be applied only for dynamically learnt entries.

Measurement Units:
N/A

See Also:
None

2.1.7. Proactive Flow Provisioning Mode

Definition:
Controller programming flows in SDN nodes based on the flow entries provisioned through controller’s northbound interface.

Discussion:
Orchestration systems and SDN applications can define the network forwarding behaviour by programming the controller using proactive flow provisioning. The controller can then program the SDN nodes with the pre-provisioned entries.

Measurement Units:
N/A

See Also:
None

2.1.8. Reactive Flow Provisioning Mode

Definition:
Controller programming flows in SDN nodes based on the traffic received from SDN nodes through controller’s southbound interface

Discussion:
The SDN controller dynamically decides the forwarding behaviour based on the incoming traffic from the SDN nodes. The controller then programs the SDN nodes using Reactive Flow Provisioning.

Measurement Units:
N/A

See Also:
None
2.1.9. Path

Definition:
A path is a sequence of SDN nodes and links traversed by a flow.

Discussion:
As defined in RFC 2330, path is a sequence of the form < h0, l1, h1, ...
...ln, hn >, where n >=0, h0 and hn is a Host, h1...hn-1 is an SDN
Node, each li is a link between hi-1 and hi. A pair <li, hi> is
termed a ‘hop’. Note that path is a unidirectional concept.

Measurement Units:
N/A

See Also:
None

2.1.10. Standalone Mode

Definition:
Single controller handling all control plane functionalities without
redundancy, or the ability to provide high availability and/or
automatic failover.

Discussion:
In standalone mode, one controller manages one or more network
domains.

Measurement Units:
N/A

See Also:
None

2.1.11. Cluster/Redundancy Mode

Definition:
A group of 2 or more controllers handling all control plane
functionalities.

Discussion:
In cluster mode, multiple controllers are teamed together for the
purpose of load sharing and/or high availability. The controllers in
the group may work in active/standby (master/slave) or active/active (equal) mode depending on the intended purpose.

Measurement Units:
N/A

See Also:
None

2.1.12. Asynchronous Message

Definition:
Any message from the SDN node that is generated for network events.

Discussion:
Control messages like flow setup request and response message is classified as asynchronous message. The controller has to return a response message. Note that the SDN node will not be in blocking mode and continues to send/receive other control messages.

Measurement Units:
N/A

See Also:
None

2.1.13. Test Traffic Generator

Definition:
Test Traffic Generator is an entity that generates/receives network traffic.

Discussion:
Test Traffic Generator can be an entity that interfaces with SDN Nodes to send/receive real-time network traffic.

Measurement Units:
N/A

See Also:
None
2.2. Test Configuration/Setup Terms

2.2.1. Number of SDN Nodes

Definition:
The number of SDN nodes present in the defined test topology.

Discussion:
The SDN nodes defined in the test topology can be deployed using real hardware or emulated in hardware platforms.

Measurement Units:
N/A

See Also:
None

2.2.2. Test Iterations

Definition:
The number of times the test needs to be repeated.

Discussion:
The test needs to be repeated for multiple iterations to obtain a reliable metric. It is recommended that this test SHOULD be performed for at least 10 iterations to increase the confidence in measured result.

Measurement Units:
N/A

See Also:
None

2.2.3. Test Duration

Definition:
Defines the duration of test trails for each iteration.

Discussion:
Test duration forms the basis for stop criteria for benchmarking tests. Test not completed within this time interval is considered as incomplete.
2.2.4. Number of Cluster nodes

**Definition:**
Defines the number of controllers present in the controller cluster.

**Discussion:**
This parameter is relevant when testing the controller performance in clustering/teaming mode. The number of nodes in the cluster MUST be greater than 1.

**Measurement Units:**
N/A

**See Also:**
None

2.3. Benchmarking Terms

This section defines metrics for benchmarking the SDN controller. The procedure to perform the defined metrics is defined in the accompanying methodology document.

2.3.1. Performance

2.3.1.1. Network Topology Discovery Time

**Definition:**
To measure the time taken to discover the network topology - nodes and links by a controller.

**Discussion:**
Network topology discovery is key for the SDN controller to provision and manage the network. So it is important to measure how quickly the controller discovers the topology to learn the current network state. This benchmark is obtained by presenting a network topology (Tree, Mesh or Linear) with the given number of nodes to the controller and wait for the discovery process to complete. It is expected that the controller supports network discovery mechanism and uses protocol messages for its discovery process.
2.3.1.2. Asynchronous Message Processing Time

Definition:
To measure the time taken by the controller to process an asynchronous message.

Discussion:
For SDN to support dynamic network provisioning, it is important to measure how quickly the controller responds to an event triggered from the network. The event could be any notification messages generated by an SDN node upon arrival of a new flow, link down etc. This benchmark is obtained by sending asynchronous messages from every connected SDN nodes one at a time for the defined test duration. This test assumes that the controller will respond to the received asynchronous message.

Measurement Units:
milliseconds

See Also:
None

2.3.1.3. Asynchronous Message Processing Rate

Definition:
To measure the maximum number of asynchronous messages that a controller can process within the test duration.

Discussion:
As SDN assures flexible network and agile provisioning, it is important to measure how many network events that the controller can handle at a time. This benchmark is obtained by sending asynchronous messages from every connected SDN nodes at full connection capacity for the given test duration. This test assumes that the controller will respond to all the received asynchronous messages.

Measurement Units:
Messages processed per second.
2.3.1.4. Reactive Path Provisioning Time

Definition:
The time taken by the controller to setup a path reactively between source and destination node, expressed in milliseconds.

Discussion:
As SDN supports agile provisioning, it is important to measure how fast that the controller provisions an end-to-end flow in the dataplane. The benchmark is obtained by sending traffic from a source endpoint to the destination endpoint, finding the time difference between the first and the last flow provisioning message exchanged between the controller and the SDN nodes for the traffic path.

Measurement Units:
milliseconds.

See Also:
None

2.3.1.5. Proactive Path Provisioning Time

Definition:
The time taken by the controller to setup a path proactively between source and destination node, expressed in milliseconds.

Discussion:
For SDN to support pre-provisioning of traffic path from application, it is important to measure how fast that the controller provisions an end-to-end flow in the dataplane. The benchmark is obtained by provisioning a flow on controller’s northbound interface for the traffic to reach from a source to a destination endpoint, finding the time difference between the first and the last flow provisioning message exchanged between the controller and the SDN nodes for the traffic path.

Measurement Units:
milliseconds.

See Also:
None
2.3.1.6. Reactive Path Provisioning Rate

Definition:
Measure the maximum number of independent paths a controller can concurrently establish between source and destination nodes reactively within the test duration, expressed in paths per second.

Discussion:
For SDN to support agile traffic forwarding, it is important to measure how many end-to-end flows that the controller could setup in the dataplane. This benchmark is obtained by sending traffic each with unique source and destination pairs from the source SDN node and determine the number of frames received at the destination SDN node.

Measurement Units:
Paths provisioned per second.

See Also:
None

2.3.1.7. Proactive Path Provisioning Rate

Definition:
Measure the maximum number of independent paths a controller can concurrently establish between source and destination nodes proactively within the test duration, expressed in paths per second.

Discussion:
For SDN to support pre-provisioning of traffic path for a larger network from the application, it is important to measure how many end-to-end flows that the controller could setup in the dataplane. This benchmark is obtained by sending traffic each with unique source and destination pairs from the source SDN node. Program the flows on controller’s northbound interface for traffic to reach from each of the unique source and destination pairs and determine the number of frames received at the destination SDN node.

Measurement Units:
Paths provisioned per second.

See Also:
None

2.3.1.8. Network Topology Change Detection Time

Definition:
The amount of time required for the controller to detect any changes in the network topology.

Discussion:
In order for the controller to support fast network failure recovery, it is critical to measure how fast the controller is able to detect any network-state change events. This benchmark is obtained by triggering a topology change event and measuring the time controller takes to detect and initiate a topology re-discovery process.

Measurement Units:
milliseconds

See Also:
None

2.3.2. Scalability

2.3.2.1. Control Sessions Capacity

Definition:
To measure the maximum number of control sessions the controller can maintain.

Discussion:
Measuring the controller’s control sessions capacity is important to determine the controller’s system and bandwidth resource requirements. This benchmark is obtained by establishing control session with the controller from each of the SDN node until it fails. The number of sessions that were successfully established will provide the Control Sessions Capacity.

Measurement Units:
N/A

See Also:
None

2.3.2.2. Network Discovery Size

Definition:
To measure the network size (number of nodes, links and hosts) that a controller can discover.

Discussion:
For optimal network planning, it is key to measure the maximum network size that the controller can discover. This benchmark is obtained by presenting an initial set of SDN nodes for discovery to the controller. Based on the initial discovery, the number of SDN nodes is increased or decreased to determine the maximum nodes that the controller can discover.

Measurement Units:
N/A

See Also:
None

2.3.2.3. Forwarding Table Capacity

Definition:
The maximum number of flow entries that a controller can manage in its Forwarding table.

Discussion:
It is significant to measure the capacity of controller’s Forwarding Table to determine the number of flows that controller could forward without flooding/dropping. This benchmark is obtained by continuously presenting the controller with new flow entries through reactive or proactive flow provisioning mode until the forwarding table becomes full. The maximum number of nodes that the controller can hold in its Forwarding Table will provide Forwarding Table Capacity.

Measurement Units:
Maximum number of flow entries managed.

See Also:
None

2.3.3. Security

2.3.3.1. Exception Handling

Definition:
To determine the effect of handling error packets and notifications on performance tests.

Discussion:
This benchmark test is to be performed after obtaining the baseline performance of the performance tests defined in Section 2.3.1. This
benchmark determines the deviation from the baseline performance due to the handling of error or failure messages from the connected SDN nodes.

Measurement Units:
N/A

See Also:
None

2.3.3.2. Denial of Service Handling

Definition:
To determine the effect of handling denial of service (DoS) attacks on performance and scalability tests.

Discussion:
This benchmark test is to be performed after obtaining the baseline performance of the performance and scalability tests defined in section 2.3.1 and section 2.3.1. This benchmark determines the deviation from the baseline performance due to the handling of denial of service attacks on controller.

Measurement Units:
Deviation of baseline metrics while handling Denial of Service Attacks.

See Also:
None

2.3.4. Reliability

2.3.4.1. Controller Failover Time

Definition:
The time taken to switch from an active controller to the backup controller, when the controllers work in redundancy mode and the active controller fails.

Discussion:
This benchmark determine the impact of provisioning new flows when controllers are teamed and the active controller fails.

Measurement Units:
milliseconds.
2.3.4.2. Network Re-Provisioning Time

Definition:
The time taken to re-route the traffic by the Controller, when there is a failure in existing traffic paths.

Discussion:
This benchmark determines the controller’s re-provisioning ability upon network failures. This benchmark test assumes the following:
   i. Network topology supports redundant path between source and destination endpoints.
   ii. Controller does not pre-provision the redundant path.

Measurement Units:
milliseconds.

3. Test Coverage

<table>
<thead>
<tr>
<th>Setup</th>
<th>Speed</th>
<th>Scalability</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Network Topology Discovery</td>
<td>1. Network Discovery Size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Reactive Path Provisioning</td>
<td>Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Proactive Path Provisioning</td>
<td>Time</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Reactive Path Provisioning</td>
<td>Rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Proactive Path Provisioning</td>
<td>Rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>Tear Down</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>----------------------------------</td>
<td>-------------------------</td>
<td></td>
</tr>
<tr>
<td>2. Asynchronous Message Processing Time</td>
<td>2. Forwarding Table Capacity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Control Sessions Capacity</td>
<td>2. Exception Handling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Network Topology Change Detection Time</td>
<td>3. Denial of Service Handling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Network Re-Provisioning Time</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. References

4.1. Normative References


4.2. Informative References

http://opencontrail.org/opencontrail-architecture-documentation

[OpenDaylight] OpenDaylight Controller: Architectural Framework,
https://wiki.opendaylight.org/view/OpenDaylight_Controller

5. IANA Considerations

This document does not have any IANA requests.

6. Security Considerations

Security issues are not discussed in this memo.

7. Acknowledgements

The authors would like to acknowledge Sandeep Gangadharan (HP) for the significant contributions to the earlier versions of this document. The authors would like to thank the following individuals for providing their valuable comments to the earlier versions of this document: Al Morton (AT&T), M. Georgescu (NAIST), Andrew McGregor (Google), Scott Bradner (Harvard University), Jay Karthik (Cisco), Ramakrishnan (Brocade), Khasanov Boris (Huawei).

8. Authors’ Addresses

Bhuvaneswaran Vengainathan
Veryx Technologies Inc.
1 International Plaza, Suite 550
Philadelphia
PA 19113

Email: bhuvaneswaran.vengainathan@veryxtech.com

Anton Basil
Veryx Technologies Inc.
1 International Plaza, Suite 550
Abstract

Software Defined Networking proposes the separation of the control plane from the data plane in the network nodes and its logical centralization on a control entity. Most of the network intelligence is moved to this functional entity. Typically, such entity is seen as a compendium of interacting control functions in a vertical, tight integrated fashion. The relocation of the control functions from a number of distributed network nodes to a logical central entity conceptually places together a number of control capabilities with different purposes. As a consequence, the existing solutions do not provide a clear separation between transport control and services that rely upon transport capabilities.

This document describes a proposal named Cooperating Layered Architecture for SDN. The idea behind that is to differentiate the control functions associated to transport from those related to services, in such a way that they can be provided and maintained independently, and can follow their own evolution path.

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 http://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 April 21, 2016.

Copyright Notice

Copyright (c) 2015 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. 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 ........................................ 3
2. Terminology ...................................... 4
3. Architecture overview .............................. 5
   3.1. Functional strata ............................... 8
       3.1.1. Transport stratum ........................... 8
       3.1.2. Service stratum ............................. 9
       3.1.3. Recursiveness ............................... 9
   3.2. Plane separation ............................... 9
       3.2.1. Control Plane ............................... 9
       3.2.2. Management Plane ......................... 10
       3.2.3. Resource Plane ............................. 10
4. Required features .................................. 10
5. Communication between SDN Controllers ......... 11
6. Deployment scenarios .............................. 11
   6.1. Full SDN environments ........................ 11
       6.1.1. Multiple Service strata associated to a single
              Transport stratum .......................... 11
       6.1.2. Single service stratum associated to multiple
              Transport strata ............................ 12
   6.2. Hybrid environments .......................... 12
       6.2.1. SDN Service stratum associated to a legacy Transport
              stratum ..................................... 12
       6.2.2. Legacy Service stratum associated to an SDN Transport
              stratum ..................................... 12
   6.3. Multi-domain scenarios in Transport Stratum .... 12
7. Use cases ......................................... 13
1. Introduction

Software Defined Networking (SDN) proposes the separation of the control plane from the data plane in the network nodes and its logical centralization on a control entity. A programmatic interface is defined between such entity and the network nodes, which functionality is supposed to perform traffic forwarding (only). Through that interface, the control entity instructs the nodes involved in the forwarding plane and modifies their traffic forwarding behavior accordingly.

Most of the intelligence is moved to such functional entity. Typically, such entity is seen as a compendium of interacting control functions in a vertical, tight integrated fashion.

This approach presents a number of issues:

- Unclear responsibilities between actors involved in a service provision and delivery.
- Complex reuse of functions for the provision of services.
- Closed, monolithic control architectures.
- Difficult interoperability and interchangeability of functional components.
- Blurred business boundaries among providers.
- Complex service/network diagnosis and troubleshooting, particularly to determine which segment is responsible for a failure.

The relocation of the control functions from a number of distributed network nodes to another entity conceptually places together a number of control capabilities with different purposes. As a consequence, the existing solutions do not provide a clear separation between services and transport control.
This document describes a proposal named Cooperating Layered Architecture for SDN (CLAS). The idea behind that is to differentiate the control functions associated to transport from those related to services, in such a way that they can be provided and maintained independently, and can follow their own evolution path.

Despite such differentiation it is required a close cooperation between service and transport layers and associated components to provide an efficient usage of the resources.

2. Terminology

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 RFC2119 [RFC2119].

This document makes use of the following terms:

- **Transport**: denotes the transfer capabilities offered by a networking infrastructure. The transfer capabilities can rely upon pure IP techniques, or other means such as MPLS or optics.

- **Service**: denote a logical construct that make use of transport capabilities. This document does not make any assumption on the functional perimeter of a service that can be built above a transport infrastructure. As such, a service can be an offering that is offered to customers or be invoked for the delivery of another (added-value) service.

- **SDN intelligence**: refers to the decision-making process that is hosted by a node or a set of nodes. The intelligence can be centralized or distributed. Both schemes are within the scope of this document. The SDN intelligence relies on inputs form various functional blocks such as: network topology discovery, service topology discovery, resource allocation, business guidelines, customer profiles, service profiles, etc. The exact decomposition of an SDN intelligence, apart from the layering discussed in this document, is out of scope.

Additionally, the following acronyms are used in this document.

- **CLAS**: Cooperating Layered Architecture for SDN

- **FCAPS**: Fault, Configuration, Accounting, Performance and Security

- **SDN**: Software Defined Networking
3. Architecture overview

Current operator networks support multiple services (e.g., VoIP, IPTV, mobile VoIP, critical mission applications, etc.) on a variety of transport technologies. The provision and delivery of a service independently of the underlying transport capabilities requires a separation of the service related functionalities and an abstraction of the transport network to hide the specificities of underlying transfer techniques while offering a common set of capabilities.

Such separation can provide configuration flexibility and adaptability from the point of view of either the services or the transport network. Multiple services can be provided on top of a common transport infrastructure, and similarly, different technologies can accommodate the connectivity requirements of a certain service. A close coordination among them is required for a consistent service delivery (inter-layer cooperation).

This document focuses particularly on:

- Means to expose transport capabilities to external services.
- Means to capture service requirements of services.
- Means to notify service intelligence with underlying transport events, for example to adjust service decision-making process with underlying transport events.
- Means to instruct the underlying transport capabilities to accommodate new requirements, etc.

An example is to guarantee some Quality of Service (QoS) levels. Different QoS-based offerings could be present at both service and transport layers. Vertical mechanisms for linking both service and transport QoS mechanisms should be in place to provide the quality guarantees to the end user.

CLAS architecture assumes that the logically centralized control functions are separated in two functional blocks or layers. One of the functional blocks comprises the service-related functions, whereas the other one contains the transport-related functions. The cooperation between the two layers is considered to be implemented through standard interfaces.

Figure 1 shows the CLAS architecture. It is based on functional separation in the NGN architecture defined by the ITU-T in [Y.2011].
Two strata of functionality are defined, namely the Service Stratum, comprising the service-related functions, and the Transport Stratum, covering the transport ones. The functions on each of these layers are further grouped on control, management and user (or data) planes.
In the CLAS architecture both the control and management functions are the ones logically centralized in one or a set of SDN controllers, in such a way that separated SDN controllers are present.
in the Service and Transport strata. Furthermore, the generic user or data plane functions included in the NGN architecture are referred here as resource plane functions. The resource plane in each stratum is controlled by the corresponding SDN controller through a standard interface.

The SDN controllers cooperate for the provision and delivery of services. There is a hierarchy in which the Service SDN controller requests transport capabilities to the Transport SDN controller. Furthermore, the Transport SDN controller interacts with the Service SDN controller to inform it about events in the transport network that can motivate actions in the service layer.

The Service SDN controller acts as a client of the Transport SDN controller.

Despite it is not shown in the figure, the Resource planes of each stratum could be connected. This will depend on the kind of service provided. Furthermore, the Service stratum could offer an interface towards external applications to expose network service capabilities to those applications or customers.

This document does assume that SDN techniques can be enabled jointly with other distributed means (e.g., IGP).

3.1. Functional strata

As described before, the functional split separates transport-related functions from service-related functions. Both strata cooperate for a consistent service delivery.

Consistency is determined and characterized by the service layer.

Communication between these two components could be implemented using a variety of means (such as [I-D.boucadair-connectivity-provisioning-protocol], Intermediate-Controller Plane Interface (I-CPI) [ONFArch], etc).

3.1.1. Transport stratum

The Transport stratum comprises the functions focused on the transfer of data between the communication end points (e.g., between end-user devices, between two service gateways, etc.). The data forwarding nodes are controlled and managed by the Transport SDN component. The Control plane in the SDN controller is in charge of instructing the forwarding devices to build the end to end data path for each communication or to make sure forwarding service is appropriately setup. Forwarding may not be rely on the sole pre-configured
entries; dynamic means can be enabled so that involved nodes can build dynamically routing and forwarding paths. Finally, the Management plane performs management functions (i.e., FCAPS) on those devices, like fault or performance management, as part of the Transport stratum capabilities.

3.1.2. Service stratum

The Service stratum contains the functions related to the provision of services and the capabilities offered to external applications. The Resource plane consists of the resources involved in the service delivery, such as computing resources, registries, databases, etc. The Control plane is in charge of controlling and configuring those resources, as well as interacting with the Control plane of the Transport stratum in client mode for requesting transport capabilities for a given service. In the same way, the Management plane implements management actions on the service-related resources and interacts with the Management plane in the Transport stratum for a cooperating management between layers.

3.1.3. Recursiveness

Recursive layering can happen in some usage scenarios in which the Transport Stratum is itself structured in Service and Transport Stratum. This could be the case of the provision of a transport services complemented with advanced capabilities additional to the pure data transport (e.g., maintenance of a given SLA [RFC7297]).

3.2. Plane separation

The CLAS architecture leverages on the SDN proposition of plane separation. As mentioned before, three different planes are considered for each stratum. The communication among these three planes (and with the corresponding plane in other strata) is based on open, standard interfaces.

3.2.1. Control Plane

The Control plane logically centralizes the control functions of each stratum and directly controls the corresponding resources. [RFC7426] introduces the role of the control plane in a SDN architecture. This plane is part of an SDN controller, and can interact with other control planes in the same or different strata for accomplishing control functions.
3.2.2. Management Plane

The Management plane logically centralizes the management functions for each stratum, including the management of the Control and Resource planes. [RFC7426] describes the functions of the management plane in a SDN environment. This plane is also part of the SDN controller, and can interact with the corresponding management planes residing in SDN controllers of the same or different strata.

3.2.3. Resource Plane

The Resource plane comprises the resources for either the transport or the service functions. In some cases the service resources can be connected to the transport ones (e.g., being the terminating points of a transport function) whereas in other cases it can be decoupled from the transport resources (e.g., one database keeping some register for the end user). Both forwarding and operational planes proposed in [RFC7426] would be part of the Resource plane in this architecture.

4. Required features

A number of features are required to be supported by the CLAS architecture.

- Abstraction: the mapping of physical resources into the corresponding abstracted resources.

- Service parameter translation: translation of service parameters (e.g., in the form of SLAs) to transport parameters (or capabilities) according to different policies.

- Monitoring: mechanisms (e.g. event notifications) available in order to dynamically update the (abstracted) resources’ status taking in to account e.g. the traffic load.

- Resource computation: functions able to decide which resources will be used for a given service request. As an example, functions like PCE could be used to compute/select/decide a certain path.

- Orchestration: ability to combine diverse resources (e.g., IT and network resources) in an optimal way.

- Accounting: record of resource usage.

- Security: secure communication among components, preventing e.g. DoS attacks.
5. Communication between SDN Controllers

The SDN Controller residing respectively in the Service and the Transport Stratum need to establish a tight coordination. Mechanisms for transfer relevant information for each stratum should be defined.

From the Service perspective, the Service SDN controller needs to easily access transport resources through well defined APIs to access the capabilities offered by the Transport Stratum. There could be different ways of obtaining such transport-aware information, i.e., by discovering or publishing mechanisms. In the former case the Service SDN Controller could be able of handling complete information about the transport capabilities (including resources) offered by the Transport Stratum. In the latter case, the Transport Stratum exposes available capabilities e.g. through a catalog, reducing the amount of detail of the underlying network.

On the other hand, the Transport Stratum requires to properly capture Service requirements. These can include SLA requirements with specific metrics (such as delay), level of protection to be provided, max/min capacity, applicable resource constraints, etc.

The communication between controllers should be also secure, preventing denial of service.

6. Deployment scenarios

Different situations can be found depending on the characteristics of the networks involved in a given deployment.

6.1. Full SDN environments

This case considers the fact that the networks involved in the provision and delivery of a given service have SDN capabilities.

6.1.1. Multiple Service strata associated to a single Transport stratum

A single Transport stratum can provide transfer functions to more than one Service strata. The Transport stratum offers a standard interface to each of the Service strata. The Service strata are the clients of the Transport stratum. Some of the capabilities offered by the Transport stratum can be isolation of the transport resources (slicing), independent routing, etc.
6.1.2. Single service stratum associated to multiple Transport strata

A single Service stratum can make use of different Transport strata for the provision of a certain service. The Service stratum interfaces each of the Transport strata with standard protocols, and orchestrates the provided transfer capabilities for building the end to end transport needs.

6.2. Hybrid environments

This case considers scenarios where one of the strata is legacy totally or in part.

6.2.1. SDN Service stratum associated to a legacy Transport stratum

An SDN service stratum can interact with a legacy Transport stratum through some interworking function able to adapt SDN-based control and management service-related commands to legacy transport-related protocols, as expected by the legacy Transport stratum. The SDN controller in the Service stratum is not aware of the legacy nature of the underlying Transport stratum.

6.2.2. Legacy Service stratum associated to an SDN Transport stratum

A legacy Service stratum can work with an SDN-enabled Transport stratum through the mediation of and interworking function capable to interpret commands from the legacy service functions and translate them into SDN protocols for operating with the SDN-enabled Transport stratum.

6.3. Multi-domain scenarios in Transport Stratum

The Transport Stratum can be composed by transport resources being part of different administrative, topological or technological domains. The Service Stratum can yet interact with a single entity in the Transport Stratum in case some abstraction capabilities are provided in the transport part to emulate a single stratum.

Those abstraction capabilities constitute a service itself offered by the Transport Stratum to the services making use of it. This service is focused on the provision of transport capabilities, then different of the final communication service using such capabilities.

In this particular case this recursion allows multi-domain scenarios at transport level.
Multi-domain situations can happen in both single-operator and multi-operator scenarios. Multi-operator scenarios will be addressed in future versions of the document.

In single operator scenarios a multi-domain or end-to-end abstraction component can provide an homogeneous abstract view of the underlying heterogeneous transport capabilities for all the domains.

7. Use cases

This section presents a number of use cases as examples of applicability of this proposal

7.1. Network Function Virtualization

To be completed

7.2. Abstraction and Control of Transport Networks

To be completed.

8. IANA Considerations

TBD.

9. Security Considerations

TBD. Security in the communication between strata to be addressed.

10. References

10.1. Normative References


10.2. Informative References
[I-D.boucadair-connectivity-provisioning-protocol]


Authors’ Addresses

Luis M. Contreras
Telefonica
Ronda de la Comunicacion, s/n
Sur-3 building, 3rd floor
Madrid  28050
Spain

Email: luismiguel.contrerasmurillo@telefonica.com
URI:   http://people.tid.es/LuisM.Contreras/

Carlos J. Bernardos
Universidad Carlos III de Madrid
Av. Universidad, 30
Leganes, Madrid 28911
Spain

Phone: +34 91624 6236
Email: cjbc@it.uc3m.es
URI:   http://www.it.uc3m.es/cjbc/
Abstract

In certain networks deployment scenarios, service providers would like to keep all the existing MPLS functionalities in both MPLS and GMPLS network while removing the complexity of existing signaling protocols such as LDP and RSVP-TE. In this document, we propose to use the PCE as a central controller so that LSP can be calculated/signaled/initiated/downloaded/managed through a centralized PCE server to each network devices along the LSP path while leveraging the existing PCE technologies as much as possible.

This draft describes the use cases for using the PCE as the central controller where LSPs are calculated/setup/initiated/downloaded/maintained through extending the current PCE architectures and extending the 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 http://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."
Internet-Draft                    PCECC                        July 2015

This Internet-Draft will expire on January 7, 2016.

Copyright Notice

Copyright (c) 2015 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.  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 ................................................. 3
   1.1.  Background .............................................. 3
   1.2.  Using the PCE as the Central Controller (PCECC)
         Approach .............................................. 4
2.  Terminology .................................................. 7
3.  PCEP Requirements ........................................... 7
4.  Use Cases of PCECC for Label Resource Reservations .......... 8
5.  Using PCECC for SR without the IGP Extension ............... 9
   5.1.  Use Cases of PCECC for SR Best Effort (BE) Path ........ 10
   5.2.  Use Cases of PCECC for SR Traffic Engineering (TE) Path 11
6.  Use Cases of PCECC for TE LSP ................................ 12
7.  Use Cases of PCECC for Multicast LSPs ....................... 14
   7.1.  Using PCECC for P2MP/MP2MP LSPs’ Setup ................ 14
   7.2.  Use Cases of PCECC for the Resiliency of P2MP/MP2MP
         LSPs .................................................. 15
       7.2.1.  PCECC for the End-to-End Protection of the
               P2MP/MP2MP LSPs .................................. 15
       7.2.2.  PCECC for the Local Protection of the P2MP/MP2MP
               LSPs .............................................. 16
8.  Use Cases of PCECC for LSP in the Network Migration ........ 17
9.  Use Cases of PCECC for L3VPN and PWE3 ..................... 19
10. The Considerations for PCECC Procedure and PCEP extensions. 19
11. IANA Considerations ......................................... 20
12. Security Considerations ...................................... 20
13. Acknowledgments ............................................. 20
14. References .................................................. 20
   14.1. Normative References .................................... 20
   14.2. Informative References ................................. 20
1. Introduction

1.1. Background

In certain network deployment scenarios, service providers would like to have the ability to dynamically adapt to a wide range of customer's requests for the sake of flexible network service delivery, SDN has provides additional flexibility in how the network is operated comparing the traditional network.

The existing networking ecosystem has become awfully complex and highly demanding in terms of robustness, performance, scalability, flexibility, agility, etc. By migrating to the SDN enabled network from the existing network, service providers and network operators must have a solution which they can evolve easily from the existing network into the SDN enabled network while keeping the network services remain scalable, guarantee robustness and availability etc.

Taking the smooth transition between traditional network and the new SDN enabled network into account, especially from a cost impact assessment perspective, using the existing PCE components from the current network to function as the central controller of the SDN network is one choice, which not only achieves the goal of having a centralized controller to provide the functionalities needed for the central controller, but also leverages the existing PCE network components.

The Path Computation Element communication Protocol (PCEP) provides mechanisms for Path Computation Elements (PCEs) to perform route computations in response to Path Computation Clients (PCCs) requests. PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model draft [I-D. draft-ietf-pce-stateful-pce] describes a set of extensions to PCEP to enable active control of MPLS-TE and GMPLS tunnels.

[I-D.crabbe-pce-pce-initiated-lsp] describes the setup and teardown of PCE-initiated LSPs under the active stateful PCE model, without the need for local configuration on the PCC, thus allowing for a dynamic MPLS network that is centrally controlled and deployed.

[I-D.ali-pce-remote-initiated-gmpls-lsp] complements [I-D. draft-crabbe-pce-pce-initiated-lsp] by addressing the requirements for remote-initiated GMPLS LSPs.

SR technology leverages the source routing and tunneling paradigms. A source node can choose a path without relying on hop-by-hop signaling protocols such as LDP or RSVP-TE. Each path is specified as a set of "segments" advertised by link-state routing protocols.
(IS-IS or OSPF). [I-D.filsfils-spring-segment-routing] provides an introduction to SR technology. The corresponding IS-IS and OSPF extensions are specified in [I-D.ietf-isis-segment-routing-extensions] and [I-D.psenak-ospf-segment-routing-extensions], respectively.

A Segment Routed path (SR path) can be derived from an IGP Shortest Path Tree (SPT). Segment Routed Traffic Engineering paths (SR-TE paths) may not follow IGP SPT. Such paths may be chosen by a suitable network planning tool and provisioned on the source node of the SR-TE path.

It is possible to use a stateful PCE for computing one or more SR-TE paths taking into account various constraints and objective functions. Once a path is chosen, the stateful PCE can instantiate an SR-TE path on a PCC using PCEP extensions specified in [I-D.crabbe-pce-pce-initiated-lsp] using the SR specific PCEP extensions described in [I-D.sivabalan-pce-segment-routing].

By using the solutions provided from above drafts, LSP in both MPLS and GMPLS network can be setup/delete/maintained/synchronized through a centrally controlled dynamic MPLS network. Since in these solutions, the LSP is need to be signaled through the head end LER to the tail end LER, there are either RSVP-TE signaling protocol need to be deployed in the MPLS/GMPLS network, or extend TGP protocol with node/adjacency segment identifiers signaling capability to be deployed.

The PCECC solution proposed in this document allow for a dynamic MPLS network that is eventually controlled and deployed without the deployment of RSVP-TE protocol or extended IGP protocol with node/adjacency segment identifiers signaling capability while providing all the key MPLS functionalities needed by the service providers. These key MPLS features include MPLS P2P LSP, P2MP/MP2MP LSP, MPLS protection mechanism etc. In the case that one LSP path consists legacy network nodes and the new network nodes which are centrally controlled, the PCECC solution provides a smooth transition step for users.

1.2. Using the PCE as the Central Controller (PCECC) Approach

With PCECC, it not only removes the existing MPLS signaling totally from the control plane without losing any existing MPLS functionalities, but also PCECC achieves this goal through utilizing the existing PCEP without introducing a new protocol into the network.

The following diagram illustrates the PCECC architecture.
Through the draft, we call the combination of the functionality for global label range signaling and the functionality of LSP setup/download/cleanup using the combination of global labels and local labels as PCECC functionality.

Current MPLS label has local meaning. That is, MPLS label allocated locally and signaled through the LDP/RSVP-TE/BGP etc dynamic signaling protocol.

As the SDN (Service-Driven Network) technology develops, MPLS global label has been proposed again for new solutions. [I-D.li-mpls-global-label-usecases] proposes possible use cases of MPLS global label. MPLS global label can be used for identification of the location, the service and the network in different application scenarios. From these use cases we can see that no matter SDN or traditional application scenarios, the new solutions based on MPLS global label can gain advantage over the existing solutions to facilitate service provisions. The solution choices are described in [I-D.li-mpls-global-label-framework].

To ease the label allocation and signaling mechanism, also with the new applications such as concentrated LSP controller is introduced, PCE can be conveniently used as a central controller and MPLS global label range negotiator.
The later section of this draft describes the user cases for PCE server and PCE clients to have the global label range negotiation and local label range negotiation functionality.

To empower networking with centralized controllable modules, there are many choices for downloading the forwarding entries to the data plane, one way is the use of the OpenFlow protocol, which helps devices populate their forwarding tables according to a set of instructions to the data plane. There are other candidate protocols to convey specific configuration information towards devices also. Since the PCEP protocol is already deployed in some of the service network, to leverage the PCEP to populated the MPLS forwarding table is a possible good choice.

For the centralized network, the performance achieved through distributed system can not be easy matched if all of the forwarding path is computed, downloaded and maintained by the centralized controller. The performance can be improved by supporting part of the forwarding path in the PCECC network through the segment routing mechanism except that the adjacency IDs for all the network nodes and links are propagated through the centralized controller instead of using the IGP extension.

The node and link adjacency IDs can be negotiated through the PCECC with each PCECC clients and these IDs can be just taken from the global label range which has been negotiated already.

With the capability of supporting SR within the PCECC architecture, all the p2p forwarding path protection use cases described in the draft [I-D.ietf-spring-resiliency-use-cases] will be supported too within the PCECC network. These protection alternatives include end-to-end path protection, local protection without operator management and local protection with operator management.

With the capability of global label and local label existing at the same time in the PCECC network, PCECC will use compute, setup and maintain the P2MP and MP2MP LSP using the local label range for each network nodes.

With the capability of setting up/maintaining the P2MP/MP2MP LSP within the PCECC network, it is easy to provide the end-end managed path protection service and the local protection with the operation management in the PCECC network for the P2MP/MP2MP LSP, which includes both the RSVP-TE P2MP based LSP and also the mLDP based LSP.
2. Terminology

The following terminology is used in this document.


PCC: Path Computation Client: any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

TE: Traffic Engineering.

3. PCEP Requirements

Following key requirements associated PCECC should be considered when designing the PCECC based solution:

1. Path Computation Element (PCE) clients supporting this draft MUST have the capability to advertise its PCECC capability to the PCECC.

2. Path Computation Element (PCE) supporting this draft MUST have the capability to negotiate a global label range for a group of clients.

3. Path Computation Client (PCC) MUST be able ask for global label range assigned in path request message.

4. PCE are not required to support label reserve service. Therefore, it MUST be possible for a PCE to reject a Path Computation Request message with a reason code that indicates no support for label reserve service.

5. PCEP SHOULD provide a means to return global label range and LSP label assignments of the computed path in the reply message.

6. PCEP SHOULD provide a means to download the MPLS forwarding entry to the PCECC’s clients.
4. Use Cases of PCECC for Label Resource Reservations

Example 1 to 2 are based on network configurations illustrated using the following figure:

![Diagram of network configuration]

**Figure 2: Using PCECC for Global Label Allocation**

**Example 1: Shared Global Label Range Reservation**

- PCECC Clients nodes report MPLS label capability to the central controller PCECC.
- The central controller PCECC collects MPLS label capability of all nodes. Then PCECC can calculate the shared MPLS global label range for all the PCECC client nodes.
- In the case that the shared global label range need to be negotiated across multiple domains, the central controllers of these domains need to communicate to negotiate a common global label range.
- The central controller PCECC notifies the shared global label range to all PCECC client nodes.

**Example 2: Global Label Allocation**
PCECC Client node1 send global label allocation request to the central controller PCECC1.

- The central controller PCECC1 allocates the global label for FEC1 from the shared global label range and sends the reply to the client node1.

- The central controller PCECC1 notifies the allocated label for FEC1 to all PCECC client nodes within domain 1.

5. Using PCECC for SR without the IGP Extension

For the centralized network, the performance achieved through distributed system can not be easy matched if all of the forwarding path is computed, downloaded and maintained by the centralized controller. The performance can be improved by supporting part of the forwarding path in the PCECC network through the segment routing mechanism except that node segment ids and adjacency segment IDs for all the network are allocated dynamically and propagated through the centralized controller instead of using the IGP extension.

When the PCECC is used for the distribution of the node segment ID and adjacency segment ID, the node segment ID is allocated from the global label pool. For the allocation of adjacency segment ID, there are two choices, the first choice is that it is allocated from the local label pool, the second choice is that it is allocated from the global label pool. The advantage for the second choice is that the depth of the label stack for the forwarding path encoding will be reduced since adjacency segment ID can signal the forwarding path without adding the node segment ID in front of it. In this version of the draft, we use the fist choice for now. We may update the draft to reflect the use of the second choice.

Same as the SR solutions, when PCECC is used as the central controller, the support of FRR on any topology can be pre-computated and setup without any additional signaling (other than the regular IGP/BGP protocols) including the support of shared risk constraints, support of node and link protection and support of microloop avoidance.

The following example illustrate the use case where the node segment ID and adjacency segment ID are allocated from the global label allocated for SR path.
In this mode of the solution, the PCECC just need to allocate the node segment ID and adjacency ID without calculating the explicit path for the SR path. The ingress of the forwarding path just need to encapsulate the destination node segment ID on top of the packet. All the intermediate nodes will forward the packet based on the final destination node segment ID. It is similar to the LDP LSP forwarding except that label swapping is using the same global label both for the in segment and out segment in each hop.

The p2p SR BE path examples are explained as below:

Note that the node segment id for each node from the shared global labels ranges negotiated already.

Example 1:

R1 may send a packet to R8 simply by pushing an SR header with segment list (1008). The path can be: R1-R2-R3-R8 or R1-R2-R5-R8
depending on the route calculation on node R2.

Example 2: local link/node protection:

For the packet which has destination of R3 and after that, R2 may preinstalled the backup forwarding entry to protect the R4 node, the pre-installed the backup path can go through either node5 or link1 or link2 between R2 and R3. The backup path calculation is locally decided by R2 and any existing IP FRR algorithms can be used here.

5.2. Use Cases of PCECC for SR Traffic Engineering (TE) Path

In the case of traffic engineering path is needed, the PCECC need to allocate the node segment ID and adjacency ID, and at the same time PCECC calculates the explicit path for the SR path and pass this explicit path represented with a sequence of node segment id and adjacency id. The ingress of the forwarding path need to encapsulate the stack of node segment id and adjacency id on top of the packet. For the case where strict traffic engineering path is needed, all the intermediate nodes and links will be specified through the stack of labels so that the packet is forwarded exactly as it is wanted.

Even though it is similar to TE LSP forwarding where forwarding path is engineered, but the Qos is only guaranteed through the enforce of the bandwidth admission control. As for the RSVP-TE LSP case, Qos is guaranteed through the link bandwidth reservation in each hop of the forwarding path.

The p2p SR traffic engineering path examples are explained as bellow:

Note that the node segment id for each node is allocated from the shared global labels ranges negotiated already and adjacency segment ids for each link are allocated from the local label pool for each node.

Example 1:

R1 may send a packet P1 to R8 simply by pushing an SR header with segment list (1008). The path should be: R1-R2-R3-R8.

Example 2:

R1 may send a packet P2 to R8 by pushing an SR header with segment list (1002, 9001, 1008). The path should be: R1-R2-(1)link-R3-R8.

Example 3:

R1 may send a packet P3 to R8 while avoiding the links between R2 and
R3 by pushing an SR header with segment list \{1004, 1008\}. The path should be: R1-R2-R4-R3-R8

The p2p local protection examples for SR TE path are explained as below:

Example 4: local link protection:

- R1 may send a packet P4 to R8 by pushing an SR header with segment list \{1002, 9001, 1008\}. The path should be: R1-R2-(1)link-R3-R8.

- When node R2 receives the packet from R1 which has the header of R2-(1)link-R3-R8, and also finds out there is a link failure of link1, then it will send out the packet with header of R3-R8 through link2.

Example 5: local node protection:

- R1 may send a packet P5 to R8 by pushing an SR header with segment list \{1004, 1008\}. The path should be: R1-R2-R4-R3-R8.

- When node R2 receives the packet from R1 which has the header of \{1004, 1008\}, and also finds out there is a node failure for node4, then it will send out the packet with header of \{1005, 1008\} to node5 instead of node4.

6. Use Cases of PCECC for TE LSP

In the previous sections, we have discussed the cases where the SR path is setup through the PCECC. Although those cases give the simplicity and scalability, but there are existing functionalities for the traffic engineering path such as the bandwidth guarantee through the full forwarding path and the multicast forwarding path which SR based solution cannot solve. Also there are cases where the depth of the label stack may have been an issue for existing deployment and certain vendors.

So to address these issues, PCECC architecture should also support the TE LSP and multicast LSP functionalities. To achieve this, the existing PCEP can be used to communicate between the PCE server and PCE’s client PCC for exchanging the path request and reply information regarding to the TE LSP info. In this case, the TE LSP info is not only the path info itself, but it includes the full forwarding info. Instead of letting the ingress of LSP to initiate the LSP setup through the RSVP-TE signaling protocol, with minor extensions, we can use the PCEP to download the complete TE LSP forwarding entries for each node in the network.
TE LSP Setup Example

- Node1 sends a path request message for the setup of TE LSP from R1 to R8.

- PCECC program each node along the path from R1 to R8 with the primary path: (R1, link1, 6001), (R2, link3, 7002), (R4, link0, 9001), (R3, link1, 3001), (R8).

- For the end to end protection, PCECC program each node along the path from R1 to R8 with the secondary path: (R1, link2, 6002), (R2, link4, 7001), (R5, link1, 9002), (R3, link2, 3002), (R8).

- It is also possible to have a secondary backup path for the local node protection setup by PCECC. For example, the primary path is still the same as what we have setup so far, then to protect...
the node R4 locally, PCECC can program the secondary path like this: (R1, link1, 6001), (R2, link1, 5001), (R3, link1, 3001), (R8). By doing this, the node R4 is locally protected.

7. Use Cases of PCECC for Multicast LSPs

The current multicast LSPs are setup either using the RSVP-TE P2MP or mLDP protocols. The setup of these LSPs not only need a lot of manual configurations, but also it is also complex when the protection is considered. By using the PCECC solution, the multicast LSP can be computed and setup through centralized controller which has the full picture of the topology and bandwidth usage for each link. It not only reduces the complex configurations comparing the distributed RSVP-TE P2MP or mLDP signal lings, but also it can compute the disjoint primary path and secondary path efficiently.

7.1. Using PCECC for P2MP/MP2MP LSPs’ Setup

With the capability of global label and local label existing at the same time in the PCECC network, PCECC will use compute, setup and maintain the P2MP and MP2MP lsp using the local label range for each network nodes.

```
+----------+
|    R1    | Root node of the multicast LSP
|----------+
|   6000   |
|----------+
|Transit Node|    R2    |
|----------+
|    *     |   *     |
|   9001*   |   *     |
|    *     |   *     |
|    +-----|   +-----|
|    | R4    |   | R5    |
|    | +-----|   | +-----|
|    |  *   |   |  *   |
|    |  *   |   |  *   |
|    |  +   |   |  +   |
|    +-----+   +-----+
|    | R3    |   | R5    |
|    | +-----+   | +-----+
|    |  *   |   |  +   |
|    |  +   |   |  +   |
|    +-----+   +-----+
|    | R8    |   | Leaf Node
|    +-----+
|         |
|        |
```

Figure 5: Using PCECC for P2MP TE LSP
The P2MP examples are explained here:

Step1: R1 may send a packet P1 to R2 simply by pushing an label of 6000 to the packet.

Step2: After R2 receives the packet with label 6000, it will forwarding to R4 by pushing header of 9001 and R5 by pusing header of 9002.

Step3: After R4 receives the packet with label 9001, it will forwarding to R3 by pushing header of 9003. After R5 receives the packet with label 9002, it will forwarding to R5 by pushing header of 9004.

Step3: After R3 receives the packet with label 9003, it will forwarding to R8 by pushing header of 9005

7.2. Use Cases of PCECC for the Resiliency of P2MP/MP2MP LSPs

7.2.1. PCECC for the End-to-End Protection of the P2MP/MP2MP LSPs

In this section we describe the end-end managed path protection service and the local protection with the operation management in the PCECC network for the P2MP/MP2MP LSP, which includes both the RSVP-TE P2MP based LSP and also the mLDP based LSP.

An end-to-end protection (for nodes and links) principle can be applied for computing backup P2MP or MP2MP LSPs. During computation of the primarily multicast trees, PCECC server may also be taken into consideration to compute a secondary tree. A PCE may compute the primary and backup P2MP or MP2Mp LSP together or sequentially.
In the example above, when the PCECC setup the primary multicast tree from the root node R1 to the leafs, which is R1->R2->(R4, R5), at same time, it can setup the backup tree, which is R11->R3->(R4, R5). Both the these two primary forwarding tree and secondary forwarding tree will be downloaded to each routers along the primary path and the secondary path. The traffic will be forwarded through the R1->R2->(R4, R5) path normally, and when there is a node in the primary tree, then the root node R1 will switch the flow to the backup tree, which is R11->R3->(R4, R5). By using the PCECC, the path computation and forwarding path downloading can all be done without the complex signaling used in the P2MP RSVP-TE or mLDP.

7.2.2. PCECC for the Local Protection of the P2MP/MP2MP LSPs

In this section we describe the local protection service in the PCECC network for the P2MP/MP2MP LSP.

While the PCECC sets up the primary multicast tree, it can also build the back LSP among PLR, the protected node, and MPs (the downstream nodes of the protected node). In the cases where the amount of downstream nodes are huge, this mechanism can avoid unnecessary packet duplication on PLR, so that protect the network from traffic congestion risk.
In the example above, when the PCECC setup the primary multicast path around the PLR node R10 to protect node R20, which is R10->R20->{R40, R50}, at same time, it can setup the backup path R10->R30->{R40, R50}. Both the these two primary forwarding path and secondary forwarding path will be downloaded to each routers along the primary path and the secondary path. The traffic will be forwarded through the R10->R20->{R40, R50} path normally, and when there is a node failure for node R20, then the PLR node R10 will switch the flow to the backup path, which is R10->R30->{R40, R50}. By using the PCECC, the path computation and forwarding path downloading can all be done without the complex signaling used in the P2MP RSVP-TE or mLDP.

8. Use Cases of PCECC for LSP in the Network Migration

One of the main advantages for PCECC solution is that it has backward compatibility naturally since the PCE server itself can function as a proxy node of MPLS network for all the new nodes which don’t support the existing MPLS signaling protocol anymore.
As it is illustrated in the following example, the current network will migrate to a total PCECC controlled network gradually by replacing the legacy nodes. During the migration, the legacy nodes still need to signal using the existing MPLS protocol such as LDP and RSVP-TE, and the new nodes setup their portion of the forwarding path through PCECC directly. With the PCECC function as the proxy of these new nodes, MPLS signaling can populate through network as normal.

Example described in this section is based on network configurations illustrated using the following figure:

![Figure 8: Using PCECC During Migration](image)

Example: PCECC Initiated LSP Setup In the Network Migration

In this example, there are five nodes for the TE LSP from head end (node1) to the tail end (node5). Where the NodeX is central controlled and other nodes are legacy nodes.

- Node1 sends a path request message for the setup of LSP destinating to Node5.

- PCECC sends a reply message for LSP setup with path (node1, if1), (node2, if22), (node-PCECC, if44), (node4, if4), Nnode5.

- Node1, Node2, Node-PCECC, Node 5 will setup the LSP to Node5 normally using the local label as normal.
Then the PCECC will program the outsegment of Node2, the insegment of Node4, and the insegment/outsegment for NodeX.

9. Use Cases of PCECC for L3VPN and PWE3

The existing services using MPLS LSP tunnels based on MPLS signalling mechanism such L3VPN, PWE3 and IPv6 can be simplified by using the PCECC to negotiate the label assignments for the L3VPN, PWE3 and IPv6.

In the case of L3VPN, VPN labels can be negotiated and distributed through the PCECC PCEP among the PE router instead of using the BGP protocols.

```
+-------------------------------------------+
|                   PCE DOMAIN              |
|    +-----------------------------------+  |
|    |                PCECC              |  |
|    +-----------------------------------+  |
|           ^          ^              ^     |
|PWE3/L3VPN | PCEP PCEP|LSP PWE3/L3VPN|PCEP |
|           V          V              V     |
|           +--------+ |     +--------+   +--------+   +--------+  |  +--------+  |
|  CE    | |     | PE1    |   | NODE x |   | PE2    |  |  |   CE   |
|        |...... |        |...|        |...|        |.....|        |
| Legacy | |if1  | PCECC  |if2|PCCEC   |if3| PCECC  |if4  | Legacy |
|  Node  | |     | Enabled|   |Enabled |   |Enabled |  |  |  Node  |
|        |     +--------+   +--------+   +--------+  |  +--------+  |
+-------------------------------------------+
```

Figure 9: Using PCECC for L3VPN and PWE3

In the cast PWE3, instead of using the LDP signalling protocols, the label and port pairs assigned to each pseudowire can be negotiated through PCECC among the PE routers and the corresponding forwarding entries will be distributed into each PE routers through the extended PCEP protocols.

10. The Considerations for PCECC Procedure and PCEP extensions

The PCECC’s procedures and PCEP extensions is defined in [I-D.zhao-pce-pcep-extension-for-pce-controller].
11. IANA Considerations

This document does not require any action from IANA.

12. Security Considerations

TBD.

13. Acknowledgments

We would like to thank Robert Tao, Changjiang Yan, Tieying Huang for their useful comments and suggestions.

14. References

14.1. Normative References


14.2. Informative References


[I-D.crabbe-pce-pce-initiated-lsp] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE"
Internet-Draft                    PCECC                        July 2015

Model", draft-
crabbe-pce-pce-
initiated-lsp-03
(work in progress),
October 2013.

[I-D.ali-pce-remote-initiated-gmpls-lsp]  
Ali, Z., Sivabalan,
S., Filsfils, C.,
Varga, R., Lopez,
V., Dios, O., and
X. Zhang, "Path
Computation Element
Communication
Protocol (PCEP)
Extensions for
remote-initiated
GMPLS LSP Setup", d
raft-ali-pce-
remote-initiated-
gmpls-lsp-03 (work
in progress),
February 2014.

[I-D.ietf-isis-segment-routing-extensions]  
Previdi, S.,
Filsfils, C.,
Bashandy, A.,
Gredler, H.,
Litkowski, S.,
Decraene, B., and
J. Tantsura, "IS-IS
Extensions for
Segment Routing", d
raft-ietf-isis-
segment-routing-
extensions-05 (work
in progress),
June 2015.

[I-D.psenak-ospf-segment-routing-extensions]  
Psenak, P.,
Previdi, S.,
Filsfils, C.,
Gredler, H.,
Shakir, R.,
Henderickx, W., and
J. Tantsura, "OSPF
Extensions for
Segment Routing", d
raft-psenak-ospf-
Internet-Draft                  PCECC                        July 2015

segment-routing-
extensions-05 (work
in progress),
June 2014.

[I-D.sivabalan-pce-segment-routing] Sivabalan, S.,
Medved, J.,
Filsfils, C.,
Crabbe, E., Raszuk,
R., Lopez, V., and
J. Tantsura, "PCEP
Extensions for
Segment Routing", d
raft-sivabalan-pce-
segment-routing-03
(work in progress),
July 2014.

[I-D.li-mpls-global-label-usecases] Li, Z., Zhao, Q.,
Yang, T., and R.
Raszuk, "Use Cases
of MPLS Global
Label", draft-li-
mpls-global-label-
usecases-02 (work
in progress),
July 2014.

[I-D.li-mpls-global-label-framework] Li, Z., Zhao, Q.,
Chen, X., Yang, T.,
and R. Raszuk, "A
Framework of MPLS
Global Label", draf
t-li-mpls-global-
label-framework-02
(work in progress),
July 2014.

[I-D.zhao-pce-pcep-extension-for-pce-controller] Zhao, Q., Zhao, K.,
Li, Z., Dhody, D.,
Palle, U., and T.
Communications,
"PCEP Procedures
and Protocol
Extensions for
Using PCE as a
Central Controller
(PCECC) of LSPs", d
[I-D.ietf-spring-resiliency-use-cases]


Authors’ Addresses

Quintin Zhao
Huawei Technologies
125 Nagog Technology Park
Acton, MA 01719
US
EMail: quintin.zhao@huawei.com

Robin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
EMail: lizhenbin@huawei.com

King Ke
Tencent Holdings Ltd.
Shenzhen
China
EMail: kinghe@tencent.com
Luyuan Fang
Microsoft

EMail: lufang@microsoft.com

Chao Zhou
Cisco Systems

EMail: chao.zhou@cisco.com

Boris Zhang
Telus Communications
200 Consilium Pl Floor 15
Toronto, ON M1H 3J3
Canada

EMail: Boris.Zhang@telus.com

Katherine Zhao
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

EMail: Katherine.zhao@huawei.com

Richard Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

EMail: renwei.li@huawei.com
PCEP Procedures and Protocol Extensions for Using PCE as a Central Controller (PCECC) of LSPs
draft-zhao-pce-pcep-extension-for-pce-controller-02

Abstract

In certain networks deployment scenarios, service providers would like to keep all the existing MPLS functionalities in both MPLS and GMPLS while removing the complexity of existing signaling protocols such as LDP and RSVP-TE. In [I-D.zhao-pce-central-controller-user-cases], we propose to use the PCE [RFC5440] as a central controller (PCECC) so that LSP can be calculated/ signaled/initiated and label forwarding entries are downloaded through a centralized PCE server to each network devices along the LSP path while leveraging the existing PCE technologies as much as possible.

This draft specify the procedures and PCEP protocol extensions for using the PCE as the central controller and user cases where LSPs are calculated/setup/initiated and label forwarding entries are downloaded through extending the existing PCE architectures and PCEP.

This document also discuss the role of PCECC in Segment Routing (SR).

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 http://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, because Internet-Drafts are subject to change without notice. There is no guarantee that documents can be retrieved from a particular location after a certain time, despite any effort or record keeping that may be undertaken by a distributed Internet-Draft distribution system. Authors can advertise their Internet-Drafts through informal means outside the IETF, but once their Internet-Drafts are placed in Internet-Drafts directories, they are expected to maintain that presence and make their Internet-Drafts available through the distribution system.
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 April 9, 2016.

Copyright Notice

Copyright (c) 2015 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. 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 ............................................. 3
   1.1. Requirements Language ............................... 5
2. Terminology .............................................. 5
3. PCECC Modes ............................................. 5
4. PCEP Requirements ....................................... 6
5. Procedures for Using the PCE as the Central Controller
   (PCECC) ................................................ 6
   5.1. Stateful PCE Model ................................... 7
   5.2. New LSP Functions .................................... 7
   5.3. PCECC Capability Advertisement ...................... 8
   5.4. Label Range Reservation .............................. 8
   5.5. PCEP session IP address and TEDB Router ID ......... 9
   5.6. LSP Operations ...................................... 10
       5.6.1. Basic PCECC Mode ............................... 10
           5.6.1.1. PCECC LSP Setup .......................... 10
           5.6.1.2. Label Download .......................... 11
           5.6.1.3. Label Cleanup ............................ 11
           5.6.1.4. PCE Initiated PCECC LSP ................. 12
           5.6.1.5. PCECC LSP Update ........................ 13
           5.6.1.6. PCECC LSP State Report ................... 14
       5.6.2. PCECC Segment Routing (SR) .................... 15
           5.6.2.1. PCECC SR-BE ................................ 15
           5.6.2.2. PCECC SR-TE ............................... 16
6. PCEP messages ............................................ 17
   6.1. The PCLRResv message ............................... 17
   6.2. The PCLabelUpd message ............................. 18
1. Introduction

In certain network deployment scenarios, service providers would like to have the ability to dynamically adapt to a wide range of customer’s requests for the sake of flexible network service delivery. Software Defined Networks (SDN) has provided additional flexibility in how the network is operated compared to the traditional network.

The existing networking ecosystem has become awfully complex and highly demanding in terms of robustness, performance, scalability, flexibility, agility, etc. By migrating to the SDN enabled network from the existing network, service providers and network operators must have a solution which they can evolve easily from the existing network into the SDN enabled network while keeping the network services remain scalable, guarantee robustness and availability etc.

Taking the smooth transition between traditional network and the new SDN enabled network into account, especially from a cost impact assessment perspective, using the existing PCE components from the current network to function as the central controller of the SDN network is one choice, which not only achieves the goal of having a centralized controller, but also leverages the existing PCE network components.
The Path Computation Element communication Protocol (PCEP) provides mechanisms for Path Computation Elements (PCEs) to perform route computations in response to Path Computation Clients (PCCs) requests. PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model [I-D.ietf-pce-stateful-pce] describes a set of extensions to PCEP to enable active control of MPLS-TE and GMPLS tunnels.

[I-D.ietf-pce-pce-initiated-lsp] describes the setup and teardown of PCE-initiated LSPs under the active stateful PCE model, without the need for local configuration on the PCC, thus allowing for a dynamic MPLS network that is centrally controlled and deployed.

[I-D.ietf-pce-remote-initiated-gmpls-lsp] complements [I-D.ietf-pce-pce-initiated-lsp] by addressing the requirements for remote-initiated GMPLS LSPs.

Segment Routing (SR) technology leverages the source routing and tunneling paradigms. A source node can choose a path without relying on hop-by-hop signaling protocols such as LDP or RSVP-TE. Each path is specified as a set of "segments" advertised by link-state routing protocols (IS-IS or OSPF). [I-D.ietf-spring-segment-routing] provides an introduction to SR technology. The corresponding IS-IS and OSPF extensions are specified in [I-D.ietf-isis-segment-routing-extensions] and [I-D.ietf-ospf-segment-routing-extensions], respectively.

A Segment Routed path (SR path) can be derived from an IGP Shortest Path Tree (SPT). Segment Routed Traffic Engineering paths (SR-TE paths) may not follow IGP SPT. Such paths may be chosen by a suitable network planning tool and provisioned on the source node of the SR-TE path.

It is possible to use a stateful PCE for computing one or more SR-TE paths taking into account various constraints and objective functions. Once a path is chosen, the stateful PCE can instantiate an SR-TE path on a PCC using PCEP extensions specified in [I-D.ietf-pce-pce-initiated-lsp] using the SR specific PCEP extensions described in [I-D.ietf-pce-segment-routing].

PCECC may further use PCEP protocol for SR label distribution instead of IGP extensions with some benefits.

Current MPLS label has local meaning. That is, MPLS label is always allocated by the downstream node to the upstream node. Then the MPLS label is only identified by the neighboring upstream node and downstream node. The label allocation is done locally and signaled through the LDP/RSVP-TE/BGP protocol. To ease the label allocation and signaling mechanism, PCE can be conveniently used as a central
controller with Label download capability. Further PCE can also be used to manage the label range and SRGB etc.

The PCECC solution introduced in [I-D.zhao-pce-central-controller-user-cases] allow for a dynamic MPLS network that is eventually controlled and deployed without the deployment of RSVP-TE protocol or extended IGP protocol with node/adjacency segment identifiers signaling capability while providing all the key MPLS functionalities needed by the service providers.

This draft specify the procedures and PCEP protocol extensions for using the PCE as the central controller and user cases where LSPs are calculated/setup/initiated/downloaded through extending the existing PCE architectures and PCEP.

1.1. Requirements Language

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 [RFC2119].

2. Terminology

The following terminology is used in this document.


PCC: Path Computation Client: any client application requesting a path computation to be performed by a Path Computation Element.

PCE: Path Computation Element. An entity (component, application, or network node) that is capable of computing a network path or route based on a network graph and applying computational constraints.

TE: Traffic Engineering.

3. PCECC Modes

The following PCECC modes are supported -

- Basic PCECC.
- PCECC SR.
  - PCECC SR-BE (Best Effort).
*  PCECC SR-TE (Traffic Engineered).

In basic PCECC mode, the forwarding is similar to RSVP-TE signalled LSP without the RSVP-TE signaling. The PCECC allocates and download the label entries along the LSP. The rest of processing is similar to the existing stateful PCE mechanism.

In case of SR, there are two modes for SR-BE and SR-TE. For SR-BE, the forwarding is similar to LDP LSP without LDP signaling or IGP-SR extension. The SR Node label are allocated and distributed in the domain centrally by the PCE via PCEP. Each node (PCC) rely on local IGP for the nexthop calculation. For SR-TE, the forwarding uses label stack similar to IGP based SR-TE without IGP-SR extension. The SR node and adj labels are allocated and distributed in the domain centrally by the PCE via PCEP by PCECC. Rest of the processing is similar to existing stateful PCE with SR mechanism.

4. PCEP Requirements

Following key requirements associated PCECC should be considered when designing the PCECC based solution:

1. PCEP speaker supporting this draft MUST have the capability to advertise its PCECC capability to its peers.

2. Path Computation Client (PCC) supporting this draft MUST have a capability to communicate local label range or global label range or both to PCE.

3. Path Computation Element (PCE) supporting this draft SHOULD have the capability to negotiate a global label range for a group of clients and communicate the final global label range to PCC.

4. PCEP speaker not supporting this draft MUST be able to reject PCECC related message with a reason code that indicates no support for PCECC.

5. PCEP SHOULD provide a means to identify PCECC based LSP in the PCEP messages.

6. PCEP SHOULD provide a means to update (or cleanup) the label-download or label-map entry to the PCC.

5. Procedures for Using the PCE as the Central Controller (PCECC)
5.1. Stateful PCE Model

Active stateful PCE is described in [I-D.ietf-pce-stateful-pce]. PCE as a central controller (PCECC) reuses existing Active stateful PCE mechanism as much as possible to control the LSP.

5.2. New LSP Functions

This document defines the following new PCEP messages and extends the existing messages to support PCECC:

(PCLRResv): a PCEP message sent by a PCC to a PCE to ask for the label range reservation or a PCE to a PCC to send the reserved label range. The PCLRResv message described in Section 6.1.

(PCLabelUpd): a PCEP message sent by a PCE to a PCC to download or cleanup the Label entry. The PCLabelUpd message described in Section 6.2.

(PCRpt): a PCEP message described in [I-D.ietf-pce-stateful-pce]. PCRpt message MAYBE used to send PCECC LSP Reports.

(PCInitiate): a PCEP message described in [I-D.ietf-pce-pce-initiated-lsp]. PCInitiate message is used to setup PCE-Initiated LSP based on PCECC mechanism.

(PCUpd): a PCEP message described in [I-D.ietf-pce-stateful-pce]. PCUpd message is used to send PCECC LSP Update.

The new LSP functions defined in this document are mapped onto the messages as shown in the following table.

<table>
<thead>
<tr>
<th>Function</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCECC Capability advertisement</td>
<td>Open</td>
</tr>
<tr>
<td>Label Range Reservation</td>
<td>PCLRResv</td>
</tr>
<tr>
<td>Label entry Update</td>
<td>PCLabelUpd</td>
</tr>
<tr>
<td>Label entry Cleanup</td>
<td>PCLabelUpd</td>
</tr>
<tr>
<td>PCECC Initiated LSP</td>
<td>PCIInitiate</td>
</tr>
<tr>
<td>PCECC LSP Update</td>
<td>PCUpd</td>
</tr>
<tr>
<td>PCECC LSP State Report</td>
<td>PCRpt</td>
</tr>
<tr>
<td>PCECC LSP Delegation</td>
<td>PCRpt</td>
</tr>
</tbody>
</table>
5.3. PCECC Capability Advertisement

During PCEP Initialization Phase, PCEP Speakers (PCE or PCC) advertise their support of PCECC extensions. A PCEP Speaker includes the "PCECC Capability" TLV, described in Section 7.1.1 of this document, in the OPEN Object to advertise its support for PCECC extensions.

The presence of the PCECC Capability TLV in PCC’s OPEN Object indicates that the PCC is willing to function as a PCECC client.

The presence of the PCECC Capability TLV in PCE’s OPEN message indicates that the PCE is interested in function as a PCECC server.

The PCEP protocol extensions for PCECC MUST NOT be used if one or both PCEP Speakers have not included the PCECC Capability TLV in their respective OPEN message. If the PCEP Speakers support the extensions of this draft but did not advertise this capability then a PCErr message with Error-Type=19 (Invalid Operation) and Error-Value=[TBD] (Attempted LSP setup/download/label-range reservation if PCECC capability was not advertised) will be generated and the PCEP session will be terminated.

L flag and G flag defined in PCECC Capability TLV specifies the local and global label range reservation capability.

A PCC or a PCE MUST include both PCECC-CAPABILITY TLV and STATEFUL-PCE-CAPABILITY TLV in OPEN Object to support the extensions defined in this document. If PCECC-CAPABILITY TLV is advertised and STATEFUL-PCE-CAPABILITY TLV is not advertised in OPEN Object, it SHOULD send a PCErr message with Error-Type=19 (Invalid Operation) and Error-value=[TBD] (stateful pce capability was not advertised) and terminate the session.

5.4. Label Range Reservation

After PCEP initial state synchronization, the label range is reserved.

If L flag is advertised in OPEN Object by PCEP speakers, a PCC reserves a local label range and is communicated using PCLRResv message to a PCE. The PCE maintains the local label range of each node and further during LSP setup, a label is assigned to each node from the corresponding local label range reserved.

If G flag is advertised in OPEN Object by PCEP speakers, a PCC reserves a global label range and is advertised in PCLRResv message to a PCE. The PCE MAY negotiate and reserves the global label range
and also sends the negotiated global label range in PCLRResv message to the PCC. Please refer [I-D.li-mpls-global-label-framework] for MPLS global label allocation.

A PCC MUST send PCLRResv message immediately after the initial LSP synchronization completion. A PCE SHOULD not send PCLLabelUpd message to a PCC before PCLRResv message received. If the PCC received PCLLabelUpd message and not initiated label range reservation, it SHOULD send a PCErr message with Error-type=[TBD] (label range not reserved) and Error-value=[TBD].

The label range reservation sequence is shown below.

```
+-----+        +-----+
|PCC|        |PCE|
+-----+        +-----+

--- PCLRResv, label type=1 ---> local label range reserved
     (100-500)
     label type=2
     (600-1000)

<--- PCLRResv, label type=2 ---
     (700-1000)
```

[Editor’s Note: This section of the document would be updated with more details about Label Block Negotiation, Reservation, Adjustment etc in a future revision of the document.]

5.5. PCEP session IP address and TEDB Router ID

PCE may construct its TEDB by participating in the IGP ([RFC3630] and [RFC5305] for MPLS-TE; [RFC4203] and [RFC5307] for GMPLS). An alternative is offered by BGP-LS [I-D.ietf-idr-ls-distribution] and [I-D.dhodylee-pce-pcep-ls].

PCEP [RFC5440] speaker MAY use any IP address while creating a TCP session. It is important to link the session IP address with the Router ID in TEDB for successful PCECC operations.

During PCEP Initialization Phase, PCC advertise the TEDB mapping information. A PCC includes the "Local Node Descriptors TLV" described in [I-D.dhodylee-pce-pcep-ls], in the OPEN Object for this purpose. Various node descriptor sub-TLV are defined in
which can be used to map the PCEP session to the node in the TEDB.

If this TLV is not present, the session IP address is directly used for the mapping purpose.

5.6. LSP Operations

The PCEP messages pertaining to PCECC MUST include PATH-SETUP-TYPE TLV [I-D.sivabalanc-pce-lsp-setup-type] in the SRP object to clearly identify the PCECC LSP is intended.

5.6.1. Basic PCECC Mode

5.6.1.1. PCECC LSP Setup

In order to setup a LSP based on PCECC mechanism, a PCC MUST delegate the LSP by sending a PCRpt message with Path Setup Type set for basic PCECC (see Section 7.3) and D (Delegate) flag (see [I-D.ietf-pce-stateful-pce]) set in the LSP object.

The LSP-ID in LSP-IDENTIFIER TLV (which usually corresponds to the RSVP-TE LSP-ID) for PCECC LSP MUST always be generated by the PCE. In the first PCRpt message of PCECC LSP, LSP ID of LSP-IDENTIFIER TLV is set to zero.

When a PCE received PCRpt message with P and D flags set, it generates LSP ID; calculates the path and assign labels along the path; and setup the path by sending PCLabelUpd message to each node along the path of the LSP.

The PCE SHOULD send the PCUpd message with the same PLSP-ID to the Ingress PCC in response to the delegate PCRpt message.

The PCECC LSPs MUST be delegated to a PCE at all times.

LSP deletion operation for PCECC LSP is same as defined in [I-D.ietf-pce-stateful-pce]. If the PCE received PCRpt message for LSP deletion then it does Label cleanup operation as described in Section 5.6.1.3 for the corresponding LSP.

The Basic PCECC LSP setup sequence is as shown below.
The PCECC LSP are considered to be ‘up’ by default. The Ingress MAY further choose to deploy a data plane check mechanism and report the status back to the PCE via PCRpt message.

5.6.1.2. Label Download

In order to setup an LSP based on PCECC, the PCE sends a PCLabelUpd message to each node of the LSP to download the Label entry as described in Section 5.6.1.1.

The LSP object in PCLabelUpd MUST include the LSP-IDENTIFIER TLV.

If a node (PCC) received a PCLabelUpd message but failed to download the Label entry, it MUST send a PCErr message with Error-type=[TBD] (label download failed) and Error-value=[TBD].

5.6.1.3. Label Cleanup

In order to delete an LSP based on PCECC, the PCE sends a PCLabelUpd message to each node along the path of the LSP to cleanup the Label entry.

If the PCC received a PCLabelUpd message but does not recognize the label, the PCC MUST generate a PCErr message with Error-Type 19 (Invalid operation) and Error-Value=3, "Unknown Label".

The R flag in SRP object defined in [I-D.ietf-pce-pce-initiated-lsp] specifies the deletion of Label Entry in the PLabelUpd message.

5.6.1.4. PCE Initiated PCECC LSP

The LSP Instantiation operation is same as defined in [I-D.ietf-pce-pce-initiated-lsp].

In order to setup a PCE Initiated LSP based on PCECC mechanism, a PCE sends PCInitiate message with Path Setup Type set for basic PCECC (see Section 7.3) to the Ingress PCC.

The Ingress PCC MUST also set D (Delegate) flag (see [I-D.ietf-pce-stateful-pce]) and C (Create) flag (see [I-D.ietf-pce-pce-initiated-lsp]) in LSP object of PCRpt message. The PCC responds with first PCRpt message with the status as "GOING-UP" and assigned PLSP-ID.

The rest of the PCECC LSP setup operations are same as those described in Section 5.6.1.1.
The LSP deletion operation for PCE Initiated PCECC LSP is same as defined in [I-D.ietf-pce-pce-initiated-lsp]. The PCE should further perform Label entry cleanup operation as described in Section 5.6.1.3 for the corresponding LSP.

The PCE Initiated PCECC LSP setup sequence is shown below.

```
+-------+                           +-------+
|PCC    |                           |  PCE  |
|Ingress|                           +-------+
+------|       |                               |
|  PCC  +-------+                               |
| Transit| |                                   |
+------|        | |<---PCInitiate,PLSP-ID=0,P=1,D=1---| PCECC LSP
|       |     |  (LSP ID=0,GOING-UP)             | Initiate
|       |     |                                   | (LSPID=2
|PCC    | Egress |                         | assigned)
|       |       |<--- PCLabelUpd, PLSP-ID=2 --------| Label
download
|       |       |(LSP ID=2)                       |
|<----- PCLabelUpd, PLSP-ID=2 -----------| Label
download
|       |       |(LSP ID=2)                       |
|<---- PCLabelUpd, PLSP-ID=2 ------- | Label
download
|       |       |(LSP ID=2)                       |
|<-- PCUpd, PLSP-ID=2, P=1, D=1---- | PCECC LSP
|       |       |(LSP ID=2)                       | Update
```

5.6.1.5. PCECC LSP Update

In case of a modification of PCECC LSP with a new path, a PCE sends a PCUpd message to the Ingress PCC.

When a PCC received a PCUpd message for an existing LSP, a PCC MUST follow the make-before-break procedure. On successful traffic switch over to the new LSP, PCC sends a PCRpt message to the PCE for the deletion of old LSP. Further the PCE does cleanup operation for the old LSP described in Section 5.6.1.3.

The PCECC LSP Update and make-before-break sequence is shown below.
The modified PCECC LSP are considered to be 'up' by default. The Ingress MAY further choose to deploy a data plane check mechanism and report the status back to the PCE via PCRpt message.

5.6.1.6. PCECC LSP State Report

As mentioned before, an Ingress PCC MAY choose to apply any OAM mechanism to check the status of LSP in the Data plane and MAY further send its status in PCRpt message to the PCE.
5.6.2. PCECC Segment Routing (SR)

Segment Routing (SR) as described in [I-D.ietf-spring-segment-routing] depends on "segments" that are advertised by Interior Gateway Protocols (IGPs). The SR-node allocate and advertise the SID (node, adj etc) and flood via the IGP. This document proposes a new mechanism where PCE allocate the SID (label) centrally and uses PCEP to advertise the SID. In some deployments PCE (and PCEP) are better suited than IGP because of centralized nature of PCE and direct TCP based PCEP session to the node.

5.6.2.1. PCECC SR-BE

Each node (PCC) is allocated a node-SID (label) by the PCECC. The PCECC sends PCLabelUpd to update the label map of each node to all the nodes in the domain. The Node IP address is determined from the TEDB using the mapping information in the "Local Node Descriptors TLV" described in [I-D.dhodylee-pce-pcep-1s].

It is RECOMMENDED that PCEP session with PCECC SR capability to use a different session IP address during TCP session establishment than the node Router ID in TEDB, to make sure that the PCEP session does not get impacted by the SR-BE node label maps.

On receiving the label map, each node (PCC) uses the local information to determine the next-hop and download the label forwarding instructions accordingly. The PCLabelUpd message in this case MUST not have LSP object but uses new FEC object.
The forwarding behavior and the end result is similar to IGP based "Node-SID" in SR. Thus, from anywhere in the domain, it enforces the ECMP-aware shortest path forwarding of the packet towards the related node.

PCE rely on the Node label cleanup using the same PCLabelUpd message.

5.6.2.2. PCECC SR-TE

A Segment Routed Best Effort path (SR-BE path) can be derived from an IGP Shortest Path Tree (SPT) as explained above. On the other hand, SR-TE paths may not follow IGP SPT. Such paths may be chosen by a PCE and provisioned on the source node of the SR-TE path.

[I-D.ietf-pce-segment-routing] extends PCEP to allow a stateful PCE to compute and initiate SR-TE paths, as well as a PCC to request a path subject to certain constraint(s) and optimization criteria in SR networks.

For SR-TE, apart from node-SID, Adj-SID is used where each adjacency is allocated an Adj-SID (label) by the PCECC. The PCECC sends PCLabelUpd to update the label map of each Adj to the corresponding nodes in the domain. Each node (PCC) download the label forwarding instructions accordingly. Similar to SR-BE, the PCLabelUpd message in this case MUST not have LSP object but uses new FEC object.
The forwarding behavior and the end result is similar to IGP based "Adj-SID" in SR.

The Path Setup Type MUST be set for PCECC SR-TE (see Section 7.3). The rest of the PCEP procedures and mechanism are similar to [I-D.ietf-pce-segment-routing].

PCE rely on the Adj label cleanup using the same PCLabelUpd message.

6. PCEP messages

As defined in [RFC5440], a PCEP message consists of a common header followed by a variable-length body made of a set of objects that can be either mandatory or optional. An object is said to be mandatory in a PCEP message when the object must be included for the message to be considered valid. For each PCEP message type, a set of rules is defined that specify the set of objects that the message can carry. An implementation MUST form the PCEP messages using the object ordering specified in this document.

6.1. The PCLRResv message

A Label Range Reservation message (also referred to as PCLRResv message) is a PCEP message sent by a PCC to a PCE for the reservation of label range or by PCE to PCC to send reserved label range for the network. The Message-Type field of the PCEP common header for the PCLRResv message is set to [TBD].
The format of a PCLRResv message is as follows:

PCLRResv Message::= <Common Header>
                  <label-range>

Where:

<label-range> ::= <SRP>
                 <labelrange-list>

Where
<labelrange-list>::=<LABEL-RANGE>[<labelrange-list>]

There are two mandatory objects that MUST be included within each
<label-range> in the PCLRResv message: the SRP Object and LABEL-RANGE
object.

SRP object is defined in [I-D.ietf-pce-stateful-pce] and this
document extends the use of SRP object in PCLRResv message. If the
SRP object is missing, the receiving PCE MUST send a PCErr message
with Error-type=6 (Mandatory Object missing) and Error-value=10 (SRP
object missing).

PCC generates the value of SRP-ID-number in SRP object of PCLRResv
message send to a PCE. The PCE MUST include the same SRP-ID-number
in SRP object of PCLRResv message sent to the PCC in response to
PCLRResv message.

LABEL-RANGE object is defined in Section 7.2. If the LABEL-RANGE
object is missing, the receiving PCE MUST send a PCErr message with
Error-type=6 (Mandatory Object missing) and Error-value=[TBD] (Label
object missing).

[Editor’s Note: This section of the document would be updated with
more details about Label Block Negotiation, Reservation, Adjustment
etc in a future revision of the document.]

6.2. The PCLLabelUpd message

The Label Update Message (also referred to as PCLLabelUpd) is a PCEP
message sent by a PCE to a PCC to download label or update the label
map. The same message is also used to cleanup the Label entry. The
Message-Type field of the PCEP common header for the PCLLabelUpd
message is set to [TBD].

The format of the PCLLabelUpd message is as follows:
<PCLabelUpd Message> ::= <Common Header>
    <pce-label-update-list>
Where:
<pce-label-update-list> ::= <pce-label-update>
    [<pce-label-update-list>]
<pce-label-update> ::= (<pce-label-download>|<pce-label-map>)
Where:
<pce-label-download> ::= <SRP>
    <LSP>
    <label-list>
<pce-label-map> ::= <SRP>
    <LABEL>
    <FEC>
<label-list > ::=  <LABEL>
    [<label-list>]

The PCLabelUpd message is used to download label along the path of
the LSP for the basic PCECC mode, as well as to update the label map
for the Node and Adjacency Label in case of SR.

The SRP object is defined in [I-D.ietf-pce-stateful-pce] and this
document extends the use of SRP object in PCLabelUpd message. The
SRP object is mandatory and MUST be included in PCLabelUpd message.
If the SRP object is missing, the receiving PCC MUST send a PCErr
message with Error-type=6 (Mandatory Object missing) and Error-
value=10 (SRP object missing).

The LSP object is defined in [I-D.ietf-pce-stateful-pce] and this
document extends the use of LSP object in PCLabelUpd message. The
LSP is an optional object and used in the basic PCECC mode in
PCLabelUpd message. LSP Identifiers TLV is defined in
[I-D.ietf-pce-stateful-pce], it MUST be included in the LSP object in
PCLabelUpd message. If the TLV is missing, the PCC will generate a
PCErr message with Error-Type=6 (mandatory object missing) and Error-
Value=11 (LSP-IDENTIFIERS TLV missing) and close the session.

The LABEL object is defined in Section 7.4. The LABEL is the
mandatory object and MUST be included in PCLabelUpd message. If the
LABEL object is missing, the receiving PCC MUST send a PCErr message
with Error-type=6 (Mandatory Object missing) and Error-value=[TBD]
(LABEL object missing). More than one LABEL object MAY be included
in the PCLabelUpd message for the transit LSR.
The FEC object is defined in Section 7.5. The FEC is an optional object and used in PCECC SR mode in PCLabelUpd message. The FEC object encodes the Node and Adjacency information of the Label Map.

To cleanup the SRF object must set the R (remove) bit.

7. PCEP Objects

The PCEP objects defined in this document are compliant with the PCEP object format defined in [RFC5440]. The P flag and the I flag of the PCEP objects defined in this document MUST always be set to 0 on transmission and MUST be ignored on receipt since these flags are exclusively related to path computation requests.

7.1. OPEN Object

This document defines a new optional TLVs for use in the OPEN Object.

7.1.1. PCECC Capability TLV

The PCECC-CAPABILITY TLV is an optional TLV for use in the OPEN Object for PCECC capability advertisement. Advertisement of the PCECC capability implies support of LSPs that are setup through PCECC as per PCEP extensions defined in this document.

Its format is shown in the following figure:

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               Type=[TBD]      |            Length=4           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                             Flags                         |G|L|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The type of the TLV is [TBD] and it has a fixed length of 4 octets.

The value comprises a single field - Flags (32 bits):

L (LOCAL-LABEL-RANGE-CAPABILITY - 1 bit): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker is capable for local label range reservation.

G (GLOBAL-LABEL-RANGE-CAPABILITY - 1 bit): If set to 1 by a PCEP speaker, it indicates that the PCEP speaker capable for global label range reservation.
Unassigned bits are considered reserved. They MUST be set to 0 on transmission and MUST be ignored on receipt.

7.2. LABEL-RANGE Object

The LABEL-RANGE object MUST be carried within PCLRResv message. The LABEL-RANGE object is used to carry the label range information based on the label type.

LABEL-RANGE Object-Class is TBD.

LABEL-RANGE Object-Type is 1.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  label type  |                 range size                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           label  base                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
//                      Optional TLVs                          //
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

label type (8 bit): The values defined for label type are label type 1 specifies the local label. It means the label range is non negotiable. label type 2 specifies the global label. It means the label range is negotiable. Refer [I-D.li-mpls-global-label-framework] for global label.

Range size (24 bit): It specifies the size of label range.

Label base (32 bit): It specifies the minimum label of label range.

7.3. PATH-SETUP-TYPE TLV

The PATH-SETUP-TYPE TLV is defined in [I-D.sivabalan-pce-lsp-setup-type]; this document defines following new PST value:

- PST = 2: Path is setup via Basic PCECC mode.
- PST = 3: Path is setup via PCECC SR-TE mode.
On a PCRpt or PCInitiate message, the PST=2 in PATH-SETUP-TYPE TLV in SRP object indicates that this LSP was setup via a basic PCECC based mechanism; the PST=3 indicates that this LSP was setup via a PCECC SR-TE based mechanism.

7.4. Label Object

The LABEL Object is used to specify the Label information and MUST be carried within PLabelUpd message.

LABEL Object-Class is TBD.

LABEL Object-Type is 1.

\[
\begin{array}{cccccc}
& 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & A & B & C & D & E & F \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{array}
\]

The fields in the LABEL object are as follows:

Flags: is used to carry any additional information pertaining to the label. Currently, the following flag bit is defined:

* O bit (Out-label): If the bit is set, it specifies the label is the OUT label and it is mandatory to encode the nexthop information (via IPV4-ADDRESS TLV or IPV6-ADDRESS TLV or UNNUMBERED-IPV4-ID-ADDRESS TLV in LABEL object). If the bit is not set, it specifies the label is the IN label and it is optional to encode the local interface information (via IPV4-ADDRESS TLV or IPV6-ADDRESS TLV or UNNUMBERED-IPV4-ID-ADDRESS TLV in LABEL object).

Label (32-bit): The Label information encoded such that the 20 rightmost bits represent a label.
7.4.1. Address TLV

This document defines the following TLV for the LABEL object to associate the next hop information in case of an outgoing label and local interface information in case of an incoming label.

**IPV4-ADDRESS TLV:**

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type=TBD                          | Length = 8 |
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| IPv4 address                                      |
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

**IPV6-ADDRESS TLV:**

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type=TBD                          | Length = 20 |
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| // IPv6 address (16 bytes)              // |
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

**UNNUMBERED-IPV4-ID-ADDRESS TLV:**

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|             Type=TBD                          | Length = 12 |
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Node-ID                                           |
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Interface ID                                       |
+----------+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The address TLVs are as follows:

**IPV4-ADDRESS TLV:** an IPv4 address.

**IPV6-ADDRESS TLV:** an IPv6 address.
7.5. FEC Object

The FEC Object is used to specify the FEC information and MAY be carried within PCLabelUpd message.

FEC Object-Type is 1 ‘IPv4 Node ID’.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
|                      IPv4 Node ID                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```

FEC Object-Type is 2 ‘IPv6 Node ID’.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
| //                      IPv6 Node ID (16 bytes)                // |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```

FEC Object-Type is 3 ‘IPv4 Adjacency’.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
|                      Local IPv4 address                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
|                     Remote IPv4 address                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```

FEC Object-Type is 4 ‘IPv6 Adjacency’.

```
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
| //                      Local IPv6 address (16 bytes)          // |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```
The FEC objects are as follows:

IPv4 Node ID: where IPv4 Node ID is specified as an IPv4 address of the Node. FEC Object-type is 1, and the Object-Length is 4 in this case.

IPv6 Node ID: where IPv6 Node ID is specified as an IPv6 address of the Node. FEC Object-type is 2, and the Object-Length is 16 in this case.

IPv4 Adjacency: where Local and Remote IPv4 address is specified as pair of IPv4 address of the adjacency. FEC Object-type is 3, and the Object-Length is 8 in this case.

IPv6 Adjacency: where Local and Remote IPv6 address is specified as pair of IPv6 address of the adjacency. FEC Object-type is 4, and the Object-Length is 32 in this case.

Unnumbered Adjacency with IPv4 NodeID: where a pair of Node ID / Interface ID tuples is used. FEC Object-type is 5, and the Object-Length is 16 in this case.

8. Security Considerations

TBD
9. Manageability Considerations

9.1. Control of Function and Policy
TBD.

9.2. Information and Data Models
TBD.

9.3. Liveness Detection and Monitoring
TBD.

9.4. Verify Correct Operations
TBD.

9.5. Requirements On Other Protocols
TBD.

TBD.

10. IANA Considerations
TBD

11. Acknowledgments

We would like to thank Robert Tao, Changjing Yan, Tieying Huang for their useful comments and suggestions.

12. References

12.1. Normative References


12.2. Informative References


Filsfils, C., Previdi, S., Decraene, B., Litkowski, S., and r. rjs@rob.sh, "Segment Routing Architecture", draft-ietf-spring-segment-routing-05 (work in progress), September 2015.


Authors’ Addresses

Quintin Zhao
Huawei Technologies
125 Nagog Technology Park
Acton, MA  01719
USA

EMail: quintin.zhao@huawei.com

Katherine Zhao
Huawei Technologies
2330 Central Expressway
Santa Clara, CA  95050
USA

EMail: katherine.zhao@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

EMail: lizhenbin@huawei.com

Dhruv Dhody
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka  560037
India

EMail: dhruv.ietf@gmail.com

Udayasree Palle
Huawei Technologies
Divyashree Techno Park, Whitefield
Bangalore, Karnataka  560037
India

EMail: udayasree.palle@huawei.com
Boris Zhang
Telus Ltd.
Toronto
Canada

EMail: boris.zhang@telus.com