

MPLS Working Group
Internet Draft
Intended status: Standard Track

I. Busi (Ed)
Alcatel-Lucent
H. van Helvoort (Ed)
J. He (Ed)
Huawei

Expires: July 2012

January 11, 2012

MPLS-TP OAM based on Y.1731
draft-bhh-mpls-tp-oam-y1731-08.txt

Abstract

This document describes methods to leverage Y.1731 [2] Protocol Data Units (PDU) and procedures (state machines) to provide a set of Operation, Administration, and Maintenance (OAM) mechanisms that meets the MPLS Transport Profile (MPLS-TP) OAM requirements as defined in [8].

In particular, this document describes the MPLS-TP technology specific encapsulation mechanisms to carry these OAM PDUs within MPLS-TP packets to provide MPLS-TP OAM capabilities in MPLS-TP networks.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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".

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright Notice

Copyright (c) 2012 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.

Table of Contents

1. Introduction.....	4
1.1. Contributing Authors.....	5
2. Conventions used in this document.....	5
2.1. Terminology.....	6
3. Encapsulation of OAM PDU in MPLS-TP.....	6
4. MPLS-TP OAM Packet Formats.....	7
4.1. Continuity Check Message (CCM).....	8
4.1.1. MEG ID Formats.....	9
4.2. OAM Loopback (LBM/LBR).....	9
4.2.1. Format of MEP and MIP ID TLVs.....	12
4.3. Alarm Indication Signal (AIS).....	16
4.4. Lock Reporting (LCK).....	16
4.5. Test (TST).....	17
4.6. Loss Measurement (LMM/LMR).....	17
4.7. One-way delay measurement (1DM).....	17
4.8. Two-way delay Measurement Message/Reply (DM).....	17
4.9. Client Signal Fail (CSF).....	18
5. MPLS-TP OAM Procedures.....	18
5.1. Continuity Check Message (MT-CCM) procedures.....	18
5.2. OAM Loopback (MT-LBM/LBR) procedures.....	20
5.3. Alarm Indication Signal (MT-AIS) procedures.....	21
5.4. Lock Reporting (LCK).....	22
5.5. Test (TST).....	23
5.6. Loss Measurement (LMM/LMR).....	23
5.7. One-way delay measurement (1DM).....	23
5.8. Two-way delay Measurement Message/Reply (DM).....	23
5.9. Client Signal Fail (CSF).....	23
6. Security Considerations.....	23
7. IANA Considerations.....	23
8. Acknowledgments.....	23
9. References.....	25
9.1. Normative References.....	25
9.2. Informative References.....	25

1. Introduction

This document describes the method for leveraging Y.1731 [2] Protocol Data Units (PDUs) and procedures to provide a set of Operation, Administration, and Maintenance (OAM) mechanisms that meet the MPLS Transport Profile (MPLS-TP) OAM requirements as defined in [8].

This version of the draft does not introduce any technical change to the -06 version of this draft.

ITU-T Recommendation Y.1731 [2] specifies:

- o OAM PDUs and procedures that meet the transport networks requirements for OAM
- o Encapsulation mechanisms to carry these OAM PDUs within Ethernet frames to provide Ethernet OAM capabilities in Ethernet networks

Although Y.1731 is focused on Ethernet OAM, the definition of OAM PDUs and procedures are technology independent and can also be used in other packet technologies (e.g., MPLS-TP) provided that the technology specific encapsulation is defined.

The OAM toolset defined in Y.1731 [2] serves as a benchmark for a high performance, comprehensive suite of packet transport OAM capabilities. It can be provided by lightweight protocol design and supports operational simplicity by providing commonality with the established operation models utilized in other transport network technologies (e.g., SDH/SONET and OTN).

This document describes mechanisms for MPLS-TP OAM that reuse the same OAM PDUs and procedures defined in Y.1731 [2], together with the necessary MPLS-TP technology specific encapsulation mechanisms.

The advantages offered by this toolset are summarized below:

- o Simplify the operations for the network operators and service providers that have to test and maintain a single general OAM protocol set when operating LSP, PW and VPLS networks.
- o Accelerate the market adoption of MPLS-TP since Y.1731 is already mature, supported, and deployed.
- o Reduce the complexity and increase the reuse of code for implementation in packet transport devices that may support both

Ethernet and MPLS-TP capabilities, e.g. VPLS and H-VPLS applications.

It is worth noting that multi-vendor interoperable implementations of the OAM mechanisms described in this document already exist to meet the essential OAM requirements for MPLS-TP deployments in PTN applications as described in [9].

Ethernet OAM is also defined by IEEE 802.1ag [14]. IEEE 802.1ag and ITU-T Y.1731 have been developed in cooperation by IEEE and ITU. They support a common subset of OAM functions. ITU-T Y.1731 further extends this common subset with additional OAM mechanisms that are important for the transport network (e.g. AIS, DM, LM).

This document does not deprecate existing MPLS and PW OAM mechanisms nor preclude definition of other MPLS-TP OAM tools.

The mechanisms described in this document, when used to provide MPLS-TP PW OAM functions, are open to support the OAM message mapping procedures defined in [10]. In order to support those procedures, the PEs MUST map the states of the procedures defined in Y.1731 to the PW defect states defined in [10].

The mapping procedures are outside the scope of this document.

In the rest of this document the term "OAM PDU" is used to indicate an OAM PDU whose format and associated procedures are defined in Y.1731 [2] and that this document proposes to be used to provide MPLS-TP OAM functions.

1.1. Contributing Authors

Italo Busi, Huub van Helvoort, Jia He, Christian Addeo, Alessandro D'Alessandro, Simon Delord, John Hoffmans, Ruiquan Jing, Kam Lam, Wang Lei, Han Li, Vishwas Manral, Masahiko Mizutani, Manuel Paul, Josef Roese, Vincenzo Sestito, Yuji Tochio, Munefumi Tsurusawa, Maarten Visser, Rolf Winter

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [1].

2.1. Terminology

ACH	Associated Channel Header
G-ACh	Generic Associated Channel
GAL	G-ACh Label
ME	Maintenance Entity
MEL	MEG Level
MEG	Maintenance Entity Group
MEP	Maintenance End Point
MIP	Maintenance Intermediate Point
PTN	Packet Transport Network
TLV	Type Length Value

3. Encapsulation of OAM PDU in MPLS-TP

Although Y.1731 is focused on Ethernet OAM, the definition of OAM PDUs and procedures are technology independent.

When used to provide Ethernet OAM capabilities, these PDUs are encapsulated into an Ethernet frame where an Ethernet header is prepended to the OAM PDUs.

The MAC DA is used to identify the MEPs and MIPs where the OAM PDU needs to be processed. The EtherType is used to distinguish OAM frames from user data frames.

Within MPLS-TP OAM Framework [6], OAM packets are distinguished from user data packets using the GAL and ACH [5] construct and they are addressed to MEPs or MIPs using existing MPLS forwarding mechanisms (i.e. label stacking and TTL expiration). It is therefore possible to reuse the OAM PDUs defined in [2] within MPLS-TP and encapsulate them within ACH.

A single ACH Channel Type (0xFFFF) is required to identify the presence of Y.1731 OAM PDU. Within the OAM PDU, the OpCode field, defined in [2], allows identifying the specific OAM PDU.

OAM PDUs are encapsulated using the ACH, according to [5], as described in Figure 1 below.

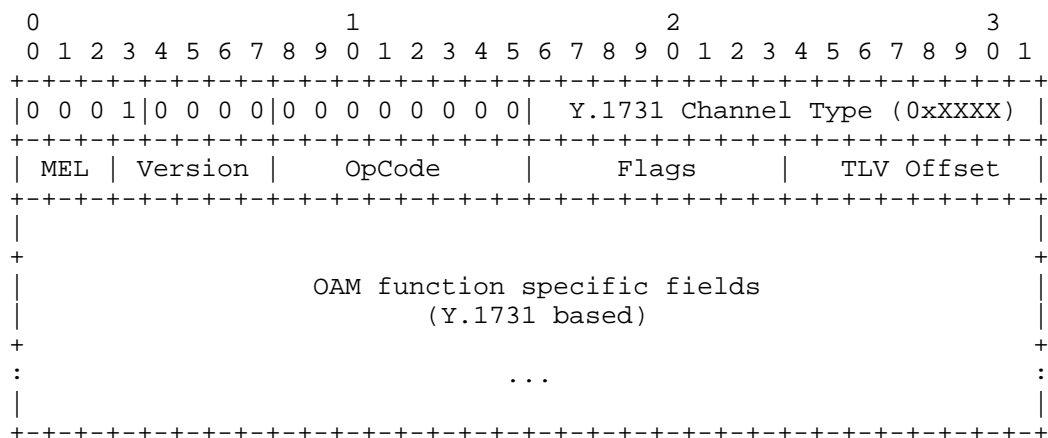


Figure 1 G-ACh Packet carrying a Y.1731 PDU

Moreover, MPLS-TP relies upon a different mechanism for supporting tandem connection monitoring (i.e. label stacking) than the fixed MEL (Maintenance Entity Group Level) field used in Ethernet.

Therefore in MPLS-TP the MEL field is allowed not to be used for supporting tandem connection monitoring.

When OAM PDUs are used in MPLS-TP, the MEL field MUST be set on transmission and checked at reception for compliancy with Y.1731 [2].

The MEL value to set and check MUST be configurable. The DEFAULT value MUST be "111". With co-routed bidirectional transport paths, the configured MEL MUST be the same in both directions.

The OpCode field identifies the type of the OAM PDU.

The setting of the Version, Flags and TLV Offset is OpCode specific and described in Y.1731 [2].

4. MPLS-TP OAM Packet Formats

This section describes the OAM functions that can be supported reusing the OAM PDUs and procedures defined in Y.1731 [2] to meet MPLS-TP OAM Requirements, as defined in [8].

This document is proposing not to use the Y.1731 MCC OAM PDU in MPLS-TP. The solution proposed in [7], where MCC PDU is directly encapsulated within an ACH with a PID, SHOULD be used instead.

The LTM/LTR OAM PDUs, as currently defined Y.1731 [2], are tracing the path for a specific MAC address: this tool is therefore addressing a different requirement than the "Route Tracing" functional requirement described in section 2.2.4 of RFC 5860 [8]. Their purpose is to test the MAC Address Forwarding tables. Due to the fact that MPLS-TP forwarding is not based on the MAC Address Forwarding tables, these tools are not applicable to MPLS-TP as currently defined.

Procedures for supporting the route tracing MPLS-TP OAM functional requirement (section 2.2.4 of RFC 5860 [8]) are outside the scope of this document.

4.1. Continuity Check Message (CCM)

The CCM PDU is defined in Y.1731 [2]. When encapsulated within MPLS-TP as described in section 3, it can be used to support the following MPLS-TP OAM functional requirements:

- o Pro-active continuity check (section 2.2.2 of RFC 5860 [8]);
- o Pro-active connectivity verification (section 2.2.3 of RFC 5860 [8]);
- o Pro-active remote defect indication (section 2.2.9 of RFC 5860 [8]);
- o Pro-active packet loss measurement (section 2.2.11 of RFC 5860 [8]).

Procedures for transmitting and receiving CCM PDUs are defined in Y.1731 [2] and described in section 5.1.

It is worth noting that the use of CCM does not require any additional status information other than the configuration parameters and defect states.

The transmission period of the CCM MUST always be the configured period and MUST not change unless the operator reconfigures it. This is a fundamental requirement to allow deterministic and predictable

protocol behavior: in transport networks the operator configures and fully controls the repetition rate of pro-active CC-V.

In order to perform pro-active Connectivity Verification, the CCM packet contains a globally unique identifier of the source MEP, as described in [6].

The source MEP for LSPs, PWs and Sections is identified by combining a globally unique MEG ID (see section 4.1.1) with a MEP ID that is unique within the scope of the Maintenance Entity Group.

4.1.1. MEG ID Formats

The generic format for MEG ID is defined in Figure A-1 of Y.1731 [2]. Different formats of MEG ID are allowed: the MEG ID format type is identified by the MEG ID Format field.

The format of the ICC-based MEG ID is defined in Annex A of Y.1731 [2]. This format is applicable to MPLS-TP Sections, LSPs and PWs.

MPLS-TP supports also IP-based format for MEG ID. These formats are still under definition in [12] and therefore outside the scope of this document.

4.2. OAM Loopback (LBM/LBR)

The LBM/LBR PDUs, defined in Y.1731 [2]. When encapsulated within MPLS-TP, as described in section 3, they can be used to support the following MPLS-TP OAM functional requirements:

- o On-demand bidirectional connectivity verification (section 2.2.3 of RFC 5860 [8]);
- o Bidirectional in-service or out-of-service diagnostic test (section 2.2.5 of RFC 5860 [8]).

Procedures for transmitting and receiving LBM/LBR PDUs are defined in Y.1731 [2] and described in section 5.2.

It is worth noticing that these OAM PDUs cover different functions than those defined in [11].

When the LBM/LBR is used for out-of-service diagnostic test, it is REQUIRED that the transport path is locked on both MEPs before the diagnostic test is performed. In transport networks, the transport

path is locked on both sides by network management operations. However, single-ended procedures as defined in [11] MAY be used.

In order to allow proper identification of the target MEP/MIP the LBM is addressed to, the LBM PDU MUST include the Target MEP/MIP ID TLV: this TLV MUST be present in an LBM PDU and MUST be located at the top of the TLVs (i.e., it MUST start at the offset indicated by the TLV Offset field).

A LBM packet with the Target MIP/MEP ID equal to the ID of receiving MIP or MEP is considered to be a valid LBM packet. Every field in the LBM packet is copied to the LBR packet, only the OpCode field is changed from LBM to LBR.

To allow proper identification of the actual MEP/MIP that has replied to an LBM PDU, the LBR PDU MUST include the Replying MEP/MIP ID TLV: this TLV MUST be present in an LBR PDU and it MUST be located at the top of the TLVs (i.e., it MUST start at the offset indicated by the TLV Offset field).

In order to simplify hardware based implementations, these TLVs have been defined to have a fixed position (as indicated by the TLV Offset field) and a fixed length (see clause 4.2.1).

It is worth noting that the MEP/MIP identifiers used in the Target MEP/MIP ID and in the Replying MEP/MIP ID TLVs SHOULD be unique within the scope of the MEG. When LBM/LBR OAM is used for connectivity verification purposes, there are some misconnectivity cases that could not be easily located by simply relying upon these TLVs. In order to locate these misconnectivity configurations, the LBM PDU SHOULD carry a Requesting MEP ID TLV that provides a globally unique identification of the MEP that has originated the LBM PDU. When the Requesting MEP ID TLV is present in the LBM PDU, the replying MIP/MEP MUST check that the received requesting MEP identifier matches with the expected requesting MEP identifier before replying. In this case, the LBR PDU MUST carry the Requesting MEP ID TLV confirming to the MEP the LBR PDU is sent to that the Requesting MEP ID TLV in the LBM PDU has been checked before replying.

When LBM/LBR OAM is used for bidirectional diagnostic tests, the Requesting MEP ID TLVs MUST NOT be included.

The format of the LBM and LBR PDUs are shown in Figure 2 and in Figure 3.

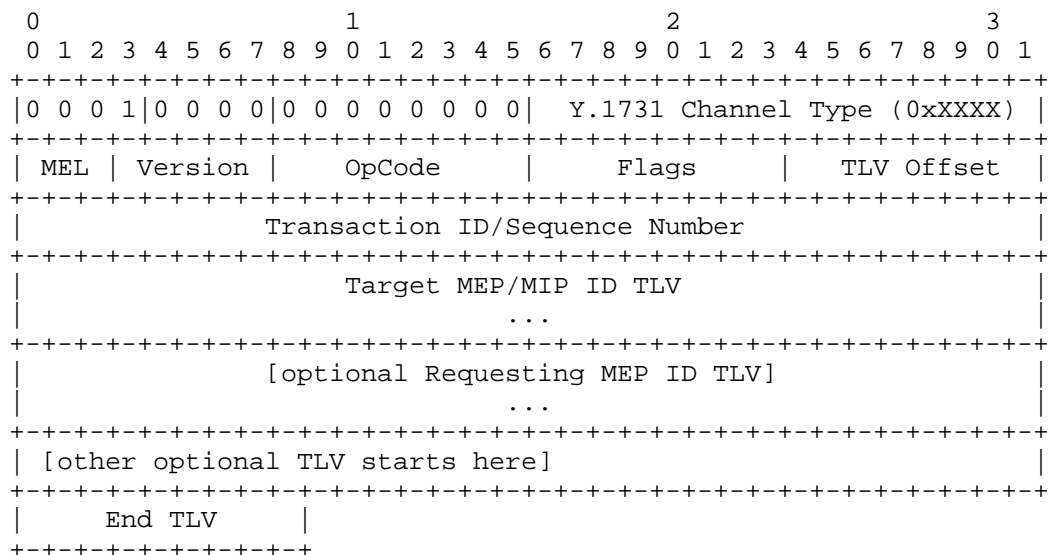


Figure 2 LBM Packet Format

The OpCode MUST be set to 0x03 (LBM). The TLV Offset MUST be set to 0x04. The formats of the Target MEP/MIP ID TLV and of the Requesting MEP ID TLV are defined in 4.2.1.

The Target MEP/MIP ID MUST be always present as the first TLV within the LBM PDU. When present, the Requesting MEP ID TLV MUST immediately follow the Target MEP/MIP ID TLV.

When the LBM packet is sent to a target MIP, the source MEP MUST know the hop count to the target MIP and set the TTL field accordingly, as described in [6].

This solution allows supporting per-node and per-interface MIP implementations as described in section 3.4 of [6]:

- o In the case of a per-node MIP implementation, the LBM packet is processed in the per-node MIP if the Target MEP/MIP ID matches the per-node MIP identifier; otherwise, the LBM packet is dropped;

- o In the case of a per-interface MIP implementation, the LBM packet is processed in the ingress MIP if the Target MEP/MIP ID matches the ingress MIP identifier; otherwise, the LBM packet is forwarded to the egress port(s) together (i.e., fate sharing) with the user data packets. The LBM packet is processed in the egress MIP if the Target MEP/MIP ID matches the egress MIP identifier; otherwise, the LBM packet is dropped.

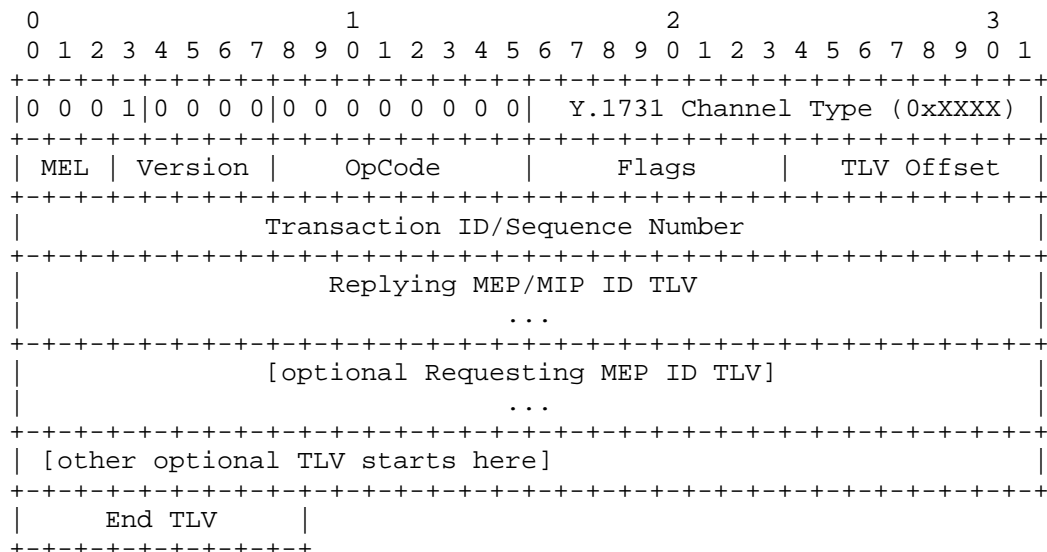


Figure 3 LBR Packet Format

The Replying MEP/MIP ID TLV MUST be present as the first TLV within the LBR PDU. When present, the Requesting MEP ID TLV MUST follow the Replying MEP/MIP ID TLV within the LBR PDU.

4.2.1. Format of MEP and MIP ID TLVs

The format of the Target and Replying MIP/MEP ID TLVs are shown in Figure 4 and Figure 5.

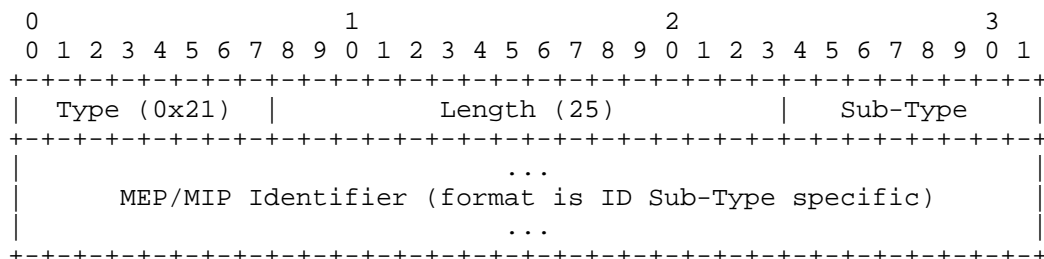


Figure 4 Target MEP/MIP ID TLV format

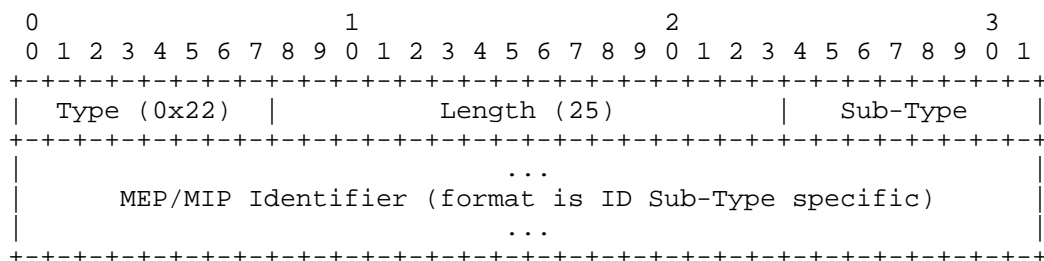


Figure 5 Replying MEP/MIP ID TLV format

Different formats of MEP/MIP identifiers MAY be used: the format type is described by the MEP/MIP ID Sub-Type field.

The "Discovery ingress/node MEP/MIP" and the "Discovery egress MEP/MIP" identifiers MAY only be used within the LBM PDU (and MUST NOT appear in an LBR PDU) for discovering the identifiers of the MEPs or of the MIPs located at a given TTL distance from the MEP originating the LBM PDU.

The format of the Target MEP/MIP ID TLV carrying a "Discovery ingress/node MEP/MIP" is shown in Figure 6.

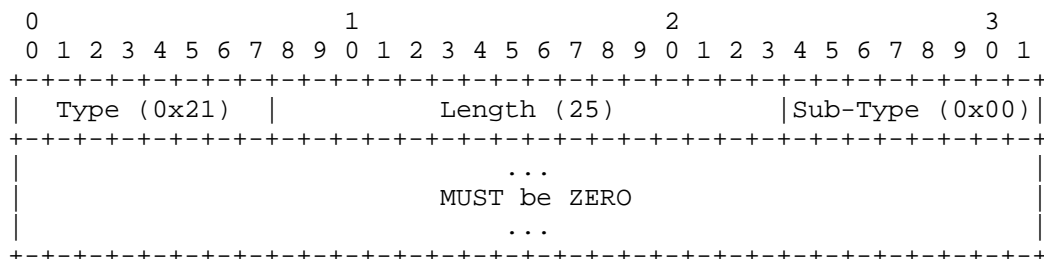


Figure 6 Target MEP/MIP ID TLV format (discovery ingress/node MEP/MIP)

The format of the Target MEP/MIP ID TLV carrying a "Discovery egress MEP/MIP" is shown in Figure 7.

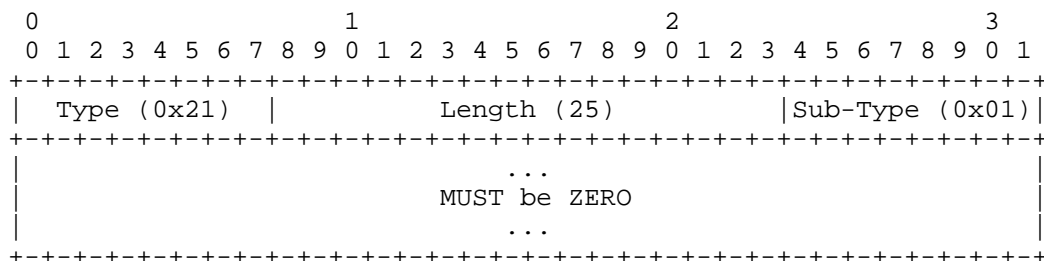


Figure 7 Target MEP/MIP ID TLV format (discovery egress MEP/MIP)

The format of the Target or Replying MEP/MIP ID TLV carrying an "ICC-based MEP ID" is shown in Figure 8.

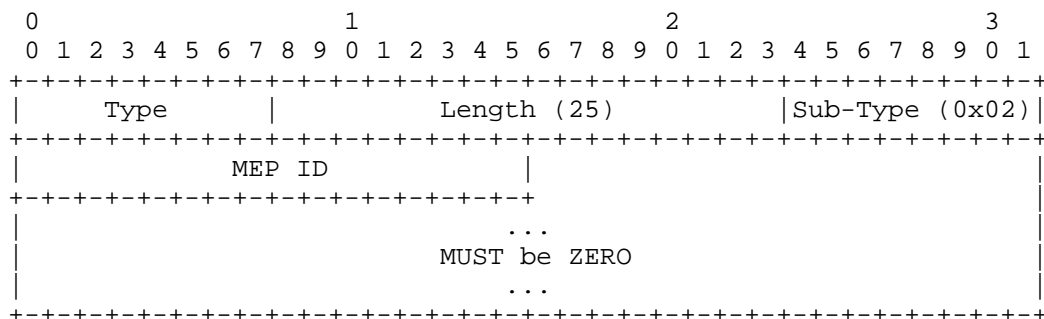


Figure 8 Target or Replying MEP/MIP ID TLV format (ICC-based MEP ID)

The MEP ID is a 16-bit integer value identifying the transmitting MEP within the MEG.

The format of the Target or Replying MEP/MIP ID TLV carrying an "ICC-based MIP ID" is shown in Figure 9.

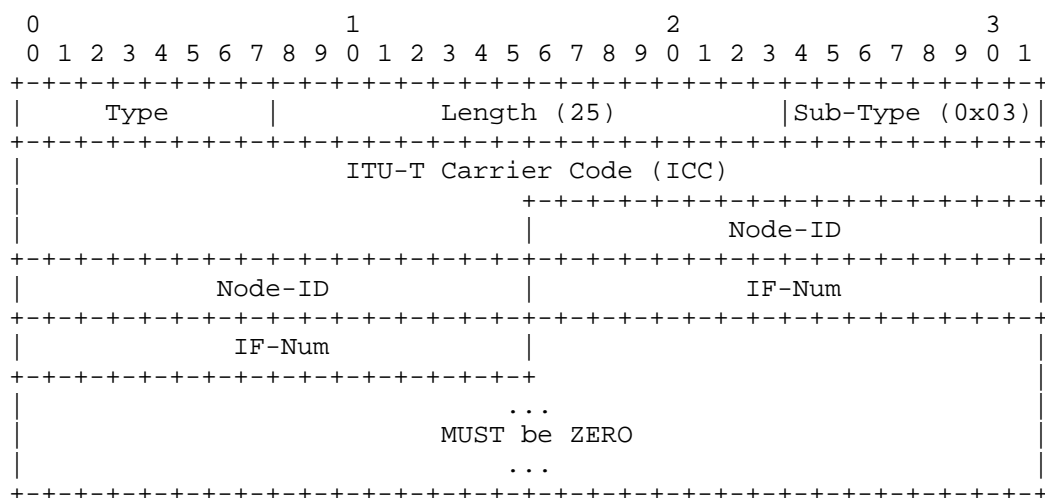


Figure 9 Target or Replying MEP/MIP ID TLV format (ICC-based MIP ID)

The ITU-T Carrier Code (ICC) is a code assigned to a network operator/service provider and maintained by the ITU-T Telecommunication Standardization Bureau (TSB) as per [13].

The Node-ID is a numeric identifier of the node where the MIP is located. Its assignment is a matter for the organization to which the ICC has been assigned, provided that uniqueness within that organization is guaranteed.

The IF-Num is a numeric identifier of the Access Point (AP) toward the server layer trail, which can be either an MPLS-TP or a non MPLS-TP server layer, where a per-interface MIP is located. Its assignment is a matter for the node the MIP is located, provided that uniqueness within that node is guaranteed. Note that the value 0 for IF-Num is reserved to identify per-node MIPs.

MPLS-TP supports also IP-based format for MIP and MEP identifiers. These formats are still under definition in [12] and therefore outside the scope of this document.

The format of the Requesting MEP ID TLVs is shown in Figure 10.

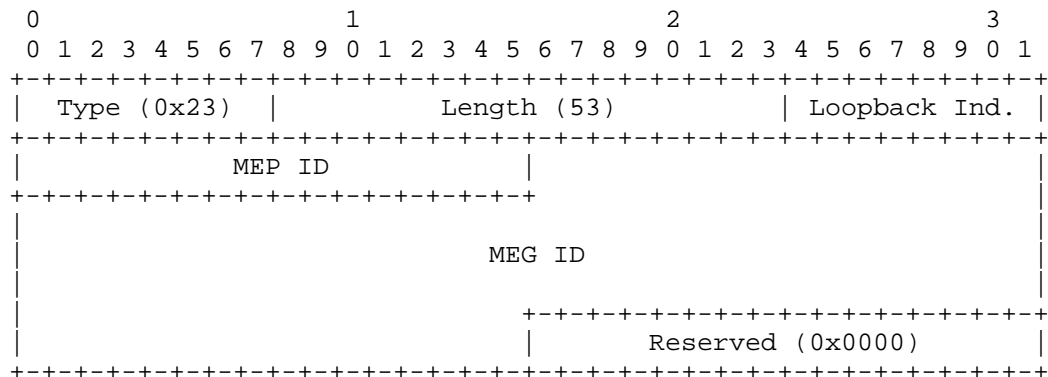


Figure 10 Requesting MEP ID TLV format

The MEP ID and MEG ID carry the globally unique MEP ID as defined in section 4.1.1.

The Reserved bits MUST be set to all-ZEROes in transmission and ignored in reception.

The Loopback Indication MUST be set to 0x0000 when this TLV is inserted in an LBM PDU and SHOULD be set to 0x0001 in the LBR PDU. This is used to indicate that the value of this TLV has been checked by the node that generated the LBR PDU.

4.3. Alarm Indication Signal (AIS)

The AIS PDU is defined in Y.1731 [2]. When encapsulated within MPLS-TP, as described in section 3, it can be used to support the alarm reporting MPLS-TP OAM functional requirement (section 2.2.8 of RFC 5860 [8]).

Procedures for transmitting and receiving AIS PDUs are defined in Y.1731 [2] and described in section 5.3.

4.4. Lock Reporting (LCK)

The LCK PDU is defined in Y.1731 [2]. When encapsulated within MPLS-TP, as described in section 3, it can be used to support the lock reporting MPLS-TP OAM functional requirement (section 2.2.7 of RFC 5860 [8]).

Procedures for transmitting and receiving LCK PDUs are defined in Y.1731 [2] and described in section 5.4.

4.5. Test (TST)

The TST PDU is defined in Y.1731 [2]. When encapsulated within MPLS-TP, as described in section 3, it can be used to support the uni-directional in-service or out-of-service diagnostic tests MPLS-TP OAM functional requirement (section 2.2.8 of RFC 5860 [8]).

Procedures for transmitting and receiving TST PDUs are defined in Y.1731 [2] and described in section 5.5.

4.6. Loss Measurement (LMM/LMR)

The LMM/LMR PDUs are defined in Y.1731 [2]. When encapsulated within MPLS-TP, as described in section 3, they can be used to support on-demand packet loss measurement MPLS-TP OAM functional requirement (section 2.2.11 of RFC 5860 [8]).

Procedures for transmitting and receiving LMM/LMR PDUs are defined in Y.1731 [2] and described in section 5.6.

4.7. One-way delay measurement (1DM)

The 1DM PDU is defined in Y.1731 [2]. When encapsulated within MPLS-TP, as described in section 3, it can be used to support the on-demand one-way packet delay measurement MPLS-TP OAM functional requirement (section 2.2.12 of RFC 5860 [8]).

It can also be used to support proactive one-way delay measurement MPLS-TP OAM functional requirement (section 2.2.12 of RFC 5860 [8]).

Procedures for transmitting and receiving 1DM PDUs are defined in Y.1731 [2] and described in section 5.7.

4.8. Two-way delay Measurement Message/Reply (DM)

The DMM/DMR PDUs are defined in Y.1731 [2]. When encapsulated within MPLS-TP, as described in section 3, they can be used to support on-demand two-ways packet delay measurement MPLS-TP OAM functional requirement (section 2.2.12 of RFC 5860 [8]).

They can also be used to support proactive two-ways packet delay measurement MPLS-TP OAM functional requirement (section 2.2.12 of RFC 5860 [8]).

Procedures for transmitting and receiving DMM/DMR PDUs are defined in Y.1731 [2] and described in section 5.8.

4.9. Client Signal Fail (CSF)

The CSF PDU is defined in Y.1731 Amendment 1 [3]. When encapsulated within MPLS-TP, as described in section 3, it can be used to support the client failure indication MPLS-TP OAM functional requirement (section 2.2.10 of RFC 5860 [8]).

Procedures for transmitting and receiving CSF PDUs are defined in Y.1731 Amendment 1 [3] and described in section 5.9.

5. MPLS-TP OAM Procedures

The high level procedures for processing Y.1731 OAM PDUs are described in [2] and [3]. The technology independent procedures are also applicable to MPLS-TP OAM.

More detailed and formal procedures for processing Y.1731 OAM PDUs are defined in G.8021 [4]. Although the description in [4] is Ethernet-specific, the technology independent procedures are also applicable to MPLS-TP OAM.

This section describes the MPLS-TP OAM procedures based on the technology independent ones defined in [2], [3] and [4].

5.1. Continuity Check Message (MT-CCM) procedures

The MT-CCM PDU format is defined in section 4.1.

When CCM generation is enabled, the MEP MUST generate CCM OAM packets with the periodicity and the PHB configured by the operator:

- o MEL field MUST be set to the configured value (see section 3);
- o Version field MUST be set to 0 (see section 3);
- o OpCode field MUST be set to 0x01 (see section 4.1);

- o RDI flag MUST be set, if the MEP asserts signal file. Otherwise, it MUST be cleared;
- o Reserved flags MUST be set to 0 (see section 4.1);
- o Period field MUST be set according to the configured periodicity (see Table 9-3 of [2]);
- o TLV Offset field MUST be set to 70 (see section 4.1);
- o Sequence Number MUST be set to 0 (see section 4.1);
- o MEP ID and MEG ID fields MUST carry the configured values;
- o The TxFCf field MUST carry the current value of the counter for in-profile data packets transmitted towards the peer MEP, when pro-active loss measurement is enabled. Otherwise it MUST be set to 0.
- o The RxFCb field MUST carry the current value of the counter for in-profile data packets received from the peer MEP, if pro-active loss measurement is enabled. Otherwise it MUST be set to 0.
- o The TxFCb field MUST carry the value of TxFCf of the last received CCM PDU from the peer MEP, if pro active loss measurement is enabled. Otherwise it MUST be set to 0.
- o Reserved field MUST be set to 0 (see section 4.1);
- o End TLV MUST be inserted after the Reserved field (see section 4.1).

The transmission period of the CCM is always the configured period and does not change unless the operator reconfigures it.

When a MEP receives a CCM OAM packet, it checks the various fields (see Figure 8-19 of [4]). The following defects are detected as described in clause 6.1 of [4]: dLOC, dUNL, dMMG, dUNM, dUNP, dUNPr and dRDI.

If the Version, MEL, MEG and MEP fields are valid and pro-active loss measurement is enabled, the values of the packet counters are processed as described in clause 8.1.7.4 of [4].

5.2. OAM Loopback (MT-LBM/LBR) procedures

The MT-LBM/LBR PDU formats are defined in section 4.2.

When an out-of-service OAM loopback function is performed, client data traffic is disrupted in the diagnosed ME. The MEP configured for the out-of-service test MUST transmit MT-LCK packets in the immediate client (sub-)layer, as described in section 5.4.

When an in-service OAM loopback function is performed, client data traffic is not disrupted and the packets with MT-LBM/LBR information are transmitted in such a manner that a limited part of the service bandwidth is utilized. The periodicity for packets with MT-LBM/LBR information is pre-determined.

When on-demand OAM loopback is enabled at a MEP, the (requesting) MEP MUST generate and send to one of the MIPs or the peer MEP MT-LBM OAM packets with the periodicity and the PHB configured by the operator:

- o MEL field MUST be set to the configured value (see section 3);
- o Version field MUST be set to 0 (see section 3);
- o OpCode field MUST be set to 0x03 (see section 4.2);
- o Flags field MUST be set to all-ZEROes (see section 4.2);
- o TLV Offset field MUST be set to 4 (see section 4.2);
- o Transaction field is a 4-octet field that contains the transaction ID/sequence number for the loop-back measurement;
- o Target MEP/MIP-ID and Originator MEP-ID fields are set to carry the configured values;
- o Optional TLV field whose length and contents are configurable at the requesting MEP. The contents can be a test pattern and an optional checksum. Examples of test patterns include pseudo-random bit sequence (PRBS) ($2^{31}-1$) as specified in sub-clause 5.8/O.150, all '0' pattern, etc. For bidirectional diagnostic test application, configuration is required for a test signal generator and a test signal detector associated with the MEP;
- o End TLV field is set to all-ZEROes (see section 4.2).

Whenever a valid MT-LBM packet is received by a (receiving) MIP or a (receiving) MEP, an MT-LBR packet is generated and transmitted by the receiving MIP/MEP to the requesting MEP:

- o MEL field MUST be copied from the received MT-LBM PDU;
- o Version field MUST be copied from the received MT-LBM PDU;
- o OpCode field MUST be set to 2 (see section 4.2);
- o Flags field MUST be copied from the received MT-LBM PDU;
- o TLV Offset field MUST be copied from the received MT-LBM PDU;
- o Transaction field MUST be copied from the received MT-LBM PDU;
- o The Target MEP/MIP-ID and Originator MEP-ID fields are set to the value which is copied from the last received MT-LBM PDU;
- o The Optional TLV field MUST be copied from the received MT-LBM PDU;
- o End TLV field MUST be inserted after the last TLV field and it MUST be copied from the last received MT-LBM PDU.

5.3. Alarm Indication Signal (MT-AIS) procedures

The MT-AIS PDU format is described in section 4.3.

When the server layer trail termination sink asserts signal fail, it notifies the server/MT_A_Sk function that raises the aAIS consequent action. The aAIS is cleared when the server layer trail termination clears the signal fail condition and notifies the server/MT_A_Sk.

When the aAIS consequent action is raised, the server/MT_A_Sk MUST continuously generate MPLS-TP OAM packets carrying the AIS PDU until the aAIS consequent action is cleared:

- o MEL field MUST be set to the configured value (see section 3):
- o Version field MUST be set to 0 (see section 3):
- o OpCode MUST be set to 0x21 (see section 4.3):
- o Reserved flags MUST be set to 0 (see section 4.3):

- o Period field MUST be set according to the configure periodicity (see Table 9-4 of [2]);
- o TLV Offset MUST be set to 0 (see section 4.3):
- o End TLV MUST be inserted after the TLV Offset field (see section 4.3).

The DEFAULT periodicity for MT-AIS is once per second.

The generated AIS packets MUST be inserted in the incoming stream, i.e., the output stream contains the incoming packets and the generated AIS packets.

When a MEP receives an AIS packet with the correct MEL value, it MUST detect the dAIS defect as described in clause 6.1 of [4].

5.4. Lock Reporting (LCK)

The MT-LCK PDU format is described in section 4.4.

When the access to the server layer trail is administratively locked by the operator, the server/MT_A_So and server/MT_A_Sk functions raise the aLCK consequent action. The aLCK is cleared when the access to the server layer trail is administratively unlocked.

When the aLCK consequent action is raised, the server/MT_A_So and server/MT_A_Sk MUST continuously generate, on both directions, MPLS-TP OAM packets carrying the LCK PDU until the aLCK consequent action is cleared:

- o MEL field MUST be set to the configured value (see section 3):
- o Version field MUST be set to 0 (see section 3):
- o OpCode MUST be set to 0x23 (see section 4.4):
- o Reserved flags MUST be set to 0 (see section 4.4):
- o Period field MUST be set according to the configure periodicity (see Table 9-4 of [2]);
- o TLV Offset MUST be set to 0 (see section 4.4):

- o End TLV MUST be inserted after the TLV Offset field (see section 4.4).

The DEFAULT periodicity for MT-LCK is once per second.

When a MEP receives an LCK packet with the correct MEL value, it detects the dLCK defect as described in clause 6.1 of [4].

5.5. Test (TST)

5.6. Loss Measurement (LMM/LMR)

5.7. One-way delay measurement (1DM)

5.8. Two-way delay Measurement Message/Reply (DM)

5.9. Client Signal Fail (CSF)

6. Security Considerations

Spurious OAM messages, such as those defined in this document, potentially could form a vector for a denial of service attack. However, since these messages are carried in a control channel, one would have to gain access to a node providing the service in order to launch such an attack. Since transport networks are usually operated as a walled garden, such threats are less likely.

7. IANA Considerations

IANA is requested to allocate a Channel Type value 0xXXXX to identify an associated channel carrying all the OAM PDUs that are defined in section 4

[Editor's note - The value 0x8902 has been proposed to keep the channel type identical to the EtherType value used in Ethernet OAM]

8. Acknowledgments

The authors gratefully acknowledge the contributions of Malcolm Betts, Zhenlong Cui, Feng Huang, Kam Lam, Jian Yang, Haiyan Zhang for the definition of extensions to LBM/LBR required for supporting on-demand connectivity verification OAM functions.

The authors would like to thank all the members of the CCSA for their comments and support.

The authors would also like to thank Brian Branscomb, Feng Huang, Kam Lam, Fang Li, Akira Sakurai and Yaakov Stein for their comments and enhancements to the text.

This document was prepared using 2-Word-v2.0.template.dot.

9. References

9.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [2] ITU-T Recommendation Y.1731 (02/08), "OAM functions and mechanisms for Ethernet based networks", February 2008
- [3] ITU-T Recommendation Y.1731 Amendment 1 (07/10), "OAM functions and mechanisms for Ethernet based networks", July 2010
- [4] ITU-T Recommendation G.8021 (12/07), "Characteristics of Ethernet transport network equipment functional blocks", December 2007
- [5] Vigoureux, M., Bocci, M., Swallow, G., Ward, D., Aggarwal, R., "MPLS Generic Associated Channel", RFC 5586, June 2009
- [6] Busi, I., Allan, D., " Operations, Administration and Maintenance Framework for MPLS-based Transport Networks", draft-ietf-mpls-tp-oam-framework-11 (work in progress), February 2011
- [7] Beller, D., Farrel, A., "An Inband Data Communication Network For the MPLS Transport Profile", RFC 5718, January 2010

9.2. Informative References

- [8] Vigoureux, M., Betts, M., Ward, D., "Requirements for OAM in MPLS Transport Networks", RFC 5860, May 2010
- [9] Li, F., Li, H., D'Alessandro, A., Jing, R., Wang, G., "Operator Considerations on MPLS-TP OAM Mechanisms", draft-fang-mpls-tp-oam-considerations-02 (work in progress), July 2011
- [10] Nadeau, T., et al., "Pseudo Wire (PW) OAM Message Mapping", draft-ietf-pwe3-oam-msg-map-16 (work in progress), April 2011
- [11] Boutros, S., et al., "Operating MPLS Transport Profile LSP in Loopback Mode", draft-ietf-mpls-tp-li-lb-02 (work in progress), June 2011

- [12] Swallow, G., Bocci, M., " MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-02 (work in progress), July 2010
- [13] ITU-T Recommendation M.1400 (07/06), " Designations for interconnections among operators' networks", July 2006
- [14] IEEE Standard 802.1ag-2007, "IEEE Standard for Local and Metropolitan Area Networks: Connectivity Fault Management", September 2007

Author's Addresses

Italo Busi (Editor)
Alcatel-Lucent

Email: Italo.Busi@alcatel-lucent.com

Huub van Helvoort (Editor)
Huawei Technologies

Email: hhelvoort@huawei.com

Jia He (Editor)
Huawei Technologies

Email: hejia@huawei.com

Contributing Authors' Addresses

Christian Addeo
Alcatel-Lucent

Email: Christian.Addeo@alcatel-lucent.com

Alessandro D'Alessandro
Telecom Italia

Email: alessandro.dalessandro@telecomitalia.it

Simon Delord
Telstra

Email: simon.a.delord@team.telstra.com

John Hoffmans
KPN

Email: john.hoffmans@kpn.com

Ruiquan Jing
China Telecom

Email: jingrq@ctbri.com.cn

Hing-Kam (Kam) Lam
Alcatel-Lucent

Email: Kam.Lam@alcatel-lucent.com

Wang Lei
China Mobile Communications Corporation

Email: wangleiyj@chinamobile.com

Han Li
China Mobile Communications Corporation

Email: lihan@chinamobile.com

Vishwas Manral
IPInfusion Inc

Email: vishwas@ipinfusion.com

Masahiko Mizutani
Hitachi, Ltd.

Email: masahiko.mizutani.ew@hitachi.com

Manuel Paul
Deutsche Telekom

Email: Manuel.Paul@telekom.de

Josef Roese
Deutsche Telekom

Email: Josef.Roese@t-systems.com

Vincenzo Sestito
Alcatel-Lucent

Email: vincenzo.sestito@alcatel-lucent.com

Yuji Tochio
Fujitsu

Email: tochio@jp.fujitsu.com

Munefumi Tsurusawa
KDDI R&D Labs

Email: tsuru@kddilabs.jp

Maarten Visser
Huawei Technologies

Email: maarten.vissers@huawei.com

Rolf Winter
NEC

Email: Rolf.Winter@nw.neclab.eu

Network Working Group
Internet Draft
Intended status: Informational
Expires: April 30, 2012

L. Fang, Ed.
Cisco Systems
N. Bitar
Verizon
R. Zhang
Alcatel Lucent
M. DAIKOKU
KDDI
P. Pan
Infinera

October 31, 2011

MPLS-TP Use Cases Studies and Design Considerations
draft-fang-mpls-tp-use-cases-and-design-04.txt

Abstract

This document provides use case studies and network design considerations for Multiprotocol Label Switching Transport Profile (MPLS-TP).

In the recent years, MPLS-TP has emerged as the technology of choice for the new generation of packet transport. Many service providers (SPs) are working to replace the legacy transport technologies, e.g. SONET/SDH, TDM, and ATM technologies, with MPLS-TP for packet transport, in order to achieve higher efficiency, lower operational cost, while maintaining transport characteristics.

The use cases for MPLS-TP include Metro Ethernet access and aggregation, Mobile backhaul, and packet optical transport. The design considerations discussed in this documents ranging from operational experience; standards compliance; technology maturity; end-to-end forwarding and OAM consistency; compatibility with IP/MPLS networks; multi-vendor interoperability; and optimization vs. simplicity design trade off discussion. The general design principle is to provide reliable, manageable, and scalable transport solutions.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

MPLS-TP Use Case and Design Considerations
Expires April 2012

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 30, 2012.

Copyright Notice

Copyright (c) 2011 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 and Motivation	3
1.2. Co-authors and contributors	5
2. Terminologies	5
3. Overview of MPLS-TP base functions	6
3.1. MPLS-TP development principles	6
3.2. Data Plane	7
3.3. Control Plane	7
3.4. OAM	7
3.5. Survivability	8
4. MPLS-TP Use Case Studies	8

4.1. Metro Access and Aggregation	8
---	---

MPLS-TP Use Case and Design Considerations
Expires April 2012

4.2. Packet Optical Transport	9
4.3. Mobile Backhaul	10
5. Network Design Considerations	11
5.1. IP/MPLS vs. MPLS-TP	11
5.2. Standards compliance	12
5.3. End-to-end MPLS OAM consistency	13
5.4. PW Design considerations in MPLS-TP networks	13
5.5. Proactive and event driven MPLS-TP OAM tools	14
5.6. MPLS-TP and IP/MPLS Interworking considerations	14
5.7. Delay and delay variation	14
5.8. More on MPLS-TP Deployment Considerations	17
6. Security Considerations	19
7. IANA Considerations	19
8. Normative References	19
9. Informative References	19
10. Author's Addresses.....	20

Requirements Language

Although this document is not a protocol specification, 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 [RFC 2119].

1. Introduction

1.1. Background and Motivation

This document provides case studies and network design considerations for Multiprotocol Label Switching Transport Profile (MPLS-TP).

In recent years, the urgency for moving from traditional transport technologies, such as SONET/SDH, TDM/ATM, to new packet technologies

has been rising. This is largely due to the tremendous success of data services, such as IPTV and IP Video for content downloading, streaming, and sharing; rapid growth of mobile services, especially smart phone applications; the continued growth of business VPNs and residential broadband. The end of live for many legacy TDM devices and the continuing network convergence effort are also key contributing factors for transport moving toward packet

MPLS-TP Use Case and Design Considerations
Expires April 2012

technologies. After several years of heated debate on which packet technology to use, MPLS-TP has emerged as the next generation transport technology of choice for many service providers worldwide.

MPLS-TP is based on MPLS technologies. MPLS-TP re-use a subset of MPLS base functions, such as MPLS data forwarding, Pseudo-wire encapsulation for circuit emulation, and GMPLS for LSP, tLDP for PW, as dynamic control plane options; MPLS-TP extended current MPLS OAM functions, such as BFD extension for Connectivity for proactive Connectivity Check (CC) and Connectivity Verification (CV), and Remote Defect Indication (RDI), LSP Ping Extension for on demand Connectivity Check (CC) and Connectivity Verification (CV), fault allocation, and remote integrity check. New tools are being defined for alarm suppression with Alarm Indication Signal (AIS), and trigger of switch over with Link Defect Indication (LDI).

The goal is to take advantage of the maturity of MPLS technology, re-use the existing component when possible and extend the existing protocols or create new procedures/protocols when needed to fully satisfy the transport requirements.

The general requirements of MPLS-TP are provided in MPLS-TP Requirements [RFC 5654], and the architectural framework are defined in MPLS-TP Framework [RFC 5921]. This document intent to provide the use case studies and design considerations from practical point of view based on Service Providers deployments plans and field implementations.

The most common use cases for MPLS-TP include Metro access and aggregation, Mobile Backhaul, and Packet Optical Transport. MPLS-TP data plane architecture, path protection mechanisms, and OAM functionalities are used to support these deployment scenarios. As part of MPLS family, MPLS-TP complements today's IP/MPLS technologies; it closes the gaps in the traditional access and aggregation transport to enable end-to-end packet technology solutions in a cost efficient, reliable, and interoperable manner.

The unified MPLS strategy, using MPLS from core to aggregation and access (e.g. IP/MPLS in the core, IP/MPLS or MPLS-TP in aggregation and access) appear to be very attractive to many SPs. It streamlines the operation, many help to reduce the overall complexity and improve end-to-end convergence. It leverages the MPLS experience, and enhances the ability to support revenue generating services.

The design considerations discussed in this document are generic. While many design criteria are commonly apply to most of SPs, each individual SP may place the importance of one aspect over another

depending on the existing operational environment, what type of applications need to be supported, the design objectives, the cost constrain, and the network evolution plans.

1.2. Co-authors and contributors

Luyuan Fang, Cisco Systems
Nabil Bitar, Verizon
Raymond Zhang, Alcatel Lucent
Masahiro DAIKOKU, KDDI
Ping Pan, Infinera
Mach(Guoyi) Chen, Huawei Technologies
Dan Frost, Cisco Systems
Kam Lee Yap, XO Communications
Henry Yu, Time W Telecom
Jian Ping Zhang, China Telecom, Shanghai
Nurit Sprecher, Nokia Siemens Networks
Lei Wang, Telenor

2. Terminologies

AIS	Alarm Indication Signal
APS	Automatic Protection Switching
ATM	Asynchronous Transfer Mode
BFD	Bidirectional Forwarding Detection
CC	Continuity Check
CE	Customer Edge device
CV	Connectivity Verification
CM	Configuration Management
DM	Packet delay measurement
ECMP	Equal Cost Multi-path
FM	Fault Management
GAL	Generic Alert Label
G-ACH	Generic Associated Channel
GMPLS	Generalized Multi-Protocol Label Switching
LB	Loopback
LDP	Label Distribution Protocol
LM	Packet loss measurement
LSP	Label Switched Path
LT	Link trace
MEP	Maintenance End Point
MIP	Maintenance Intermediate Point
MP2MP	Multi-Point to Multi-Point connections
MPLS	Multi-Protocol Label Switching
MPLS-TP	MPLS transport profile

MPLS-TP Use Case and Design Considerations
Expires April 2012

OAM	Operations, Administration, and Management
P2P	Point to Multi-Point connections
P2MP	Point to Point connections
PE	Provider-Edge device
PHP	Penultimate Hop Popping
PM	Performance Management
PW	Pseudowire
RDI	Remote Defect Indication
RSVP-TE	Resource Reservation Protocol with Traffic Engineering Extensions
SLA	Service Level Agreement
SNMP	Simple Network Management Protocol
SONET	Synchronous Optical Network
S-PE	Switching Provider Edge
SRLG	Shared Risk Link Group
SM-PW	Multi-Segment PW
SS-PW	Single-Segment PW
TDM	Time Division Multiplexing
TE	Traffic Engineering
tLDP	target LDP
TTL	Time-To-Live
T-PE	Terminating Provider Edge
VPN	Virtual Private Network

3. Overview of MPLS-TP base functions

The section provides a summary view of MPLS-TP technology, especially in comparison to the base IP/MPLS technologies. For complete requirements and architecture definitions, please refer to [RFC 5654] and [RFC 5921].

3.1. MPLS-TP development principles

The principles for MPLS-TP development are: meeting transport requirements; maintain transport characteristics; re-using the existing MPLS technologies wherever possible to avoid duplicate the effort; ensuring consistency and inter-operability of MPLS-TP and IP/MPLS networks; developing new tools as necessary to fully meet transport requirements.

MPLS-TP Technologies include four major areas: Data Plane, Control Plane, OAM, and Survivability. The short summary is provided below.

MPLS-TP Use Case and Design Considerations
Expires April 2012

3.2. Data Plane

MPLS-TP re-used MPLS and PW architecture; and MPLS forwarding mechanism;

MPLS-TP extended the LSP support from unidirectional to both bi-directional unidirectional support.

MPLS-TP defined PHP as optional, disallowed ECMP and MP2MP, only P2P and P2MP are supported.

3.3. Control Plane

MPLS-TP allowed two control plane options:

Static: Using NMS for static provisioning;

Dynamic control plane for LSP: using GMPLS, OSPF-TE, RSVP-TE for full automation;

Dynamic control plane for PW: using tLDP.

ACH concept in PW is extended to G-ACh for MPLS-TP LSP to support in-band OAM.

Both Static and dynamic control plane options must allow control plane, data plane, management plane separation.

3.4. OAM

OAM received most attention in MPLS-TP development; Many OAM functions require protocol extensions or new development to meet the transport requirements.

1) Continuity Check (CC), Continuity Verification (CV), and Remote Integrity:

- Proactive CC and CV: Extended BFD
- On demand CC and CV: Extended LSP Ping
- Proactive Remote Integrity: Extended BFD
- On demand Remote Integrity: Extended LSP Ping

2) Fault Management:

- Fault Localization: Extended LSP Ping
- Alarm Suppression: created AIS
- Remote Defect Indication (RDI): Extended BFD
- Lock reporting: Created Lock Instruct
- Link defect Indication: Created LDI
- Static PW defect indication: Use Static PW status

MPLS-TP Use Case and Design Considerations

Expires April 2012

Performance Management:

- Loss Management: Create MPLS-TP loss/delay measurement
- Delay Measurement: Create MPLS-TP loss/delay measurement

MPLS-TP OAM tool set overview can be found at [OAM Tool Set].

3.5. Survivability

- Deterministic path protection
- Switch over within 50ms
- 1:1, 1+1, 1:N protection
- Linear protection
- Ring protection
- Shared Mesh Protection

MPLS transport Profile Survivability Framework [RFC 6372] provides more details on the subject.

4. MPLS-TP Use Case Studies

4.1. Metro Access and Aggregation

The most common deployment cases observed in the field upto today is using MPLS-TP for Metro access and aggregation. Some SPs are building green field access and aggregation infrastructure, while others are upgrading/replacing the existing transport infrastructure with new packet technologies such as MPLS-TP. The access and aggregation networks today can be based on ATM, TDM, MSTP, or Ethernet technologies as later development.

Some other SPs announced their plans for replacing their ATM or TDM aggregation networks with MPLS-TP technologies, simply because their ATM / TDM aggregation networks are no longer suited to support the rapid bandwidth growth, and they are expensive to maintain or may also be and impossible expand due to End of Sale and End of Life legacy equipments. Operators have to move forward with the next generation packet technology, the adoption of MPLS-TP in access and aggregation becomes a natural choice. The statistical muxing in MPLS-TP helps to achieve higher efficiency comparing with the time division scheme in the legacy technologies.

The unified MPLS strategy, using MPLS from core to aggregation and access (e.g. IP/MPLS in the core, IP/MPLS or MPLS-TP in aggregation and access) appear to be very attractive to many SPs. It streamlines the operation, many help to reduce the overall complexity and

improve end-to-end convergence. It leverages the MPLS experience, and enhances the ability to support revenue generating services.

The current requirements from the SPs for ATM/TDM aggregation replacement often include maintaining the current operational model, with the similar user experience in NMS, supports current access network (e.g. Ethernet, ADSL, ATM, STM, etc.), support the connections with the core networks, support the same operational feasibility even after migrating to MPLS-TP from ATM/TDM and services (OCN, IP-VPN, E-VLAN, Dedicated line, etc.). MPLS-TP currently defined in IETF are meeting these requirements to support a smooth transition.

The green field network deployment is targeting using the state of art technology to build most stable, scalable, high quality, high efficiency networks to last for the next many years. IP/MPLS and MPLS-TP are both good choices, depending on the operational model.

4.2. Packet Optical Transport

Many SP's transport networks consist of both packet and optical portions. The transport operators are typically sensitive to network deployment cost and operation simplicity. MPLS-TP is therefore a natural fit in some of the transport networks, where the operators can utilize the MPLS-TP LSP's (including the ones statically provisioned) to manage user traffic as "circuits" in both packet and optical networks.

Among other attributes, bandwidth management, protection/recovery and OAM are critical in Packet/Optical transport networks. In the context of MPLS-TP, each LSP is expected to be associated with a fixed amount of bandwidth in terms of bps and/or time-slots. OAM is to be performed on each individual LSP. For some of performance monitoring (PM) functions, the OAM mechanisms need to be able transmit and process OAM packets at very high frequency, as low as several msec's.

Protection is another important element in transport networks. Typically, ring and linear protection can be readily applied in metro networks. However, as long-haul networks are sensitive to bandwidth cost and tend to have mesh-like topology, shared mesh protection is becoming increasingly important.

Packet optical deployment plans in some SPs cases are using MPLS-TP from long haul optical packet transport all the way to the aggregation and access.

4.3. Mobile Backhaul

Wireless communication is one of the fastest growing areas in communication world wide. For some regions, the tremendous rapid mobile growth is fueled with lack of existing land-line and cable infrastructure. For other regions, the introduction of Smart phones quickly drove mobile data traffic to become the primary mobile bandwidth consumer, some SPs have already seen 85% of total mobile traffic are data traffic.

MPLS-TP has been viewed as a suitable technology for Mobile backhaul.

4.3.1. 2G and 3G Mobile Backhaul Support

MPLS-TP is commonly viewed as a very good fit for 2G)/3G Mobile backhaul.

2G (GSM/CDMA) and 3G (UMTS/HSPA/1xEVDO) Mobile Backhaul Networks are dominating mobile infrastructure today.

The connectivity for 2G/3G networks are Point to point. The logical connections are hub-and-spoke. The physical construction of the networks can be star topology or ring topology. In the Radio Access Network (RAN), each mobile base station (BTS/Node B) is communicating with one Radio Controller (BSC/RNC) only. These connections are often statically set up.

Hierarchical Aggregation Architecture / Centralized Architecture are often used for pre-aggregation and aggregation layers. Each aggregation networks inter-connects with multiple access networks. For example, single aggregation ring could aggregate traffic for 10 access rings with total 100 base stations.

The technology used today is largely ATM based. Mobile providers are replacing the ATM RAN infrastructure with newer packet technologies. IP RAN networks with IP/MPLS technologies are deployed today by many SPs with great success. MPLS-TP is another suitable choice for Mobile RAN. The P2P connection from base station to Radio Controller can be set statically to mimic the operation today in many RAN environments, in-band OAM and deterministic path protection would support the fast failure detection and switch over to satisfy the SLA agreement. Bidirectional LSP may help to simplify the provisioning process. The deterministic nature of MPLS-TP LSP set up can also help packet based synchronization to maintain predictable performance regarding packet delay and jitters.

4.3.2. LTE Mobile Backhaul

One key difference between LTE and 2G/3G Mobile networks is that the logical connection in LTE is mesh while 2G/3G is P2P star connections.

In LTE, the base stations eNB/BTS can communicate with multiple Network controllers (PSW/SGW or ASNGW), and each Radio element can communicate with each other for signal exchange and traffic offload to wireless or Wireline infrastructures.

IP/MPLS may have a great advantage in any-to-any connectivity environment. The use of mature IP or L3VPN technologies is particularly common in the design of SP's LTE deployment plan.

MPLS-TP can also bring advantages with the in-band OAM and path protection mechanism. MPLS-TP dynamic control-plane with GMPLS signaling may bring additional advantages in the mesh environment for real time adaptivities, dynamic topology changes, and network optimization.

Since MPLS-TP is part of the MPLS family. Many component already shared by both IP/MPLS and MPLS-TP, the line can be further blurred by sharing more common features. For example, it is desirable for many SPs to introduce the in-band OAM developed for MPLS-TP back into IP/MPLS networks as an enhanced OAM option. Today's MPLS PW can also be set statically to be deterministic if preferred by the SPs without going through full MPLS-TP deployment.

4.3.3. WiMAX Backhaul

WiMAX Mobile backhaul shares the similar characteristics as LTE, with mesh connections rather than P2P, star logical connections.

5. Network Design Considerations

5.1. IP/MPLS vs. MPLS-TP

Questions one might hear: I have just built a new IP/MPLS network to support multi-services, including L2/L3 VPNs, Internet service, IPTV, etc. Now there is new MPLS-TP development in IETF. Do I need to move onto MPLS-TP technology to state current with technologies?

MPLS-TP Use Case and Design Considerations

Expires April 2012

The answer is no. MPLS-TP is developed to meet the needs of traditional transport moving towards packet. It is designed to support the transport behavior coming with the long history. IP/MPLS and MPLS-TP both are state of art technologies. IP/MPLS support both transport (e.g. PW, RSVP-TE, etc.) and services (e.g L2/L3 VPNs, IPTV, Mobile RAN, etc.), MPLS-TP provides transport only. The new enhanced OAM features built in MPLS-TP should be share in both flavors through future implementation.

Another common question: I need to evolve my ATM/TDM/SONET/SDH networks into new packet technologies, but my operational force is largely legacy transport, not familiar with new data technologies, and I want to maintain the same operational model for the time being, what should I do? The answer would be: MPLS-TP may be the best choice today for the transition.

A few important factors need to be considered for IP/MPLS or MPLS-TP include:

- Technology maturity (IP/MPLS is much more mature with 12 years development)
- Operation experience (Work force experience, Union agreement, how easy to transition to a new technology? how much does it cost?)
- Needs for Multi-service support on the same node (MPLS-TP provide transport only, does not replace many functions of IP/MPLS)
- LTE, IPTV/Video distribution considerations (which path is the most viable for reaching the end goal with minimal cost? but it also meet the need of today's support)

5.2. Standards compliance

It is generally recognized by SPs that standards compliance are important for driving the cost down and product maturity up, multi-vendor interoperability, also important to meet the expectation of the business customers of SP's.

MPLS-TP is a joint work between IETF and ITU-T. In April 2008, IETF and ITU-T jointly agreed to terminate T-MPLS and progress MPLS-TP as joint work [RFC 5317]. The transport requirements would be provided by ITU-T, the protocols would be developed in IETF.

Today, majority of the core set of MPLS-TP protocol definitions are published as IETF RFCs already. It is important to deploy the solutions based on the standards definitions, in order to ensure the compatibility between MPLS-TP and IP/MPLS networks, and the interoperability among different equipment by different vendors.

Note that using non-standards, e.g. experimental code point is not recommended practice, it bares the risk of code-point collision, as indicated by [RFC 3692]: It can lead to interoperability problems when the chosen value collides with a different usage, as it someday surely will.

5.3. End-to-end MPLS OAM consistency

In the case Service Providers deploy end-to-end MPLS solution with the combination of dynamic IP/MPLS and static or dynamic MPLS-TP cross core, service edge, and aggregation/access networks, end-to-end MPLS OAM consistency becomes an essential requirements from many Service Provider. The end-to-end MPLS OAM can only be achieved through implementation of IETF MPLS-TP OAM definitions.

5.4. PW Design considerations in MPLS-TP networks

In general, PW works the same as in IP/MPLS network, both SS-PW and MS-PW are supported.

For dynamic control plane, tLDP is used. For static provisioning is used, PW status is a new PW OAM feature for failure notification.

In addition, both directions of a PW must be bound to the same transport bidirectional LSP.

When multi-tier rings involved in the network topology, should S-PE be used or not? It is a design trade-off.

- . Pros for using S-PE
 - . Domain isolation, may facilitate trouble shooting
 - . the PW failure recovery may be quicker
- . Cons for using S-PE
 - . Adds more complexity
 - . If the operation simplicity is the high priority, some SPs choose not to use S-PE, simply forming longer path across primary and secondary rings.

Should PW protection for the same end points be considered? It is another design trade-off.

- . Pros for using PW protection
 - . PW is protected when both working and protect LSPs carrying the working PW fails as long as the protection PW is following a diverse LSP path from the one carrying the working PW.

- . Cons for using PW protection
 - . Adds more complexity, some may choose not to use if protection against single point of failure is sufficient.

5.5. Proactive and event driven MPLS-TP OAM tools

MPLS-TP provide both proactive tools and event drive OAM Tools.

E.g. in the proactive fashion, the BFD hellos can be sent every 3.3 ms as its lowest interval, 3 missed hellos would be trigger the failure protection switch over. BFD sessions should be configured for both working and protecting LSPs.

When Unidirectional Failure occurs, RDI will send the failure notification to the opposite direction to trigger both end switch over.

In the reactive fashion, when there is a fiber cut for example, LDI message would be generated from the failure point and propagate to MEP to trigger immediate switch over from working to protect path. And AIS would propagate from MIP to MEP for alarm suppression.

Should both proactive and event driven OAM tools be used? The answer is yes.

Should BFD timers be set as low as possible? It depends on the applications. In many cases, it is not necessary. The lower the times are, the faster the detection time, and also the higher resource utilization. It is good to choose a balance point.

5.6. MPLS-TP and IP/MPLS Interworking considerations

Since IP/MPLS is largely deployment in most networks, MPLS-TP and IP/MPLS interworking is a reality.

Typically, there is peer model and overlay model.

The inter-connection can be simply VLAN, or PW, or could be MPLS-TE. A separate document is addressing the in the interworking issues, please refer to the descriptions in [Interworking].

5.7. Delay and delay variation

Background/motivation: Telecommunication Carriers plan to replace the aging TDM Services (e.g. legacy VPN services) provided by Legacy TDM technologies/equipments to new VPN services provided by MPLS-TP technologies/equipments with minimal cost. The Carriers cannot allow any degradation of service quality, service operation Level, and service availability when migrating out of Legacy TDM technologies/equipments to MPLS-TP transport. The requirements from the customers of these carriers are the same before and after the migration.

5.7.1. Network Delay

From our recent observation, more and more Ethernet VPN customers becoming very sensitive to the network delay issues, especially the financial customers. Many of those customers has upgraded their systems in their Data Centers, e.g., their accounting systems. Some of the customers built the special tuned up networks, i.e. Fiber channel networks, in their Data Centers, this tripped more strict delay requirements to the carriers.

There are three types of network delay:

1. Absolute Delay Time

Absolute Delay Time here is the network delay within SLA contract. It means the customers have already accepted the value of the Absolute Delay Time as part of the contract before the Private Line Service is provisioned.

2. Variation of Absolute Delay Time (without network configuration changes).

The variation under discussion here is mainly induced by the buffering in network elements.

Although there is no description of Variation of Absolute Delay Time on the contract, this has no practical impact on the customers who contract for the highest quality of services available. The bandwidth is guaranteed for those customers' traffic.

3. Relative Delay Time

Relative Delay Time is the difference of the Absolute Delay Time between using working and protect path.

Ideally, Carriers would prefer the Relative Delay Time to be zero, for the following technical reasons and network operation feasibility concerns.

The following are the three technical reasons:

Legacy throughput issue

In the case that Relative Delay Time is increased between FC networks or TCP networks, the effective throughput is degraded. The effective throughput, though it may be recovered after revert back to the original working path in revertive mode.

On the other hand, in that case that Relative Delay Time is decreased between FC networks or TCP networks, buffering over flow may occur at receiving end due to receiving large number of busty packets. As a consequence, effective throughput is degraded as well. Moreover, if packet reordering is occurred due to RTT decrease, unnecessary packet resending is induced and effective throughput is also further degraded. Therefore, management of Relative Delay Time is preferred, although this is known as the legacy TCP throughput issue.

Locating Network Acceralators at CE

In order to improve effective throughput between customer's FC networks over Ethernet private line service, some customer put "WAN Accelerator" to increase throughput value. For example, some WAN Accelerators at receiving side may automatically send back "R_RDY" in order to avoid decreasing a number of BBcredit at sending side, and the other WAN Accelerators at sending side may have huge number of initial BB credit.

When customer tunes up their CE by locating WAN Accelerator, for example, when Relative Delay Time is changes, there is a possibility that effective throughput is degraded. This is because a lot of packet destruction may be occurred due to loss of synchronization, when change of Relative delay time induces packet reordering. And, it is difficult to re-tune up their CE network element automatically when Relative Delay Time is changed, because only less than 50 ms network down detected at CE.

Depending on the tuning up method, since Relative Delay Time affects effective throughput between customer's FC networks, management of Relative Delay Time is preferred.

c) Use of synchronized replication system

MPLS-TP Use Case and Design Considerations

Expires April 2012

Some strict customers, e.g. financial customers, implement "synchronized replication system" for all data back-up and load sharing. Due to synchronized replication system, next data processing is conducted only after finishing the data saving to both primary and replication DC storage. And some tuning function could be applied at Server Network to increase throughput to the replication DC and Client Network. Since Relative Delay Time affects effective throughput, management of Relative Delay Time is preferred.

The following are the network operational feasibility issues.

Some strict customers, e.g., financial customer, continuously checked the private line connectivity and absolute delay time at CEs. When the absolute delay time is changed, that is Relative delay time is increased or decreased, the customer would complain.

From network operational point of view, carrier want to minimize the number of customers complains, MPLS-TP LSP provisioning with zero Relative delay time is preferred and management of Relative Delay Time is preferred.

Obviously, when the Relative Delay Time is increased, the customer would complain about the longer delay. When the Relative Delay Time is decreased, the customer expects to keep the lesser Absolute Delay Time condition and would complain why Carrier did not provide the best solution in the first place. Therefore, MPLS-TP LSP provisioning with zero Relative Delay Time is preferred and management of Relative Delay Time is preferred.

More discussion will be added on how to manage the Relative delay time.

5.8. More on MPLS-TP Deployment Considerations

5.8.1. Network Modes Selection

When considering deployment of MPLS-TP in the network, possibly couple of questions will come into mind, for example, where should the MPLS-TP be deployed? (e.g., access, aggregation or core network?) Should IP/MPLS be deployed with MPLS-TP simultaneously? If MPLS-TP and IP/MPLS is deployed in the same network, what is the relationship between MPLS-TP and IP/MPLS (e.g., peer or overlay?) and where is the demarcation between MPLS-TP domain and IP/MPLS

MPLS-TP Use Case and Design Considerations
Expires April 2012

domain? The results for these questions depend on the real requirements on how MPLS-TP and IP/MPLS are used to provide services. For different services, there could be different choice. According to the combination of MPLS-TP and IP/MPLS, here are some typical network modes:

Pure MPLS-TP as the transport connectivity (E2E MPLS-TP), this situation more happens when the network is a totally new constructed network. For example, a new constructed packet transport network for Mobile Backhaul, or migration from ATM/TDM transport network to packet based transport network.

Pure IP/MPLS as transport connectivity (E2E IP/MPLS), this is the current practice for many deployed networks.

MPLS-TP combines with IP/MPLS as the transport connectivity (Hybrid mode)

Peer mode, some domains adopt MPLS-TP as the transport connectivity; other domains adopt IP/MPLS as the transport connectivity. MPLS-TP domains and IP/MPLS domains are interconnected to provide transport connectivity. Considering there are a lot of IP/MPLS deployments in the field, this mode may be the normal practice in the early stage of MPLS-TP deployment.

Overlay mode

b-1: MPLS-TP as client of IP/MPLS, this is for the case where MPLS-TP domains are distributed and IP/MPLS do-main/network is used for the connection of the distributed MPLS-TP domains. For examples, there are some service providers who have no their own Backhaul network, they have to rent the Backhaul network that is IP/MPLS based from other service providers.

b-2: IP/MPLS as client of MPLS-TP, this is for the case where transport network below the IP/MPLS network is a MPLS-TP based network, the MPLS-TP network provides transport connectivity for the IP/MPLS routers, the usage is analogous as today's ATM/TDM/SDH based transport network that are used for providing connectivity for IP/MPLS routers.

5.8.2. Provisioning Modes Select

ion

As stated in MPLS-TP requirements [RFC 5654], MPLS-TP network MUST be possible to work without using Control Plane. And this does not mean that MPLS-TP network has no control plane. Instead, operators could deploy their MPLS-TP with static provisioning (e.g., CLI, NMS etc.), dynamic control plane signaling (e.g., OSPF-TE/ISIS-TE, GMPLS, LDP, RSVP-TE etc.), or combination of static and dynamic provisioning (Hybrid mode). Each mode has its own pros and cons and how to determine the right mode for a specific network mainly

depends on the operators' preference. For the operators who are used to operate traditional transport network and familiar with the Transport-Centric operational model (e.g., NMS configuration without control plane) may prefer static provisioning mode. The dynamic provisioning mode is more suitable for the operators who are familiar with the operation and maintenance of IP/MPLS network where a fully dynamic control plane is used. The hybrid mode may be used when parts of the network are provisioned with static way and the other parts are controlled by dynamic signaling. For example, for big SP, the network is operated and maintained by several different departments who prefer to different modes, thus they could adopt this hybrid mode to support both static and dynamic modes hence to satisfy different requirements. Another example is that static provisioning mode is suitable for some parts of the network and dynamic provisioning mode is suitable for other parts of the networks (e.g., static for access network, dynamic for metro aggregate and core network).

6. Security Considerations

Reference to [RFC 5920]. More will be added.

7. IANA Considerations

This document contains no new IANA considerations.

8. Normative References

[RFC 5317]: Joint Working Team (JWT) Report on MPLS Architectural Considerations for a Transport Profile, Feb. 2009.

[RFC 5654], Niven-Jenkins, B., et al, "MPLS-TP Requirements," RFC 5654, September 2009.

9. Informative References

[RFC 2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC 3692] T. Narten, "Assigning Experimental and Testing Numbers Considered Useful", RFC 3692, Jan. 2004.

[RFC 5921] Bocci, M., ED., Bryant, S., ED., et al., Frost, D. ED., Levrau, L., Berger., L., "A Framework for MPLS in Transport," July 2010.

MPLS-TP Use Case and Design Considerations
Expires April 2012

[RFC 5920] Fang, L., ED., et al, "Security Framework for MPLS and GMPLS Networks," July 2010.

[RFC 6372] Sprecher, N., Ferrel, A., MPLS transport Profile Survivability Framework [RFC 6372], September 2011.

[OAM Tool Set] Sprecher, N., Fang, L., "An Overview of the OAM Tool Set for MPLS Based Transport Networks, ", draft-ietf-mpls-to-oam-analysis-06.txt, Oct. 2011, work in progress.

[Interworking] Martinotti, R., et al., "Interworking between MPLS-TP and IP/MPLS", draft-martinotti-mpls-tp-interworking-02.txt, June 2011.

10. Author's Addresses

Luyuan Fang
Cisco Systems, Inc.
111 Wood Ave. South
Iselin, NJ 08830
USA
Email: lufang@cisco.com

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02145
USA
Email: nabil.bitar@verizon.com

Raymond Zhang
British Telecom
BT Center
81 Newgate Street
London, EC1A 7AJ
United Kingdom
Email: raymond.zhang@bt.com

Masahiro DAIKOKU
KDDI corporation
3-11-11.Tidabashi, Chiyodaku, Tokyo
Japan
Email: ms-daikoku@kddi.com

MPLS-TP Use Case and Design Considerations
Expires April 2012

Kam Lee Yap
XO Communications
13865 Sunrise Valley Drive,
Herndon, VA 20171
Email: klyap@xo.com

Dan Frost
Cisco Systems, Inc.
Email: danfrost@cisco.com

Henry Yu
TW Telecom
10475 Park Meadow Dr.
Littleton, CO 80124
Email: henry.yu@twtelecom.com

Jian Ping Zhang China Telecom, Shanghai
Room 3402, 211 Shi Ji Da Dao
Pu Dong District, Shanghai
China Email: zhangjp@shtel.com.cn

Lei Wang
Telenor
Telenor Norway
Office Snaroyveien
1331 Fornebu
Email: Lai.wang@telenor.com

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd.
No. 3 Xinxin Road
Shangdi Information Industry Base
Hai-Dian District, Beijing 100085
China
Email: mach@huawei.com

Nurit Sprecher
Nokia Siemens Networks
3 Hanagar St. Neve Ne'eman B
Hod Hasharon, 45241
Israel
Email: nurit.sprecher@nsn.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 18, 2012

N. Sprecher
Nokia Siemens Networks
L. Fang
Cisco
April 17, 2012

An Overview of the OAM Tool Set for MPLS based Transport Networks
draft-ietf-mpls-tp-oam-analysis-09.txt

Abstract

This document provides an overview of the OAM toolset for MPLS based Transport Networks (MPLS-TP). The toolset consists of a comprehensive set of fault management and performance monitoring capabilities (operating in the data-plane) which are appropriate for transport networks as required in RFC 5860 and support the network and services at different nested levels. This overview includes a brief recap of MPLS-TP OAM requirements and functions, and of generic mechanisms created in the MPLS data plane to allow the OAM packets run in-band and share their fate with data packets. The protocol definitions for each of the MPLS-TP OAM tools are defined in separate documents (RFCs or Working Group drafts) which are referenced by this document.

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 October 18, 2012.

Copyright Notice

Copyright (c) 2012 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	4
1.1. Scope	4
1.2. Contributing Authors	5
1.3. Acronyms	6
2. Basic OAM Infrastructure Functionality	6
3. MPLS-TP OAM Functions	8
3.1. Continuity Check and Connectivity Verification	8
3.1.1. Documents for CC-CV tools	9
3.2. Remote Defect Indication	9
3.2.1. Documents for RDI	9
3.3. Route Tracing	9
3.3.1. Documents for Route Tracing	10
3.4. Alarm Reporting	10
3.4.1. Documents for Alarm Reporting	10
3.5. Lock Instruct	10
3.5.1. Documents for Lock Instruct	10
3.6. Lock Reporting	10
3.6.1. Documents for Lock Reporting	10
3.7. Diagnostic	11
3.7.1. Documents for Diagnostic Testing	11
3.8. Packet Loss Measurement	11
3.8.1. Documents for Packet Loss Measurement	11
3.9. Packet Delay Measurement	12
3.9.1. Documents for Delay Measurement	12
4. MPLS-TP OAM documents guide	12
5. OAM Toolset Applicability and Utilization	14
5.1. Connectivity Check and Connectivity Verification	14
5.2. Diagnostic Tests and Lock Instruct	15
5.3. Lock Reporting	16
5.4. Alarm Reporting and Link Down Indication	16
5.5. Remote Defect Indication	17
5.6. Packet Loss and Delay Measurement	17
6. IANA Considerations	18
7. Security Considerations	18
8. Acknowledgements	19
9. References	19
9.1. Normative References	19
9.2. Informative References	21
Authors' Addresses	21

1. Introduction

1.1. Scope

The MPLS Transport Profile (MPLS-TP) architectural framework is defined in [RFC 5921], and it describes common set of protocol functions that supports the operational models and capabilities typical of such networks.

OAM (Operations, Administration, and Maintenance) plays a significant role in carrier networks, providing methods for fault management and performance monitoring in both the transport and the service layers in order to improve their ability to support services with guaranteed and strict Service Level Agreements (SLAs) while reducing their operational costs.

[RFC 5654], in general, and [RFC 5860], in particular, define a set of requirements for OAM functionality for MPLS-Transport Profile (MPLS-TP) Label Switched Paths (LSPs)), Pseudowires (PWs) and sections.

The OAM solution, developed by the joint IETF and ITU-T MPLS-TP project, has three objectives:

- o The OAM toolset should be developed based on existing MPLS architecture, technology, and toolsets.
- o The OAM operational experience should be similar to that in other transport networks.
- o The OAM toolset developed for MPLS based transport networks needs to be fully inter-operable with existing MPLS OAM tools as documented in [RFC 5860].

The MPLS-TP OAM toolset is based on the following existing tools:

- o LSP-Ping as defined in [RFC 4379].
- o Bidirectional Forwarding Detection (BFD) as defined in [RFC 5880] and refined in [RFC 5884].
- o ITU-T OAM for Ethernet toolset as defined in [Y.1731]. This has been used for functionality guidelines for the performance measurement tools that were not previously supported in MPLS.

It should be noted that certain extensions and adjustments have been specified relative to the existing MPLS tools, in order to conform to the transport environment and the requirements of MPLS-TP. However,

compatibility with the existing tools has been maintained.

This document provides an overview of the MPLS-TP OAM toolset, which consists of tools for MPLS-TP fault management and performance monitoring. This overview includes a brief recap of MPLS-TP OAM requirements and functions, and of the generic mechanisms used to support the MPLS-TP OAM operation.

The protocol definitions for each individual MPLS-TP OAM tool are specified in separate RFCs (or Working Group documents while this document is work in progress), which are referenced by this document.

In addition, the document includes a table that cross-references the solution documents to the OAM functionality supported. Finally, the document presents the applicability and utilization of each tool in the MPLS-TP OAM toolset.

1.2. Contributing Authors

Elisa Bellagamba	Ericsson
Yaacov Weingarten	Nokia Siemens Networks
Dan Frost	Cisco
Nabil Bitar	Verizon
Raymond Zhang	Alcatel Lucent
Lei Wang	Telenor
Kam Lee Yap	XO Communications
John Drake	Juniper
Yaakov Stein	RAD
Anamaria Fulignoli	Ericsson
Italo Busi	Alcatel Lucent
Huub van Helvoort	Huawei
Thomas Nadeau	Computer Associate
Henry Yu	TW Telecom
Mach Chen	Huawei
Manuel Paul	Deutsche Telekom

1.3. Acronyms

This document uses the following acronyms:

ACH	Associated Channel Header
AIS	Alarm Indication Signal
BFD	Bidirectional Forwarding Detection
CC-CV	Continuity Check and Connectivity Verification
DM	Delay Measurement
FM	Fault Management
G-ACh	Generic Associated Channel
GAL	G-ACh Label
GMPLS	Generalized Multi-Protocol Label Switching
IANA	Internet Assigned Names Authority
LDI	Link Down Indication
LKR	Lock Report
LM	Loss Measurement
LOC	Loss of Continuity
LSP	Label Switched Path
MEP	Maintenance Entity Group End Point
MEG	Maintenance Entity Group
MIP	Maintenance Entity Group Intermediate Point
MPLS	MultiProtocol Label Switching
MPLS-TP	Transport Profile for MPLS
OAM	Operations, Administration, and Maintenance
PM	Performance Monitoring
PW	Pseudowire
RDI	Remote Defect Indication
SLA	Service Level Agreement
TLV	Type, Length, Value
VCCV	Virtual Circuit Connectivity Verification

2. Basic OAM Infrastructure Functionality

[RFC 5860] defines a set of requirements on OAM architecture and general principles of operations, which are evaluated below:

[RFC 5860] requires that --

- o OAM mechanisms in MPLS-TP are independent of the transmission media and of the client service being emulated by the PW ([RFC 5860], section 2.1.2).
- o MPLS-TP OAM must be able to support both an IP based and non-IP based environment. If the network is IP based, i.e. IP routing and forwarding are available, then it must be possible to choose to make use of IP capabilities. On the other hand, in

environments where IP functionality is not available, the OAM tools must still be able to operate independent of IP forwarding and routing ([RFC 5860], section 2.1.4). It is required to have OAM interoperability between distinct domains materializing the environments ([RFC 5860], section 2.1.5).

- o all OAM protocols support identification information, at least in the form of IP addressing structure and be extensible to support additional identification schemes ([RFC 5860], section 2.1.4).
- o OAM packets and the user traffic are congruent (i.e. OAM packets are transmitted in-band) and there is a need to differentiate OAM packets from user-plane packets ([RFC 5860], section 2.1.3). Inherent in this requirement is the principle that full operation of the MPLS-TP OAM must be possible independently of the control or management plane used to operate the network ([RFC 5860], section 2.1.3).
- o MPLS-TP OAM supports point-to-point bidirectional PWs, point-to-point co-routed bidirectional LSPs, point-to-point bidirectional Sections ([RFC 5860], section 2.1.1). The applicability of particular MPLS-TP OAM functions to point-to-point associated bidirectional LSPs, point-to-point unidirectional LSPs, and point-to-multipoint LSPs, is described in ([RFC 5860], section 2.2)). In addition, MPLS-TP OAM supports these LSPs and PWs when they span either a single or multiple domains ([RFC 5860], section 2.1.1).
- o OAM packets may be directed to an intermediate point of a LSP/PW ([RFC 5860], sections 2.2.3, 2.2.4 and 2.2.5).

[RFC 5860] recommends that any protocol solution, meeting one or more functional requirement(s), be the same for PWs, LSPs, and Sections (section 2.2).

The following document-set addresses the basic requirements listed above:

- o The [RFC 6371] document describes the architectural framework for conformance to the basic requirements listed above. It also defines the basic relationships between the MPLS structures, e.g. LSP, PW, and the structures necessary for OAM functionality, i.e. the Managed Entity Group, its End-points, and Intermediate Points.
- o The [RFC 5586] document specifies the use of the MPLS-TP in-band control channels. It generalizes the applicability of the Pseudowire (PW) Associated Channel Header (ACH) to MPLS LSPs and Sections, by defining a Generic Associated Channel (G-ACh). The

G-ACh allows control packets to be multiplexed transparently over LSPs and sections, similar to that of PW VCCV [RFC 5085]. The Generic Association Label (GAL) is defined by assigning a reserved MPLS label value and is used to identify the OAM control packets. The value of the ACH Channel Type field indicates the specific protocol carried on the associated control channel. Each MPLS-TP OAM protocol has an IANA assigned channel type allocated to it.

[RFC 5085] defines an Associated Channel Header (ACH) which provides a PW associated control channel between a PW's endpoints, over which OAM and other control messages can be exchanged. [RFC 5586] generalizes MPLS-TP generalized the PW Associated Channel Header (ACH) to provide common in-band control channels also at the LSP and MPLS-TP link levels. The G-ACh allows control packets to be multiplexed transparently over the same LSP or MPLS-TP link as in PW VCCV. Multiple control channels can exist between endpoints.

[RFC 5085] also defines a label-based exception mechanism that helps an LSR to identify the control packets and direct them to the appropriate entity for processing. The use of G-ACh and GAL provides the necessary mechanisms to allow OAM packets run in-band and share their fate with data packets. It is expected that all of the OAM protocols will be used in conjunction with this Generic Associated Channel.

- o The [RFC 6370] document provides an IP-based identifier set for MPLS-TP that can be used to identify the transport entities in the network and referenced by the different OAM protocols.
[MPLS TP ITU Idents] augments that set of identifiers to include identifier information in a format used by the ITU-T. Other identifier sets may be defined as well.

3. MPLS-TP OAM Functions

The following sections discuss the OAM functions that are required in [RFC 5860] and expanded upon in [RFC 6371].

3.1. Continuity Check and Connectivity Verification

Continuity Check and Connectivity Verification (CC-CV) are OAM operations generally used in tandem, and complement each other. These functions are generally run proactively, but may also be used on-demand for diagnoses of a specific condition. Proactively [RFC 5860] states that the function should allow the MEPs to monitor the liveness and connectivity of a transport path (LSP, PW or a section) between them. In on-demand mode, this function should

support monitoring between the MEPs and, in addition, between a MEP and MIP. Note that as specified in sections 3.3 and 3.4 of [RFC 6371], a MEP and a MIP can reside in an unspecified location within a node, or in a particular interface on a specific side of the forwarding engine.

The [RFC 6371] highlights the need for the CC-CV messages to include unique identification of the MEG that is being monitored and the MEP that originated the message. The function, both proactively and in on-demand mode, needs to be transmitted at regular transmission rates pre-configured by the operator.

3.1.1. Documents for CC-CV tools

[RFC 6428] defines BFD extensions to support proactive CC-CV applications.

[RFC 6426] provides LSP-Ping extensions that are used to implement on-demand Connectivity Verification.

Both of these tools will be used within the framework of the basic tools described above, in section 2.

3.2. Remote Defect Indication

Remote Defect Indication (RDI) is used by a path end-point to report that a defect is detected on a bi-directional connection to its peer end-point. [RFC 5860] points out that this function may be applied to a unidirectional LSP only if a return path exists. [RFC 6371] points out that this function is associated with the proactive CC-CV function.

3.2.1. Documents for RDI

The [RFC 6428] document includes an extension for BFD that would include the RDI indication in the BFD format, and a specification of how this indication is to be used.

3.3. Route Tracing

[RFC 5860] defines that there is a need for functionality that would allow a path end-point to identify the intermediate (if any) and end-points of the path (LSP, PW or a section). This function would be used in on-demand mode. Normally, this path will be used for bidirectional PW, LSP, and sections, however, unidirectional paths may be supported only if a return path exists.

3.3.1. Documents for Route Tracing

The [RFC 6426] document that specifies the LSP-Ping enhancements for MPLS-TP on-demand Connectivity Verification includes information on the use of LSP-Ping for route tracing of a MPLS-TP transport path.

3.4. Alarm Reporting

Alarm Reporting is a function used by an intermediate point of a path (LSP or PW), that becomes aware of a fault on the path, to report to the end-points of the path. [RFC 6371] states that this may occur as a result of a defect condition discovered at a server layer. The intermediate point generates an Alarm Indication Signal (AIS) that continues until the fault is cleared. The consequent action of this function is detailed in [RFC 6371].

3.4.1. Documents for Alarm Reporting

MPLS-TP defines a new protocol to address this functionality that is documented in [RFC 6427]. This protocol uses all of the basic mechanisms detailed in Section 2.

3.5. Lock Instruct

The Lock Instruct function is an administrative control tool that allows a path end-point to instruct its peer end-point to lock the path (LSP, PW or section). The tool is necessary to support single-side provisioning for administrative locking, according to [RFC 6371]. This function is used on-demand.

3.5.1. Documents for Lock Instruct

The [RFC 6435] document describes the details of a new ACH based protocol format for this functionality.

3.6. Lock Reporting

Lock reporting, defined in [RFC 5860], is similar to the Alarm Reporting function described above. It is used by an intermediate point to notify the end points of a transport path (LSP or PW) that an administrative lock condition exists for this transport path.

3.6.1. Documents for Lock Reporting

MPLS-TP defines a new protocol to address this functionality that is documented in [RFC 6427]. This protocol uses all of the basic mechanisms detailed in Section 2.

3.7. Diagnostic

The [RFC 5860] indicates that there is need to provide a OAM function that would enable conducting different diagnostic tests on a PW, LSP, or Section. The [RFC 6371] provides two types of specific tests to be used through this functionality:

- o Throughput Estimation - allowing the provider to verify the bandwidth/throughput of a transport path. This is an out-of-service tool, that uses special packets of varying sizes to test the actual bandwidth and/or throughput of the path.
- o Data-plane loopback - this out-of-service tool causes all traffic that reaches the target node, either a MEP or MIP, to be looped back to the originating MEP. For targeting MIPs, a co-routed bi-directional path is required.

3.7.1. Documents for Diagnostic Testing

The [RFC 6435] document describes the details of a new ACH based protocol format for the Data-plane loopback functionality.

The tool for Throughput Estimation tool is under study.

3.8. Packet Loss Measurement

Packet Loss Measurement is required by [RFC 5860] to provide a quantification of the packet loss ratio on a transport path. This is the ratio of the number of user packets lost to the total number of user packets during a defined time interval. To employ this function, [RFC 6371] defines that the two end-points of the transport path should exchange counters of messages transmitted and received within a time period bounded by loss-measurement messages. The framework warns that there may be small errors in the computation that result from various issues.

3.8.1. Documents for Packet Loss Measurement

The [RFC 6374] document describes the protocol formats and procedures for using the tool and enable efficient and accurate measurement of packet loss, delay, and throughput in MPLS networks. [RFC 6375] describes a profile of the general MPLS loss, delay, and throughput measurement techniques that suffices to meet the specific requirements of MPLS-TP. Note that the tool logic is based on the behavior of the parallel function described in [Y.1731].

3.9. Packet Delay Measurement

Packet Delay Measurement is a function that is used to measure one-way or two-way delay of a packet transmission between a pair of the end-points of a path (PW, LSP, or Section), as described in [RFC 5860]. Where:

- o One-way packet delay is the time elapsed from the start of transmission of the first bit of the packet by a source node until the reception of the last bit of that packet by the destination node.
- o Two-way packet delay is the time elapsed from the start of transmission of the first bit of the packet by a source node until the reception of the last bit of the loop-backed packet by the same source node, when the loopback is performed at the packet's destination node.

[RFC 6371] describes how the tool could be performed (both in proactive and on-demand modes) for either one-way or two-way measurement. However, it warns that the one-way delay option requires precise time synchronization between the end-points.

3.9.1. Documents for Delay Measurement

The [RFC 6374] document describes the protocol formats and procedures for using the tool and enable efficient and accurate measurement of packet loss, delay, and throughput in MPLS networks. [RFC 6375] describes a profile of the general MPLS loss, delay, and throughput measurement techniques that suffices to meet the specific requirements of MPLS-TP. Note that the tool logic is based on the behavior of the parallel function described in [Y.1731].

4. MPLS-TP OAM documents guide

The complete MPLS-TP OAM protocol suite is covered by a small set of existing IETF documents. This set of documents may be expanded in the future to cover additional OAM functionality. In order to allow the reader to understand this set of documents, a cross-reference of the existing documents (IETF RFCs or Internet drafts while this document is work in progress) for the initial phase of the specification of MPLS based transport networks is provided below.

[RFC 5586] provides a specification of the basic structure of protocol messages for in-band data plane OAM in an MPLS environment.

[RFC 6370] provides definitions of different formats that may be used

within OAM protocol messages to identify the network elements of a MPLS based transport network.

The following table (Table 1) provides the summary of proactive MPLS-TP OAM Fault Management toolset functions, associated tool/protocol, and the corresponding IETF RFCs where they are defined.

OAM Functions	OAM Tools/Protocols	RFCs
Continuity Check and Connectivity Verification	Bidirectional Forwarding Detection (BFD)	[RFC 6428]
Remote Defect Indication (RDI)	Flag in Bidirectional Forwarding Detection (BFD) message	[RFC 6428]
Alarm Indication Signal (AIS)	G-ACh bases AIS message	[RFC 6427]
Link Down Indication (LDI)	Flag in AIS message	[RFC 6427]
Lock Reporting (LKR)	G-ACh bases LKR message	[RFC 6427]

Proactive Fault Management OAM Toolset

Table 1

The following table (Table 2) provides an overview of the on-demand MPLS-TP OAM Fault Management toolset functions, associated tool/protocol, and the corresponding IETF RFCs they are defined.

OAM Functions	OAM Tools/Protocols	RFCs
Connectivity Verification	LSP Ping	[RFC 6426]
Diagnostic: Loopback and Lock Instruct	(1) G-ACh based Loopback and Lock Instruct, (2) LSP Ping	[RFC 6435]

Lock Instruct(LI)	Flag in AIS message	[RFC 6427]
-------------------	---------------------	------------

On Demand Fault Management OAM Toolset

Table 2

The following table (Table 3) provides the Performance Monitoring Functions, associated tool/protocol definitions, and corresponding RFCs.

OAM Functions	OAM Tools/Protocols	RFCs
Packet Loss Measurement (LM)	G-ACh based LM & DM query messages	[RFC 6374] [RFC 6375]
Packet Delay Measurement (DM)	G-ACh based LM & DM query messages	[RFC 6374] [RFC 6375]
Throughput Measurement	derived from Loss Measurement	[RFC 6374] [RFC 6375]
Delay Variation Measurement	derived from Delay Measurement	[RFC 6374] [RFC 6375]

Performance Monitoring OAM Toolset

Table 3

5. OAM Toolset Applicability and Utilization

The following subsections present the MPLS-TP OAM toolset from the perspective of the specified protocols and identifies which of the required functionality is supported by the particular protocol.

5.1. Connectivity Check and Connectivity Verification

Proactive Continuity Check and Connectivity Verification (CC-CV) functions are used to detect loss of continuity (LOC), and unintended connectivity between two MEPs. Loss of connectivity, mis-merging, mis-connectivity, or unexpected Maintenance Entity Group End Points (MEPs) can be detected using the CC-CV tools. See Section 3.1, 3.2, 3.3 in this document for CC-CV protocol references.

The CC-CV tools are used to support MPLS-TP fault management, performance management, and protection switching. Proactive CC-CV control packets are sent by the source MEP to sink MEP. The sink MEP monitors the arrival of the CC-CV control packets and detects the defect. For bidirectional transport paths, the CC-CV protocol is, usually, transmitted simultaneously in both directions.

The transmission interval of CC-CV control packet can be configured. For example:

- o 3.3ms is the default interval for protection switching.
- o 100ms is the default interval for performance monitoring.
- o 1s is the default interval for fault management.

5.2. Diagnostic Tests and Lock Instruct

[RFC 6435] describes a protocol that provides a mechanism is provided to Lock and unlock traffic (e.g. data and control traffic) or specific OAM traffic at a specific LSR on the path of the MPLS-TP LSP to allow loop back of the traffic to the source.

These diagnostic functions apply to associated bidirectional MPLS-TP LSPs, including MPLS-TP LSPs, bi-directional RSVP-TE tunnels (which is relevant for MPLS-TP dynamic control plane option with GMPLS), and single segment and multi-segment pseudowires. [RFC 6435] provides the protocol definition for diagnostic tests functions.

The Lock operation instruction is carried in an MPLS Loopback request message sent from a MEP to a trail-end MEP of the LSP to request that the LSP be taken out of service. In response, the Lock operation reply is carried in a Loopback response message sent from the trail-end MEP back to the originating MEP to report the result.

The loopback operations include:

- o Lock: take an LSP out of service for maintenance.
- o Unlock: Restore a previously locked LSP to service.
- o Set_Full_Loopback and Set_OAM_Loopback
- o Unset_Full_Loopback and Set_OAM_Loopback

Operators can use the loopback mode to test the connectivity or performance (loss, delay, delay variation, and throughput) of given LSP up to a specific node on the path of the LSP.

5.3. Lock Reporting

The Lock Report (LKR) function is used to communicate to the client (sub-) layer MEPs the administrative locking of a server (sub-) layer MEP, and consequential interruption of data traffic forwarding in the client (sub-) layer. See Section 3.6 in this document for Lock Reporting protocol references.

When operator is taking the LSP out of service for maintenance or other operational reason, using the LKR function can help to distinguish the condition as administrative locking from defect condition.

The Lock Report function would also serve the purpose of alarm suppression in the MPLS-TP network above the level at which the Lock has occurred. The receipt of an LKR message may be treated as the equivalent of loss of continuity at the client layer.

5.4. Alarm Reporting and Link Down Indication

Alarm Indication Signal (AIS) message serves the purpose of alarm suppression upon the failure detection in the server (-sub) layer. When the Link Down Indication (RDI) is set, the AIS message may be used to trigger recovery mechanisms.

When a server MEP detects the failure, it asserts Loss of Continuity (LOC) or signal fail which sets the flag up to generate OAM packet with AIS message. The AIS message is forwarded to downstream sink MEP in the client layer. This would enable the client layer to suppress the generation of secondary alarms.

A Link Down Indication (LDI) flag is defined in the AIS message. The LDI flag is set in the AIS message in response to detecting a fatal failure in the server layer. Receipt of an AIS message with this flag set may be interpreted by a MEP as an indication of signal fail at the client layer.

The protocols for Alarm Indication Signal (AIS) and Link Down Indication (LDI) are defined in [RFC 6427].

Fault OAM messages are generated by intermediate nodes where an LSP is switched, and propagated to the end points (MEPs).

From a practical point of view, when both proactive Continuity Check functions and LDI are used, one may consider running the proactive Continuity Check functions at a slower rate (e.g. longer BFD hello intervals), and reply on LDI to trigger fast protection switch over upon failure detection in a given LSP.

5.5. Remote Defect Indication

Remote Defect Indication (RDI) function enables an End Point to report to the other End Point that a fault or defect condition is detected on the PW, LSP, or Section for which they are the End Points.

The RDI OAM function is supported by the use of Bidirectional Forwarding Detection (BFD) Control Packets [RFC 6428]. RDI is only used for bidirectional connections and is associated with proactive CC-CV activation.

When an end point (MEP) detects a signal failure condition, it sets the flag up by setting the diagnostic field of the BFD control packet to a particular value to indicate the failure condition on the associated PW, LSP, or Section, and transmitting the BFD control packet with the failure flag up to the other end point (its peer MEP).

The RDI function can be used to facilitate protection switching by synchronizing the two end points when unidirectional failure occurs and is detected by one end.

5.6. Packet Loss and Delay Measurement

The packet loss and delay measurement toolset enables operators to measure the quality of the packet transmission over a PW, LSP, or Section. Section 3.8 in this document defined the protocols for packet loss measurement and 3.9 in defined the protocols for packet delay measurement.

The loss and delay protocols have the following characteristics and capabilities:

- o They support measurement of packet loss, delay and throughput over Label Switched Paths (LSPs), pseudowires, and MPLS sections.
- o The same LM and DM protocols can be used for both continuous/proactive and selective/on-demand measurement.
- o The LM and DM protocols use a simple query/response model for bidirectional measurement that allows a single node - the querier - to measure the loss or delay in both directions.
- o The LM and DM protocols use query messages for unidirectional loss and delay measurement. The measurement can either be carried out at the downstream node(s) or at the querier if an out-of-band return path is available.

- o The LM and DM protocols do not require that the transmit and receive interfaces be the same when performing bidirectional measurement.
- o The LM supports test-message-based measurement (i.e. inferred mode) as well as measurement based on data-plane counters (i.e. direct mode).
- o The LM protocol supports both 32-bit and 64-bit counters.
- o The LM protocol supports measurement in terms of both packet counts and octet counts although for simplicity only packet counters are currently included in the MPLS-TP profile.
- o The LM protocol can be used to measure channel throughput as well as packet loss.
- o The DM protocol supports varying the measurement message size in order to measure delays associated with different packet sizes.
- o The DM protocol uses IEEE 1588 timestamps by default but also supports other timestamp formats such as NTP.

6. IANA Considerations

This document makes no request of IANA.

The OAM tools and functions defined under G-ACh use IANA assigned code points. the codes are defined in the corresponding IETF RFCs

Note to RFC Editor:

this section may be removed on publication as an RFC.

7. Security Considerations

This document as an overview of MPLS OAM tools does not by itself raise any particular security considerations.

The general security considerations are provided in [RFC 6920] and [MPLS-TP Security Frwk]. Security considerations for each function in the OAM toolset have been documented in each document that specifies the particular functionality.

OAM in general is always an area where the security risk is high, e.g. confidential information may be intercepted for attackers to

again access to the networks, therefore authentication, authorization, and encryption need to be enforced for prevent security breach.

In addition to implement security protocol, tools, and mechanisms, following strict operation security procedures is very important, especially MPLS-TP static provisioning processes involve operator direct interactions with NMS and devices, its critical to prevent human errors and malicious attacks.

Since MPLS-TP OAM uses G-ACh, the security risks and mitigation described in [RFC 5085] apply here. In short, the G-ACh could be intercepted, or false G-ACh packets could be inserted. DoS attack could happen by flooding G-ACh messages to peer devices. To mitigate this type of attacks, throttling mechanisms can be used. For more details, please see [RFC 5085].

8. Acknowledgements

The authors would like to thank the MPLS-TP experts from both the IETF and ITU-T for their helpful comments. In particular, we would like to thank Loa Andersson, and the Area Directors for their suggestions and enhancements to the text.

Thanks to Tom Petch for useful comments and discussions.

Thanks to Rui Costa for his review and comments which helped improve this document.

9. References

9.1. Normative References

[RFC 4379]

Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

[RFC 5085]

Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.

- [RFC 5586]
Bocci, M., Bryant, S., and M. Vigoureux, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC 5654]
Niven-Jenkins, B., Nadeau, T., and C. Pignataro, "Requirements for the Transport Profile of MPLS", RFC 5654, April 2009.
- [RFC 5860]
Vigoureux, M., Betts, M., and D. Ward, "Requirements for OAM in MPLS Transport Networks", RFC 5860, April 2009.
- [RFC 5880]
Katz, D. and D. Ward, "Bidirectional Forwarding Detection", RFC 5880, February 2009.
- [RFC 5884]
Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "BFD For MPLS LSPs", RFC 5884, June 2008.
- [RFC 5921]
Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC 6370]
Bocci, M., Swallow, G., and E. Gray, "MPLS-TP Identifiers", RFC 6370, September 2011.
- [RFC 6371]
Busi, I., Niven-Jenkins, B., and D. Allan, "MPLS-TP OAM Framework and Overview", RFC 6371, September 2011.
- [RFC 6374]
Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC 6375]
Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-based Transport Networks", RFC 6375, September 2011.
- [RFC 6426]
Bahadur, N., Aggarwal, R., Boutros, S., and E. Gray, "MPLS on-demand Connectivity Verification, Route Tracing and Adjacency Verification", RFC 6426, August 2011.

- [RFC 6427]
Swallow, G., Fulignoli, A., and M. Vigoureux, "MPLS Fault Management OAM", RFC 6427, September 2011.
- [RFC 6428]
Allan, D. and G. Swallow, "Proactive Connectivity Verification, Continuity Check and Remote Defect indication for MPLS Transport Profile", RFC 6428, August 2011.
- [RFC 6435]
Boutros, S., Sivabalan, S., Aggarwal, R., Vigoureux, M., and X. Dai, "MPLS Transport Profile Lock Instruct and Loopback Functions", RFC 6435, September 2011.

9.2. Informative References

- [MPLS-TP Security Frwk]
Fang, L., Niven-Jenkins, B., and S. Mansfield, "MPLS-TP Security Framework",
ID draft-ietf-mpls-tp-security-framework-02, May 2011.
- [RFC 6920]
Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [Y.1731] International Telecommunications Union - Standardization,
"OAM functions and mechanisms for Ethernet based networks", ITU Y.1731, May 2006.
- [MPLS TP ITU Idents]
Winter, R., van Helvoort, H., and M. Betts, "MPLS-TP Identifiers Following ITU-T Conventions",
ID draft-ietf-mpls-tp-itu-t-identifiers-02, July 2011.

Authors' Addresses

Nurit Sprecher
Nokia Siemens Networks
3 Hanagar St. Neve Ne'eman B
Hod Hasharon, 45241
Israel

Email: nurit.sprecher@nsn.com

Luyuan Fang
Cisco
111 Wood Avenue South
Iselin, NJ 08830
USA

Email: lufang@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 7, 2012

L. Jin
ZTE
K. Liu
Nokia Siemens
S. Kini
Ericsson
May 6, 2012

Leaf discovery mechanism for mLDP based P2MP/MP2MP LSP
draft-jin-mppls-mldp-leaf-discovery-04.txt

Abstract

This document describes a mechanism for a root node to discover the leaf nodes of an mLDP based P2MP/MP2MP LSP. Such kind of function could be used for multiplexing/aggregating root initiated and leaf initiated application which will use mLDP based P2MP/MP2MP LSP. Examples of root initiated applications are P2MP PW [I-D.ietf-pwe3-p2mp-pw], VPLS multicast [I-D.ietf-l2vpn-vpls-mcast], L3VPN multicast [RFC6513]. And examples of leaf initiated applications are statically configured mLDP based P2MP/MP2MP LSP, mLDP in-band signaling [I-D.ietf-mppls-mldp-in-band-signaling].

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 November 7, 2012.

Copyright Notice

Copyright (c) 2012 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. Motivation and problem statement	3
3. Terminology	4
4. Leaf discovery mechanism	4
4.1. Leaf discovery mechanism based on T-LDP	5
4.1.1. Node operation	5
4.2. Leaf discovery mechanism based on MP-BGP	6
4.2.1. mLDP leaf NLRI	6
4.2.2. Node operation	7
5. Scalability	7
6. Security Considerations	7
7. IANA Considerations	8
7.1. MP-BGP	8
8. Acknowledgement	8
9. References	8
9.1. Normative references	8
9.2. Informative References	8
Authors' Addresses	9

1. Introduction

This document describes a mechanism for a root node to discover the leaf nodes of an mLDP based P2MP/MP2MP LSP. Such kind of function could be used for multiplexing/aggregating root initiated and leaf initiated application which will use mLDP based P2MP/MP2MP LSP. Examples of root initiated applications are P2MP PW [I-D.ietf-pwe3-p2mp-pw], VPLS multicast [I-D.ietf-l2vpn-vpls-mcast], L3VPN multicast [RFC6513]. And examples of leaf initiated applications are statically configured mLDP based P2MP/MP2MP LSP, mLDP in-band signaling [I-D.ietf-mpls-mldp-in-band-signaling].

This draft provides a discovery mechanism based on a signaling session between each leaf and root node. Each leaf node would signal the leaf node information to root node through this session. There are two signaling protocols to be used for root initiated application, targeted LDP [RFC5036] or BGP auto-discovery using BGP Multiprotocol Extensions [RFC4760]. In order to reuse the signaling protocol of root initiated application, this document introduces both signaling protocols for mLDP leaf discovery.

2. Motivation and problem statement

The leaf initiated application mLDP in-band signaled P2MP LSP will trigger the leaf node to join from leaf node, which means none of the members belonging to a P2MP/MP2MP LSP topology knows all the other members of the P2MP/MP2MP LSP. This means that the root node cannot get the whole P2MP/MP2MP LSP membership information. This problem may cause some limitation for multiplexing/aggregation root initiated applications using mLDP LSPs.

Multicast VPLS [I-D.ietf-l2vpn-vpls-mcast] is a root initiated application. When setting up a inclusive P-Multicast tunnel, BGP A-D is used to do the VPLS membership auto-discovery. The mLDP based P2MP/MP2MP LSP will be set up when receiving auto-discovery routes through BGP A-D. The root node will only know the mLDP LSP leaf node information which is triggered by the specific BGP A-D mechanism. Let's assume that a mLDP in-band signaling P2MP/MP2MP LSP_a (setup by leaf initiated application) already exist on the root node, and that LSP_a has the same leaf nodes as the P2MP LSP that VPLS multicast BGP A-D tries to set up. The root node does not know LSP_a leaf node information, and will set up mLDP based LSP_b triggered by BGP A-D with same root and leaf nodes.

This causes mLDP based LSP resources waste in the network as it may not be necessary to setup two mLDP LSPs with the same root and leaves in the same network.

The introduction of a leaf discovery mechanism for mLDP based P2MP/MP2MP LSP will enable leaf initiated applications to share one P2MP/MP2MP LSP with root initiated application of P2MP/MP2MP LSP by multiplexing/aggregating mechanism.

3. Terminology

mLDP: Multicast LDP.

T-LDP: Target LDP.

P2MP LSP: An LSP that has one Ingress LSR and one or more Egress LSRs.

MP2MP LSP: An LSP that connects a set of nodes, such that traffic sent by any node in the LSP is delivered to all others.

Bud LSR: An LSR that is an egress but also has one or more directly connected downstream LSRs.

Ingress LSR: Source of the P2MP LSP, also referred to as root node.

Egress LSR: One of potentially many destinations of an LSP, also referred to as leaf node in the case of P2MP and MP2MP LSPs.

Transit LSR: An LSR that has one or more directly connected downstream LSRs.

Leaf node: A Leaf node can be either an Egress or Bud LSR when referred in the context of a P2MP LSP. In the context of a MP2MP LSP, an LSR is both Ingress and Egress for the same MP2MP LSP and can also be a Bud LSR.

P2MP FEC: The P2MP FEC Element consists of the address of the root of the P2MP LSP and an opaque value.

MP2MP FEC: MP2MP FEC consists of MP2MP downstream FEC and upstream FEC Element.

MP FEC: Includes both P2MP FEC and MP2MP FEC.

4. Leaf discovery mechanism

It would be beneficial if the mLDP leaf discovery mechanism can reuse the same signaling session as the root initiated application, without requiring additional session overload. This document defines two

leaf discovery mechanisms, one is based on T-LDP, the other is based on MP-BGP. Generally, the root initiated application with LDP as the main signaling mechanism, e.g, P2MP PW [I-D.ietf-pwe3-p2mp-pw], would use leaf discovery mechanism based on T-LDP, while application with MP-BGP as main signaling mechanism, e.g, VPLS Multicast [I-D.ietf-l2vpn-vpls-mcast], L3VPN Multicast [RFC6513] may use leaf discovery mechanism based on MP-BGP.

4.1. Leaf discovery mechanism based on T-LDP

This section will introduce the discovery mechanism based on T-LDP session. Each leaf node will report the leaf node information to root through this T-LDP session. It is required that there is a T-LDP session existed between each leaf node and root node. mLDP leaf discovery function will share the same mLDP P2MP capability described in section 2.1 of [RFC6388]

A LDP Label mapping message on the T-LDP session to the root with the MP FEC Element is used to convey the addition of the leaf membership to the root. The implicit NULL label is used to indicate that the mapping is from a leaf node. The Label Withdraw message is used to convey the deletion of the leaf membership to the root.

4.1.1. Node operation

The mLDP based P2MP/MP2MP LSP leaf discovery mechanism can be operated as follows.

For every leaf node, there will be a T-LDP session to be setup between root and leaf node. This T-LDP session can be setup automatically or manually, which depends on specific implementation.

When the leaf node is triggered to join one P2MP/MP2MP LSP, by various applications, the leaf node sends label mapping message to its upstream node (root or transit node). At the same time, the leaf node sends LDP label map message with MP FEC to its root node. When the root node receives the LDP label map message over T-LDP session with MP FEC, it will store the leaf node information associated with the specified P2MP/MP2MP LSP locally.

When the leaf node is triggered to leave one P2MP/MP2MP LSP, by various applications, the leaf node sends label withdraw message to its upstream node (root or transit node). At the same time, the leaf node sends LDP label withdraw message with MP FEC to its root node. When the root node receives the LDP label withdraw message over T-LDP with MP FEC, it will delete the leaf node information associated with the specified P2MP/MP2MP LSP locally.

4.2. Leaf discovery mechanism based on MP-BGP

This section will introduce the discovery mechanism based on MP-BGP[RFC4760]. Each leaf node will report the leaf node information to root through this BGP session.

4.2.1. mLDP leaf NLRI

This document defines a new BGP NLRI, called mLDP leaf NLRI. Following is the format of the mLDP leaf NLRI:

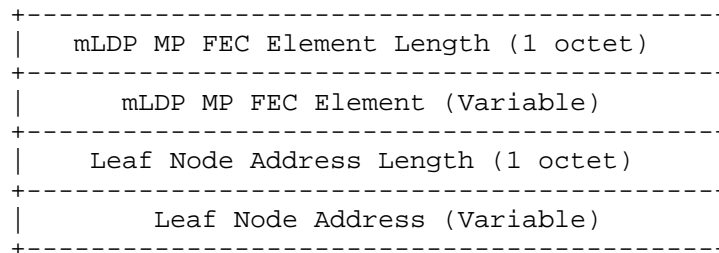


Figure 1

mLDP MP FEC Element may either contain P2MP FEC or MP2MP FEC element. Leaf Node Address field contains the leaf node IP address, and the value of length is 32 if it is IPv4 address, or 128 if it is IPv6 address. The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a mLDP MP FEC Element attached with Leaf Node Address. The mLDP leaf NLRI is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes [RFC4760]. The [AFI, SAFI] value pair used to identify this NLRI is (AFI=26 (AFI for MPLS Multicast, pending, IANA allocation), SAFI=8 (SAFI for mLDP leaf discovery, pending IANA allocation)).

In order for two BGP speakers to exchange mLDP leaf NLRI, they must use BGP Capabilities Advertisement to ensure that they both are capable of properly processing such NLRI. This is done as specified in [RFC4760], by using capability code 1 (multiprotocol BGP) with an AFI of 26 and an SAFI of mLDP leaf discovery.

The Next Hop field of MP_REACH_NLRI attribute shall be interpreted as an IPv4 address, whenever the length of the NextHop address is 4 octets, and as a IPv6 address, whenever the length of the NextHop address is 16 octets.

4.2.2. Node operation

The mLDP based P2MP/MP2MP LSP leaf discovery mechanism can be operated as follows.

When the leaf node is triggered to join one P2MP/MP2MP LSP, by various applications, the leaf node sends label mapping message to its upstream node (root or transit node). At the same time, the leaf node sends BGP UPDATE messages with MP_REACH_NLRI to its root node. The mLDP leaf NLRI will set Leaf Node Address to leaf node IP address, and next hop field to leaf node identifier. When the root node receives BGP UPDATE messages with MP_REACH_NLRI, it will store the leaf node information associated with the specified P2MP/MP2MP LSP locally.

When the leaf node is triggered to leave one P2MP/MP2MP LSP, by various applications, the leaf node sends label withdraw message to its upstream node (root or transit node). At the same time, the leaf node sends BGP UPDATE messages with MP_UNREACH_NLRI to its root node. The mLDP leaf NLRI will set Leaf Node Address to leaf node IP address, and next hop field to leaf node identifier. When the root node receives BGP UPDATE messages with MP_UNREACH_NLRI, it will delete the leaf node information associated with the specified P2MP/MP2MP LSP locally.

To constrain distribution of the mLDP leaf NLRI to the AS of the advertising PE the BGP Update message originated by the advertising PE SHOULD carry the NO_EXPORT Community [RFC1997].

5. Scalability

As recommended in section 4, leaf discovery will reuse the same signaling session as application, and will not setup additional sessions. For the application that uses T-LDP to do leaf discovery, all the leaf nodes have to setup T-LDP session to root node. There may be too many T-LDP sessions that have to be setup on the root node in the network, which will cause some scalability problem. This problem is caused by the application and out of scope of this draft.

6. Security Considerations

The same security considerations apply as for the multicast LDP specification, as described in [I-D. draft-ietf-mpls-ldp-p2mp], and MP-BGP, as described in [RFC4760].

7. IANA Considerations

7.1. MP-BGP

This document requires allocation of a new BGP AFI and SAFI.

A new AFI is allocated for MPLS Multicast function, the requested value has been pre-allocated as 26.

A new BGP SAFI for "Network Layer Reachability Information used for mLDP leaf discovery" from the IANA "Subsequence Address Family Identifiers (SAFI)" registry. The requested value has been pre-allocated as 8.

8. Acknowledgement

The author would like to thank Rahul Aggarwal, Dimitri Papadimitriou, IJsbrand Wijnands, Sandeep Bishnoi, Frederic Jounay and Simon DeLord for their valuable comments.

9. References

9.1. Normative references

- [RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

9.2. Informative References

- [I-D.ietf-l2vpn-vpls-mcast] Aggarwal, R., Rekhter, Y., Kamite, Y., and L. Fang, "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-10 (work in progress), February 2012.
- [I-D.ietf-mpls-mldp-in-band-signaling] Wijnands, I., Eckert, T., Leymann, N., and M. Napierala, "Multipoint LDP in-band signaling for Point-to-Multipoint and Multipoint- to-Multipoint Label Switched Paths", draft-ietf-mpls-mldp-in-band-signaling-05 (work in progress), February 2012.

progress), December 2011.

[I-D.ietf-pwe3-p2mp-pw]

Sivabalan, S., Boutros, S., and L. Martini, "Signaling Root-Initiated Point-to-Multipoint Pseudowire using LDP", draft-ietf-pwe3-p2mp-pw-04 (work in progress), March 2012.

[I-D.ietf-pwe3-p2mp-pw-requirements]

Bocci, M., Heron, G., and Y. Kamite, "Requirements and Framework for Point-to-Multipoint Pseudowires over MPLS PSNs", draft-ietf-pwe3-p2mp-pw-requirements-05 (work in progress), September 2011.

[RFC6388]

Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

[RFC6513]

Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

Authors' Addresses

Lizhong Jin
ZTE Corporation
889, Bibo Road
Shanghai, 201203, China

Email: lizhong.jin@zte.com.cn

Kebo Liu
Nokia Siemens Networks
1122 North Qinzhou Road
Shanghai, 200233, China

Email: kebo.liu@nsn.com

Sriganesh Kini
Ericsson
300 Holger Way
San Jose, CA 95134

Email: sriganesh.kini@ericsson.com

MPLS Working Group
Internet Draft

A. D'Alessandro
Telecom Italia
M. Paul
Deutsche Telekom
S. Ueno
NTT Communications
Y. Koike
NTT

Intended status: Informational

Expires: April 30, 2012

October 31, 2011

Temporal and hitless path segment monitoring
draft-koike-mpls-tp-temporal-hitless-psm-04.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 30, 2011.

Copyright Notice

Copyright (c) 2011 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.

Abstract

The MPLS transport profile (MPLS-TP) is being standardized to enable carrier-grade packet transport and complement converged packet network deployments. Among the most attractive features of MPLS-TP are OAM functions, which enable network operators or service providers to provide various maintenance characteristics, such as fault location, survivability, performance monitoring, and preliminary or in-service measurements.

One of the most important mechanisms which is common for transport network operation is fault location. A segment monitoring function of a transport path is effective in terms of extension of the maintenance work and indispensable particularly when the OAM function is effective only between end points. However, the current approach defined for MPLS-TP for the segment monitoring (SPME) has some fatal drawbacks.

This document elaborates on the problem statement for the Sub-path Maintenance Elements (SPMEs) which provides monitoring of a portion of a set of transport paths (LSPs or MS-PWs). Based on the problems, this document specifies new requirements to consider a new improved mechanism of hitless transport path segment monitoring.

This document is a product of a joint Internet Engineering Task Force (IETF) / International Telecommunications Union Telecommunications Standardization Sector (ITU-T) effort to include an MPLS Transport Profile within the IETF MPLS and PWE3 architectures to support the capabilities and functionalities of a packet transport network.

Table of Contents

1. Introduction	4
2. Conventions used in this document.....	4
2.1. Terminology	5
2.2. Definitions	5
3. Network objectives for monitoring.....	5

4. Problem statement	5
5. OAM functions for segment monitoring	9
6. Further consideration of requirements for enhanced segment monitoring	10
6.1. Necessity of on-demand single-layer monitoring.....	10
6.2. Necessity of on-demand monitoring independent from proactive monitoring	11
6.3. On-demand diagnostic procedures	12
7. Conclusion	13
8. Security Considerations.....	14
9. IANA Considerations	14
10. References	14
10.1. Normative References.....	14
10.2. Informative References.....	15
11. Acknowledgments	15

1. Introduction

A packet transport network will enable carriers or service providers to use network resources efficiently, reduce operational complexity and provide carrier-grade network operation. Appropriate maintenance functions, supporting fault location, survivability, performance monitoring and preliminary or in-service measurements, are essential to ensure quality and reliability of a network. They are essential in transport networks and have evolved along with TDM, ATM, SDH and OTN.

Unlike in SDH or OTN networks, where OAM is an inherent part of every frame and frames are also transmitted in idle mode, it is not per se possible to constantly monitor the status of individual connections in packet networks. Packet-based OAM functions are flexible and selectively configurable according to operators' needs.

According to the MPLS-TP OAM requirements [1], mechanisms MUST be available for alerting a service provider of a fault or defect affecting the service(s) provided. In addition, to ensure that faults or degradations can be localized, operators need a method to analyze or investigate the problem. From the fault localization perspective, end-to-end monitoring is insufficient. Using end-to-end OAM monitoring, when one problem occurs in an MPLS-TP network, the operator can detect the fault, but is not able to localize it.

Thus, a specific segment monitoring function for detailed analysis, by focusing on and selecting a specific portion of a transport path, is indispensable to promptly and accurately localize the fault.

For MPLS-TP, a path segment monitoring function has been defined to perform this task. However, as noted in the MPLS-TP OAM Framework[5], the current method for segment monitoring function of a transport path has implications that hinder the usage in an operator network.

This document elaborates on the problem statement for the path segment monitoring function and proposes to consider a new improved method of the segment monitoring, following up the work done in [5]. Moreover, this document explains detailed requirements on the new temporal and hitless segment monitoring function which are not covered in [5].

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [1].

2.1. Terminology

HSPM Hitless Path Segment Monitoring

LSP Label Switched Path

OTN Optical Transport Network

PST Path Segment Tunnel

TCM Tandem connection monitoring

SDH Synchronous Digital Hierarchy

SPME Sub-path Maintenance Element

2.2. Definitions

None

3. Network objectives for monitoring

There are two indispensable network objectives for MPLS-TP networks as described in section 3.8 of [5].

(1) The monitoring and maintenance of current transport paths has to be conducted in-service without traffic disruption.

(2) Segment monitoring must not modify the forwarding of the segment portion of the transport path.

It is common in transport networks that network objective (1) is mandatory and that regarding network objective (2) the monitoring shall not change the forwarding behavior.

4. Problem statement

To monitor, protect, or manage portions of transport paths, such as LSPs in MPLS-TP networks, the Sub-Path Maintenance Element (SPME) is defined in [2]. The SPME is defined between the edges of the portion of the transport path that needs to be monitored, protected, or managed. This SPME is created by stacking the shim header (MPLS header)[3] and is defined as the segment where the header is stacked. OAM messages can be initiated at the edge of the SPME and sent to the peer edge of the SPME or to a MIP along the SPME by setting the TTL value of the label stack entry (LSE) and interface identifier value at the corresponding hierarchical LSP level in case of per-node model.

This method has the following general issues, which are fatal in terms of cost and operation.

(P-1) Increasing the overhead by the stacking of shim header(s)

(P-2) Increasing the address management complexity, as new MEPs and MIPs need to be configured for the SPME in the old MEG

Problem (P-1) leads to decreased efficiency as bandwidth is wasted only for maintenance purposes. As the size of monitored segments increases, the size of the label stack grows. Moreover, if the operator wants to monitor the portion of a transport path without service disruption, one or more SPMEs have to be set in advance until the end of life of a transport path, which is not temporal or on-demand. Consuming additional bandwidth permanently for only the monitoring purpose should be avoided to maximize the available bandwidth.

Problem (P-2) is related to an identifier-management issue. The identification of each layer in case of LSP label stacking is required in terms of strict sub-layer management for the segment monitoring in a MPLS-TP network. There is no standardized way to identify a layer, however a possible rule of differentiating layers will be necessary at least within an administrative domain, if SPME is applied for on-demand OAM functions. This enforces operators to create an additional permanent layer identification policy only for temporal path segment monitoring. Moreover, from the perspective of operation, increasing the managed addresses and the managed layer is not desirable in terms of simplified operation featured by current transport networks. Reducing the managed identifiers and managed layers should be the fundamental direction in designing the architecture.

The most familiar example for SPME in transport networks is Tandem Connection Monitoring (TCM), which can for example be used for a carrier's carrier solution, as shown in Fig. 17 of the framework document[2]. However, in this case, the SPMEs have to be pre-configured. If this solution is applied to specific segment monitoring within one operator domain, all the necessary specific segments have to be pre-configured. This setting increases the managed objects as well as the necessary bandwidth, shown as Problem (P-1) and (P-2). Moreover, as a result of these pre-configurations, they impose operators to pre-design the structure of sub-path maintenance elements, which is not preferable in terms of operators' increased burden. These concerns are summarized in section 3.8 of [5].

Furthermore, in reality, all the possible patterns of path segment cannot be set in SPME, because overlapping of path segments is limited to nesting relationship. As a result, possible SPME patterns of portions of an original transport path are limited due to the characteristic of SPME shown in Figure.1, even if SPMEs are pre-configured. This restriction is inconvenient when operators have to fix issues in an on-demand manner. To avoid these issues, the temporal and on-demand setting of the SPME(s) is needed and more efficient for monitoring in MPLS-TP transport network operation.

However, using currently defined methods, the temporal setting of SPMEs also causes the following problems due to label stacking, which are fatal in terms of intrinsic monitoring and service disruption.

(P'-1) Changing the condition of the original transport path by changing the length of all the MPLS frames and changing label value(s)

(P'-2) Disrupting client traffic over a transport path, if the SPME is temporally configured.

Problem (P'-1) is a fatal problem in terms of intrinsic monitoring. As shown in network objective (2), the monitoring function needs to monitor the status without changing any conditions of the targeted monitored segment or the transport path. If the conditions of the transport path change, the measured value or observed data will also change. This can make the monitoring meaningless because the result of the monitoring would no longer reflect the reality of the connection where the original fault or degradation occurred.

Another aspect is that changing the settings of the original shim header should not be allowed because those changes correspond to creating a new portion of the original transport path, which differs from the original data plane conditions.

Figure 1 shows an example of SPME setting. In the figure, X means the one label expected on the tail-end node D of the original transport path. "210" and "220" are label allocated for SPME. The label values of the original path are modified as well as the values of stacked label. As shown in Fig.1, SPME changes the length of all the MPLS frames and changes label value(s). This is no longer the monitoring of the original transport path but the monitoring of a different path. Particularly, performance monitoring measurement (Delay measurement and loss measurement) are sensitive to those changes.

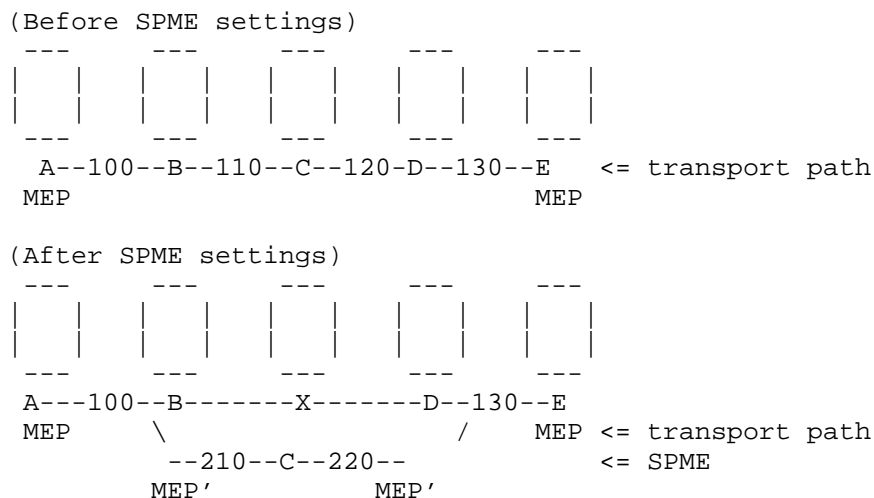


Figure 1 : An Example of a SPME setting

Problem (P'-2) was not fully discussed, although the make-before-break procedure in the survivability document [4] seemingly supports the hitless configuration for monitoring according to the framework document [2]. The reality is the hitless configuration of SPME is impossible without affecting the conditions of the targeted transport path, because the make-before-break procedure is premised on the change of the inner label value. This means changing one of the settings in MPLS shim header.

Moreover, this might not be effective under the static model without a control plane because the make-before-break is a restoration application based on the control plane. The removal of SPME whose segment is monitored could have the same impact (disruption of client traffic) as the creation of an SPME on the same LSP.

Note: (P'-2) will be removed when non-disruptive make-before-break (in both with and without C-plane environment) is specified in other MPLS-TP documents. However, (P'-2) could be replaced with the following issue. 'Non-disruptive MBB, in other words, taking an action similar to switching just for monitoring is not an ideal operation in transport networks.

The other potential risks are also envisaged. Setting up a temporal SPME will result in the LSRs within the monitoring segment only looking at the added (stacked) labels and not at the labels of the original LSP. This means that problems stemming from incorrect (or unexpected) treatment of labels of the original LSP by the nodes

within the monitored segment could not be found when setting up SPME. This might include hardware problems during label look-up, mis-configuration etc. Therefore operators have to pay extra attention to correctly setting and checking the label values of the original LSP in the configuration. Of course, the inversion of this situation is also possible, .e.g., incorrect or unexpected treatment of SPME labels can result in false detection of a fault where none of the problem originally existed.

The utility of SPMEs is basically limited to inter-carrier or inter-domain segment monitoring where they are typically pre-configured or pre-instantiated. SPME instantiates a hierarchical transport path (introducing MPLS label stacking) through which OAM packets can be sent. SPME construct monitoring function is particularly important mainly for protecting bundles of transport paths and carriers' carrier solutions. SPME is expected to be mainly used for protection purpose within one administrative domain.

To summarize, the problem statement is that the current sub-path maintenance based on a hierarchical LSP (SPME) is problematic for pre-configuration in terms of increasing bandwidth by label stacking and managing objects by layer stacking and address management. A on-demand/temporal configuration of SPME is one of the possible approaches for minimizing the impact of these issues. However, the current method is unfavorable because the temporal configuration for monitoring can change the condition of the original monitored transport path(and disrupt the in-service customer traffic). From the perspective of monitoring in transport network operation, a solution avoiding those issues or minimizing their impact is required. Another monitoring mechanism is therefore required that supports temporal and hitless path segment monitoring. Hereafter it is called on-demand hitless path segment monitoring (HPSM).

Note: The above sentence ''and disrupt the in-service customer traffic'' might need to be modified depending on the result of future discussion about (P'-2).

5. OAM functions using segment monitoring

OAM functions in which on-demand HPSM is required are basically limited to on-demand monitoring which are defined in OAM framework document [5], because those segment monitoring functions are used to locate the fault/degraded point or to diagnose the status for detailed analyses, especially when a problem occurred. In other words, the characteristic of "on-demand" is generally temporal for maintenance operation. Conversely, this could be a good reason that operations should not be based on pre-configuration and pre-design.

Packet loss and packet delay measurements are OAM functions in which hitless and temporal segment monitoring are strongly required because these functions are supported only between end points of a transport path. If a fault or defect occurs, there is no way to locate the defect or degradation point without using the segment monitoring function. If an operator cannot locate or narrow the cause of the fault, it is quite difficult to take prompt action to solve the problem. Therefore, on-demand HPSM for packet loss and packet delay measurements are indispensable for transport network operation.

Regarding other on-demand monitoring functions path segment monitoring is desirable, but not as urgent as for packet loss and packet delay measurements.

Regarding out-of-service on-demand monitoring functions, such as diagnostic tests, there seems no need for HPSM. However, specific segment monitoring should be applied to the OAM function of diagnostic test, because SPME doesn't meet network objective (2) in section 3. See section 6.3.

Note:

The solution for temporal and hitless segment monitoring should not be limited to label stacking mechanisms based on pre-configuration, such as PST/TCM(label stacking), which can cause the issues (P-1) and (P-2) described in Section 4.

The solution for HPSM has to cover both per-node model and per-interface model which are specified in [5].

6. Further consideration of requirements for enhanced segment monitoring

6.1. Necessity of on-demand single-level monitoring

The new segment monitoring function is supposed to be applied mainly for diagnostic purpose on-demand. We can differentiate this monitoring from the proactive segment monitoring as on-demand multi-level monitoring. The most serious problem at the moment is that there is no way to localize the degradation point on a path without changing the conditions of the original path. Therefore, as a first step, single layer segment monitoring not affecting the monitored path is required for a new on-demand and hitless segment monitoring function.

A combination of multi-level and simultaneous monitoring is the most powerful tool for accurately diagnosing the performance of a transport path. However, considering the substantial benefits to operators, a strict monitoring function which is required in such as a test environment of a laboratory does not seem to be necessary in the field. To summarize, on-demand and in-service (hitless) single-level segment monitoring is required, on-demand and in-service multi-level segment monitoring is desirable. Figure 2 shows an example of a multi-level on-demand segment monitoring.

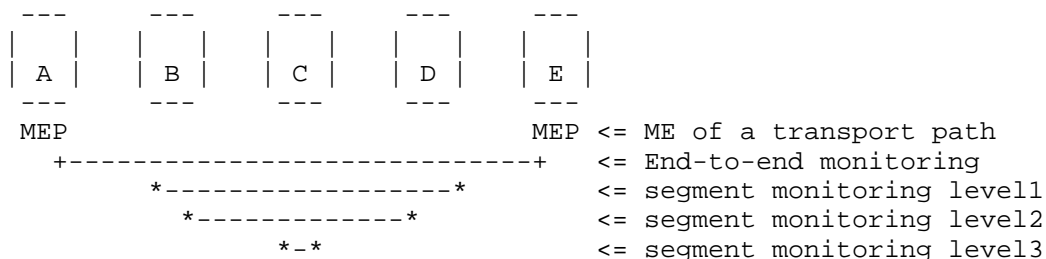


Figure 2 : An Example of a multi-level on-demand segment monitoring

6.2. Necessity of on-demand monitoring independent from end-to-end proactive monitoring

As multi-level simultaneous monitoring only for on-demand new path segment monitoring was already discussed in section 6.1, next we consider the necessity of simultaneous monitoring of end-to-end current proactive monitoring and new on-demand path segment monitoring. Normally, the on-demand path segment monitoring is configured in a segment of a maintenance entity of a transport path. In this environment, on-demand single-level monitoring should be done without disrupting pro-active monitoring of the targeted end-to-end transport path.

If operators have to disable the pro-active monitoring during the on-demand hitless path segment monitoring, the network operation system might miss any performance degradation of user traffic. This kind of inconvenience should be avoided in the network operations.

Accordingly, the on-demand single level path segment monitoring is required without changing or interfering the proactive monitoring of the original end-to-end transport path.

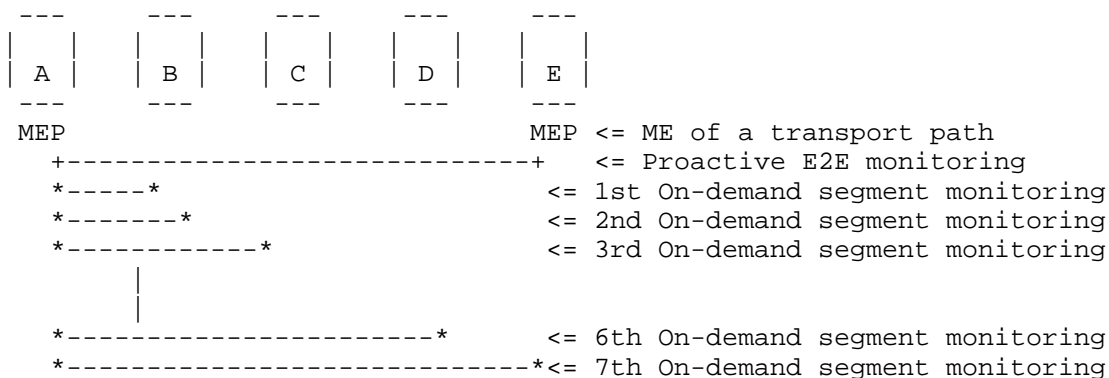


Figure 4 : One possible procedure to localize a fault point by sequential on-demand segment monitoring

Accordingly, on-demand monitoring of arbitrary segments is mandatory in the case in Fig. 5. As a result, on-demand HSPM should be set in an arbitrary segment of a transport path and diagnostic packets should be inserted from at least any of intermediate maintenance points of the original ME.

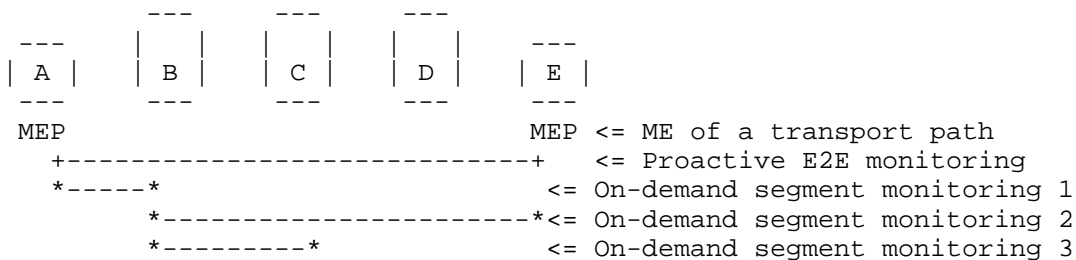


Figure 5 : Example where on-demand monitoring has to be configured in arbitrary segments

7. Conclusion

It is requested that another monitoring mechanism is required to support temporal and hitless segment monitoring which meets the two network objectives mentioned in Section 3 of this draft that are described also in section 3.8 of [5].

The enhancements should minimize the issues described in Section 4,, i.e., P-1, P-2, P'-1(and P'-2,) to meet those two network objectives.

The solution for the temporal and hitless segment monitoring has to cover both per-node model and per-interface model which are specified in [5]. In addition, the following requirements should be considered for an enhanced temporal and hitless path segment monitoring function.

Note: (P'-2) needs to be reconsidered.- An on-demand and in-service ''single-level'' segment monitoring is mandatory. Multi-level segment monitoring is optional.

- ''On-demand and in-service'' single level segment should be done without changing or interfering any condition of pro-active monitoring of an original ME of a transport path.

- On-demand and in-service segment monitoring should be able to be set in an arbitrary segment of a transport path.

The followings are specific requirements on each on-demand OAM function. Mandatory: Packet Loss Measurement and Packet Delay Measurement

Option: Connectivity verification, Diagnostic Tests (Throughput test), Route tracing

8. Security Considerations

This document does not by itself raise any particular security considerations.

9. IANA Considerations

There are no IANA actions required by this draft.

10. References

10.1. Normative References

- [1] Vigoureux, M., Betts, M., Ward, D., "Requirements for OAM in MPLS Transport Networks", RFC5860, May 2010
- [2] Bocci, M., et al., "A Framework for MPLS in Transport Networks", RFC5921, July 2010
- [3] Rosen, E., et al., "MPLS Label Stack Encoding", RFC 3032, January 2001

- [4] Sprecher, N., Farrel, A. , ''Multiprotocol Label Switching Transport Profile Survivability Framework'', draft-ietf-mpls-tp-survive-fwk-06.txt(work in progress), June 2010
- [5] Busi, I., Dave, A. , "Operations, Administration and Maintenance Framework for MPLS-based Transport Networks ", draft-ietf-mpls-tp-oam-framework-11.txt(work in progress), February 2011

10.2. Informative References

None

11. Acknowledgments

The author would like to thank all members (including MPLS-TP steering committee, the Joint Working Team, the MPLS-TP Ad Hoc Group in ITU-T) involved in the definition and specification of MPLS Transport Profile.

The authors would also like to thank Alexander Vainshtein, Dave Allan, Fei Zhang, Huub van Helvoort, Italo Busi, Maarten Vissers, Malcolm Betts and Nurit Sprecher for their comments and enhancements to the text.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Alessandro D'Alessandro
Telecom Italia
Email: alessandro.dalessandro@telecomitalia.it

Manuel Paul
Deutsche Telekom
Email: Manuel.Paul@telekom.de

Satoshi Ueno
NTT Communications
Email: satoshi.ueno@ntt.com

Yoshinori Koike
NTT
Email: koike.yoshinori@lab.ntt.co.jp

Network Working Group
Internet-Draft
Updates: 3031 (if approved)
Intended status: Standards Track
Expires: September 7, 2011

K. Kompella
J. Drake
Juniper Networks
S. Amante
Level 3 Communications, LLC
W. Henderickx
Alcatel-Lucent
L. Yong
Huawei USA
March 6, 2011

The Use of Entropy Labels in MPLS Forwarding
draft-kompella-mpls-entropy-label-02

Abstract

Load balancing is a powerful tool for engineering traffic across a network. This memo suggests ways of improving load balancing across MPLS networks using the concept of "entropy labels". It defines the concept, describes why entropy labels are useful, enumerates properties of entropy labels that allow maximal benefit, and shows how they can be signaled and used for various applications.

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 September 7, 2011.

Copyright Notice

Copyright (c) 2011 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. Conventions used	4
1.2. Motivation	5
2. Approaches	6
3. Entropy Labels	7
4. Data Plane Processing of Entropy Labels	8
4.1. Ingress LSR	8
4.2. Transit LSR	9
4.3. Egress LSR	9
5. Signaling for Entropy Labels	9
5.1. LDP Signaling	10
5.2. BGP Signaling	11
5.3. RSVP-TE Signaling	12
6. Operations, Administration, and Maintenance (OAM) and Entropy Labels	13
7. MPLS-TP and Entropy Labels	14
8. Point-to-Multipoint LSPs and Entropy Labels	15
9. Entropy Labels and Applications	15
9.1. Tunnels	15
9.2. LDP Pseudowires	17
9.3. BGP Applications	18
9.3.1. Inter-AS BGP VPNs	19
9.4. Multiple Applications	20
10. Security Considerations	21
11. IANA Considerations	22
11.1. LDP Entropy Label TLV	22
11.2. BGP Entropy Label Attribute	22
11.3. Attribute Flags for LSP_Attributes Object	22
11.4. Attributes TLV for LSP_Attributes Object	22
12. Acknowledgments	23
13. References	23
13.1. Normative References	23
13.2. Informative References	23
Appendix A. Applicability of LDP Entropy Label sub-TLV	24
Authors' Addresses	25

1. Introduction

Load balancing, or multi-pathing, is an attempt to balance traffic across a network by allowing the traffic to use multiple paths. Load balancing has several benefits: it eases capacity planning; it can help absorb traffic surges by spreading them across multiple paths; it allows better resilience by offering alternate paths in the event of a link or node failure.

As providers scale their networks, they use several techniques to achieve greater bandwidth between nodes. Two widely used techniques are: Link Aggregation Group (LAG) and Equal-Cost Multi-Path (ECMP). LAG is used to bond together several physical circuits between two adjacent nodes so they appear to higher-layer protocols as a single, higher bandwidth 'virtual' pipe. ECMP is used between two nodes separated by one or more hops, to allow load balancing over several shortest paths in the network. This is typically obtained by arranging IGP metrics such that there are several equal cost paths between source-destination pairs. Both of these techniques may, and often do, co-exist in various parts of a given provider's network, depending on various choices made by the provider.

A very important requirement when load balancing is that packets belonging to a given 'flow' must be mapped to the same path, i.e., the same exact sequence of links across the network. This is to avoid jitter, latency and re-ordering issues for the flow. What constitutes a flow varies considerably. A common example of a flow is a TCP session. Other examples are an L2TP session corresponding to a given broadband user, or traffic within an ATM virtual circuit.

To meet this requirement, a node uses certain fields, termed 'keys', within a packet's header as input to a load balancing function (typically a hash function) that selects the path for all packets in a given flow. The keys chosen for the load balancing function depend on the packet type; a typical set (for IP packets) is the IP source and destination addresses, the protocol type, and (for TCP and UDP traffic) the source and destination port numbers. An overly conservative choice of fields may lead to many flows mapping to the same hash value (and consequently poorer load balancing); an overly aggressive choice may map a flow to multiple values, potentially violating the above requirement.

For MPLS networks, most of the same principles (and benefits) apply. However, finding useful keys in a packet for the purpose of load balancing can be more of a challenge. In many cases, MPLS encapsulation may require fairly deep inspection of packets to find these keys at transit LSRs.

One way to eliminate the need for this deep inspection is to have the ingress LSR of an MPLS Label Switched Path extract the appropriate keys from a given packet, input them to its load balancing function, and place the result in an additional label, termed the 'entropy label', as part of the MPLS label stack it pushes onto that packet.

The packet's MPLS entire label stack can then be used by transit LSRs to perform load balancing, as the entropy label introduces the right level of "entropy" into the label stack.

There are four key reasons why this is beneficial:

1. at the ingress LSR, MPLS encapsulation hasn't yet occurred, so deep inspection is not necessary;
2. the ingress LSR has more context and information about incoming packets than transit LSRs;
3. ingress LSRs usually operate at lower bandwidths than transit LSRs, allowing them to do more work per packet, and
4. transit LSRs do not need to perform deep packet inspection and can load balance effectively using only a packet's MPLS label stack.

This memo describes why entropy labels are needed and defines the properties of entropy labels; in particular how they are generated and received, and the expected behavior of transit LSRs. Finally, it describes in general how signaling works and what needs to be signaled, as well as specifics for the signaling of entropy labels for LDP ([RFC5036]), BGP ([RFC3107], [RFC4364]), and RSVP-TE ([RFC3209]).

1.1. Conventions used

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].

The following acronyms are used:

LSR: Label Switching Router;

LER: Label Edge Router;

PE: Provider Edge router;

CE: Customer Edge device; and

FEC: Forwarding Equivalence Class.

The term ingress (or egress) LSR is used interchangeably with ingress (or egress) LER. The term application throughout the text refers to an MPLS application (such as a VPN or VPLS).

A label stack (say of three labels) is denoted by <L1, L2, L3>, where L1 is the "outermost" label and L3 the innermost (closest to the payload). Packet flows are depicted left to right, and signaling is shown right to left (unless otherwise indicated).

The term 'label' is used both for the entire 32-bit label and the 20-bit label field within a label. It should be clear from the context which is meant.

1.2. Motivation

MPLS is very successful generic forwarding substrate that transports several dozen types of protocols, most notably: IP, PWE3, VPLS and IP VPNs. Within each type of protocol, there typically exist several variants, each with a different set of load balancing keys, e.g., for IP: IPv4, IPv6, IPv6 in IPv4, etc.; for PWE3: Ethernet, ATM, Frame-Relay, etc. There are also several different types of Ethernet over PW encapsulation, ATM over PW encapsulation, etc. as well. Finally, given the popularity of MPLS, it is likely that it will continue to be extended to transport new protocols.

Currently, each transit LSR along the path of a given LSP has to try to infer the underlying protocol within an MPLS packet in order to extract appropriate keys for load balancing. Unfortunately, if the transit LSR is unable to infer the MPLS packet's protocol (as is often the case), it will typically use the topmost (or all) MPLS labels in the label stack as keys for the load balancing function. The result may be an extremely inequitable distribution of traffic across equal-cost paths exiting that LSR. This is because MPLS labels are generally fairly coarse-grained forwarding labels that typically describe a next-hop, or provide some of demultiplexing and/or forwarding function, and do not describe the packet's underlying protocol.

On the other hand, an ingress LSR (e.g., a PE router) has detailed knowledge of an packet's contents, typically through a priori configuration of the encapsulation(s) that are expected at a given PE-CE interface, (e.g., IPv4, IPv6, VPLS, etc.). They also have more flexible forwarding hardware. PE routers need this information and these capabilities to:

- a) apply the required services for the CE;
- b) discern the packet's CoS forwarding treatment;
- c) apply filters to forward or block traffic to/from the CE;
- d) to forward routing/control traffic to an onboard management processor; and,
- e) load-balance the traffic on its uplinks to transit LSRs (e.g., P routers).

By knowing the expected encapsulation types, an ingress LSR router can apply a more specific set of payload parsing routines to extract the keys appropriate for a given protocol. This allows for significantly improved accuracy in determining the appropriate load balancing behavior for each protocol.

If the ingress LSR were to capture the flow information so gathered in a convenient form for downstream transit LSRs, transit LSRs could remain completely oblivious to the contents of each MPLS packet, and use only the captured flow information to perform load balancing. In particular, there will be no reason to duplicate an ingress LSR's complex packet/payload parsing functionality in a transit LSR. This will result in less complex transit LSRs, enabling them to more easily scale to higher forwarding rates, larger port density, lower power consumption, etc. The idea in this memo is to capture this flow information as a label, the so-called entropy label.

Ingress LSRs can also adapt more readily to new protocols and extract the appropriate keys to use for load balancing packets of those protocols. This means that deploying new protocols or services in edge devices requires fewer concomitant changes in the core, resulting in higher edge service velocity and at the same time more stable core networks.

2. Approaches

There are two main approaches to encoding load balancing information in the label stack. The first allocates multiple labels for a particular Forwarding Equivalence Class (FEC). These labels are equivalent in terms of forwarding semantics, but having multiple labels allows flexibility in assigning labels to flows belonging to the same FEC. This approach has the advantage that the label stack has the same depth whether or not one uses label-based load balancing; and so, consequently, there is no change to forwarding operations on transit and egress LSRs. However, it has a major

drawback in that there is a significant increase in both signaling and forwarding state.

The other approach encodes the load balancing information as an additional label in the label stack, thus increasing the depth of the label stack by one. With this approach, there is minimal change to signaling state for a FEC; also, there is no change in forwarding operations in transit LSRs, and no increase of forwarding state in any LSR. The only purpose of the additional label is to increase the entropy in the label stack, so this is called an "entropy label". This memo focuses solely on this approach.

3. Entropy Labels

An entropy label (as used here) is a label:

1. that is not used for forwarding;
2. that is not signaled; and
3. whose only purpose in the label stack is to provide 'entropy' to improve load balancing.

Entropy labels are generated by an ingress LSR, based entirely on load balancing information. However, they MUST NOT have values in the reserved label space (0-15). Entropy labels MUST be at the bottom of the label stack, and thus the 'Bottom of Stack' (S) bit ([RFC3032]) in the label should be set. To ensure that they are not used inadvertently for forwarding, entropy labels SHOULD have a TTL of 0.

Since entropy labels are generated by an ingress LSR, an egress LSR MUST be able to tell unambiguously that a given label is an entropy label. If any ambiguity is possible, the label above the entropy label MUST be an 'entropy label indicator' (ELI), which indicates that the following Label is an entropy label. An ELI is typically signaled by an egress LSR and is added to the MPLS label stack along with an entropy label by an ingress LSR. For many applications, the use of entropy labels is unambiguous, and an ELI is not needed. If used, an ELI MUST have S = 0 and SHOULD have a TTL of 0.

Applications for MPLS entropy labels include pseudowires ([RFC4447]), Layer 3 VPNs ([RFC4364]), VPLS ([RFC4761], [RFC4762]) and Tunnel LSPs carrying, say, IP traffic. [I-D.ietf-pwe3-fat-pw] explains how entropy labels can be used for RFC 4447-style pseudowires, and thus is complementary to this memo, which focuses on several other applications of entropy labels.

4. Data Plane Processing of Entropy Labels

4.1. Ingress LSR

Suppose that for a particular application (or service or FEC), an ingress LSR X is to push label stack <TL, AL>, where TL is the 'tunnel label' and AL is the 'application label'. (Note the use of the convention for label stacks described in Section 1.1. The use of a two-label stack is just for illustrative purposes.) Suppose furthermore that the egress LSR Y has told X that it is capable of processing entropy labels for this application. If X can insert entropy labels, it may use a label stack of <TL, AL, EL> for this application, where EL is the entropy label.

When a packet for this application arrives at X, X does the following:

1. X identifies the application to which the packet belongs, identifies the egress LSR as Y, and thereby picks the outgoing label stack <TL, AL> to push onto the packet to send to Y;
2. X determines which keys that it will use for load balancing;
3. X, having kept state that Y can process entropy labels for this application, generates an entropy label EL (based on the output of the load balancing function), and
4. X pushes <TL, AL, EL> on to the packet before forwarding it to the next LSR on its way to Y.

EL is a 'regular' 32-bit label whose S bit MUST be 1 and whose TTL field SHOULD be 0. The load balancing information is encoded in the 20-bit label field. If X is told (via signaling) that it must use an entropy label indicator with label value E, then X instead pushes <TL, AL, ELI, EL> onto the packet, where ELI is a label whose S bit MUST be 0, whose TTL SHOULD be 0, and whose 20-bit label field MUST be E. The CoS fields for EL and ELI can be set to any values.

Note that ingress LSR X MUST NOT include an entropy label unless the egress LSR Y for this application has indicated that it is ready to receive entropy labels. Furthermore, if Y has signaled that an ELI is needed, then X MUST include the ELI before the entropy label.

Note that the signaling and use of entropy labels in one direction (signaling from Y to X, and data path from X to Y) has no bearing on the behavior in the opposite direction (signaling from X to Y, and data path from Y to X).

4.2. Transit LSR

Transit LSRs have virtually no change in forwarding behavior. For load balancing, transit LSRs SHOULD use the whole label stack as keys for the load balancing function. Transit LSRs MAY choose to look beyond the label stack for further keys; however, if entropy labels are being used, this may not be very useful. Looking beyond the label stack may be the simplest approach in an environment where some ingress LSRs use entropy labels and others don't, or for backward compatibility. Thus, other than using the full label stack as input to the load balancing function, transit LSRs are almost unaffected by the use of entropy labels.

4.3. Egress LSR

If egress LSR Y signals that it is capable of processing entropy labels without an ELI for an application, then when Y receives a packet with the application label, then Y looks to see if the S bit is set. If so, Y applies its usual processing rules to the packet, including popping the application label. If the S bit is not set, Y assumes that the label below the application label is an entropy label and pops both the application label and the entropy label. Y SHOULD ensure that the entropy label has its S bit set. Y then processes the packet as usual. Implementations may choose the order in which they apply these operations, but the net result should be as specified.

If Y signals that it is capable of processing entropy labels but that an ELI is necessary for a given application, then when Y receives a packet with the application label, Y processes the application label as usual, then pops it. Y then checks whether the S bit on the application label is set. If not, Y looks to see if the label below the application label is the ELI. If so, Y further pops both the ELI and the label below (which should be the entropy label). Y SHOULD ensure that the ELI has its S bit unset, and that the entropy label has its S bit set. If the S bit of the application label is set, or the label below is not the ELI, Y processes the packet as usual (there is no entropy label).

5. Signaling for Entropy Labels

An egress LSR Y may signal to ingress LSR(s) its ability to process entropy labels on a per-application (or per-FEC) basis. As part of this signaling, Y also signals the ELI to use, if any.

In cases where an application label is used and must be the bottommost label in the label stack, Y MAY signal that no ELI is

needed for that application.

In cases where no application label exists, or where the application label may not be the bottommost label in the label stack, Y MUST signal a valid ELI to be used in conjunction with the entropy label for this FEC. In this case, an ingress LSR will either not add an entropy label, or push the ELI before the entropy label. This makes the use or non-use of an entropy label by the ingress LSR unambiguous. Valid ELI label values are strictly greater than 15.

It should be noted that egress LSR Y may use the same ELI value for all applications for which an ELI is needed. The ELI MUST be a label that does not conflict with any other labels that Y has advertised to other LSRs for other applications. Furthermore, it should be noted that the ability to process entropy labels (and the corresponding ELI) may be asymmetric: an LSR X may be willing to process entropy labels, whereas LSR Y may not be willing to process entropy labels. The signaling extensions below allow for this asymmetry.

For an illustration of signaling and forwarding with entropy labels, see Figure 9.

5.1. LDP Signaling

When using LDP for signaling tunnel labels ([RFC5036]), a Label Mapping Message sub-TLV (Entropy Label sub-TLV) is used to signal an egress LSR's ability to process entropy labels.

The presence of the Entropy Label sub-TLV in the Label Mapping Message indicates to ingress LSRs that the egress LSR can process an entropy label. In addition, the Entropy Label sub-TLV contains a label value for the ELI. If the ELI is zero, this indicates the egress doesn't need an ELI for the signaled application; if not, the egress requires the given ELI with entropy labels. An example where an ELI is needed is when the signaled application is an LSP that can carry IP traffic.

The structure of the Entropy Label sub-TLV is shown below.

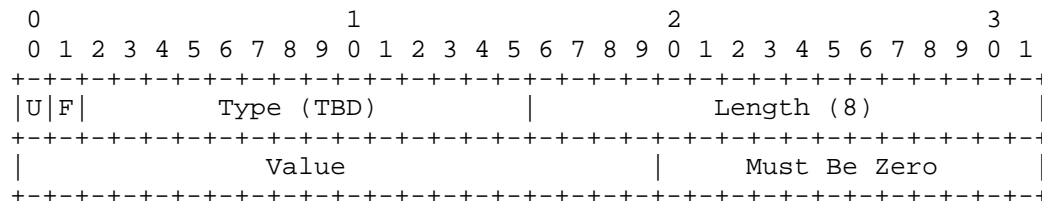


Figure 1: Entropy Label sub-TLV

where:

U: Unknown bit. This bit MUST be set to 1. If the Entropy Label sub-TLV is not understood, then the TLV is not known to the receiver and MUST be ignored.

F: Forward bit. This bit MUST be set to 1. Since this sub-TLV is going to be propagated hop-by-hop, the sub-TLV should be forwarded even by nodes that may not understand it.

Type: sub-TLV Type field, as specified by IANA.

Length: sub-TLV Length field. This field specifies the total length in octets of the Entropy Label sub-TLV.

Value: value of the Entropy Label Indicator Label.

5.2. BGP Signaling

When BGP [RFC4271] is used for distributing Network Layer Reachability Information (NLRI) as described in, for example, [RFC3107], [RFC4364] and [RFC4761], the BGP UPDATE message may include the Entropy Label attribute. This is an optional, transitive BGP attribute of type TBD. The inclusion of this attribute with an NLRI indicates that the advertising BGP router can process entropy labels as an egress LSR for that NLRI. If the attribute length is less than three octets, this indicates that the egress doesn't need an ELI for the signaled application. If the attribute length is at least three octets, the first three octets encode an ELI label value as the high order 20 bits; the egress requires this ELI with entropy labels. An example where an ELI is needed is when the NLRI contains unlabeled IP prefixes.

A BGP speaker S that originates an UPDATE should only include the Entropy Label attribute if both of the following are true:

A1: S sets the BGP NEXT_HOP attribute to itself; AND

A2: S can process entropy labels for the given application.

If both A1 and A2 are true, and S needs an ELI to recognize entropy labels, then S MUST include the ELI label value as part of the Entropy Label attribute. An UPDATE SHOULD contain at most one Entropy Label attribute.

Suppose a BGP speaker T receives an UPDATE U with the Entropy Label attribute ELA. T has two choices. T can simply re-advertise U with the same ELA if either of the following is true:

B1: T does not change the NEXT_HOP attribute; OR

B2: T simply swaps labels without popping the entire label stack and processing the payload below.

An example of the use of B1 is Route Reflectors; an example of the use of B2 is illustrated in Section 9.3.1.2.

However, if T changes the NEXT_HOP attribute for U and in the data plane pops the entire label stack to process the payload, T MUST remove ELA. T MAY include a new Entropy Label attribute ELA' for UPDATE U' if both of the following are true:

C1: T sets the NEXT_HOP attribute of U' to itself; AND

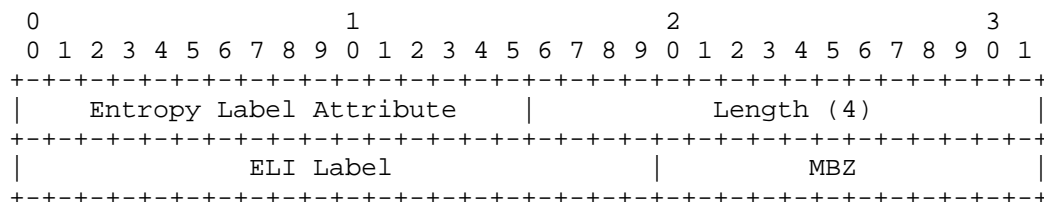
C2: T can process entropy labels for the given application.

Again, if both C1 and C2 are true, and T needs an ELI to recognize entropy labels, then T MUST include the ELI label value as part of the Entropy Label attribute.

5.3. RSVP-TE Signaling

Entropy Label support is signaled in RSVP-TE [RFC3209] using an Entropy Label Attribute TLV (Type TBD) of the LSP_ATTRIBUTES object [RFC5420]. The presence of this attribute indicates that the signaler (the egress in the downstream direction using Resv messages; the ingress in the upstream direction using Path messages) can process entropy labels. The Entropy Label Attribute contains a value for the ELI. If the ELI is zero, this indicates that the signaler doesn't need an ELI for this application; if not, then the signaler requires the given ELI with entropy labels. An example where an ELI is needed is when the signaled LSP can carry IP traffic.

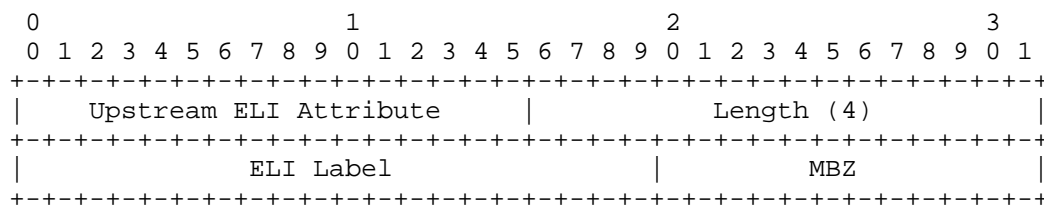
The format of the Entropy Label Attribute is as follows:



An egress LSR includes the Entropy Label Attribute in a Resv message to indicate that it can process entropy labels in the downstream direction of the signaled LSP.

An ingress LSR includes the Entropy Label Attribute in a Path message for a bi-directional LSP to indicate that it can process entropy labels in the upstream direction of the signaled LSP. If the signaled LSP is not bidirectional, the Entropy Label Attribute SHOULD NOT be included in the Path message, and egress LSR(s) SHOULD ignore the attribute, if any.

As described in Section 8, there is also the need to distribute an ELI from the ingress (upstream label allocation). In the case of RSVP-TE, this is accomplished using the Upstream ELI Attribute TLV of the LSP_ATTRIBUTES object, as shown below:



6. Operations, Administration, and Maintenance (OAM) and Entropy Labels

Generally OAM comprises a set of functions operating in the data plane to allow a network operator to monitor its network infrastructure and to implement mechanisms in order to enhance the general behavior and the level of performance of its network, e.g., the efficient and automatic detection, localization, diagnosis and handling of defects.

Currently defined OAM mechanisms for MPLS include LSP Ping/Traceroute [RFC4379] and Bidirectional Failure Detection (BFD) for MPLS [RFC5884]. The latter provides connectivity verification between the endpoints of an LSP, and recommends establishing a separate BFD session for every path between the endpoints.

The LSP traceroute procedures of [RFC4379] allow an ingress LSR to obtain label ranges that can be used to send packets on every path to the egress LSR. It works by having ingress LSR sequentially ask the transit LSRs along a particular path to a given egress LSR to return a label range such that the inclusion of a label in that range in a packet will cause the replying transit LSR to send that packet out the egress interface for that path. The ingress provides the label range returned by transit LSR N to transit LSR N + 1, which returns a label range which is less than or equal in span to the range provided to it. This process iterates until the penultimate transit LSR replies to the ingress LSR with a label range that is acceptable to it and to all LSRs along path preceding it for forwarding a packet along the path.

However, the LSP traceroute procedures do not specify where in the label stack the value from the label range is to be placed, whether deep packet inspection is allowed and if so, which keys and key values are to be used.

This memo updates LSP traceroute by specifying that the value from the label range is to be placed in the entropy label. Deep packet inspection is thus not necessary, although an LSR may use it, provided it do so consistently, i.e., if the label range to go to a given downstream LSR is computed with deep packet inspection, then the data path should use the same approach and the same keys.

In order to have a BFD session on a given path, a value from the label range for that path should be used as the EL value for BFD packets sent on that path.

As part of the MPLS-TP work, an in-band OAM channel is defined in [RFC5586]. Packets sent in this channel are identified with a reserved label, the Generic Associated Channel Label (GAL) placed at the bottom of the MPLS label stack. In order to use the inband OAM channel with entropy labels, this memo relaxes the restriction that the GAL must be at the bottom of the MPLS label stack. Rather, the GAL is placed in the MPLS label stack above the entropy label so that it effectively functions as an application label.

7. MPLS-TP and Entropy Labels

Since MPLS-TP does not use ECMP, entropy labels are not applicable to an MPLS-TP deployment.

8. Point-to-Multipoint LSPs and Entropy Labels

Point-to-Multipoint (P2MP) LSPs [RFC4875] typically do not use ECMP for load balancing, as the combination of replication and multipathing can lead to duplicate traffic delivery. However, P2MP LSPs can traverse Bundled Links [RFC4201] and LAGs. In both these cases, load balancing is useful, and hence entropy labels can be of some value for P2MP LSPs.

There are two potential complications with the use of entropy labels in the context of P2MP LSPs, both a consequence of the fact that the entire label stack below the P2MP label must be the same for all egress LSRs. First, all egress LSRs must be willing to receive entropy labels; if even one egress LSR is not willing, then entropy labels MUST NOT be used for this P2MP LSP. Second, if an ELI is required, all egress LSRs must agree to the same value of ELI. This can be achieved by upstream allocation of the ELI; in particular, for RSVP-TE P2MP LSPs, the ingress LSR distributes the ELI value using the Upstream ELI Attribute TLV of the LSP_ATTRIBUTES object, defined in Section 5.3.

With regard to the first issue, the ingress LSR MUST keep track of the ability of each egress LSR to process entropy labels, especially since the set of egress LSRs of a given P2MP LSP may change over time. Whenever an existing egress LSR leaves, or a new egress LSR joins the P2MP LSP, the ingress MUST re-evaluate whether or not to include entropy labels for the P2MP LSP.

In some cases, it may be feasible to deploy two P2MP LSPs, one to entropy label capable egress LSRs, and the other to the remaining egress LSRs. However, this requires more state in the network, more bandwidth, and more operational overhead (tracking EL-capable LSRs, and provisioning P2MP LSPs accordingly). Furthermore, this approach may not work for some applications (such mVPNs and VPLS) which automatically create and/or use P2MP LSPs for their multicast requirements.

9. Entropy Labels and Applications

This section describes the usage of entropy labels in various scenarios with different applications.

9.1. Tunnels

Tunnel LSPs, signaled with either LDP or RSVP-TE, typically carry other MPLS applications such as VPNs or pseudowires. This being the case, if the egress LSR of a tunnel LSP is willing to process entropy

labels, it would signal the need for an Entropy Label Indicator to distinguish between entropy labels and other application labels.

In the figures below, the following convention is used to depict information signaled between X and Y:

```

X ----- ... ----- Y
app:  <--- [label L, ELI value]
```

This means Y signals to X label L for application app. The ELI value can be one of:

-: meaning entropy labels are NOT accepted;

0: meaning entropy labels are accepted, no ELI is needed; or

E: entropy labels are accepted, ELI label E is required.

The following illustrates a simple intra-AS tunnel LSP.

```

X ----- A --- ... --- B ----- Y
tunnel LSP L:  [TL, E] <--- ... <--- [TL0, E]

IP pkt:       push <TL, E, EL> ----->
```

Figure 2: Tunnel LSPs and Entropy Labels

Tunnel LSPs may cross Autonomous System (AS) boundaries, usually using BGP ([RFC3107]). In this case, the AS Border Routers (ASBRs) MAY simply propagate the egress LSR's ability to process entropy labels, or they MAY declare that entropy labels may not be used. If an ASBR (say A2 below) chooses to propagate the egress LSR Y's ability to process entropy labels, A2 MUST also propagate Y's choice of ELI.

```

X ---- ... ---- A1 ----- A2 ---- ... ---- Y
intra-AS LSP A2-Y:                               <--- [TL0, E]
inter-AS LSP A1-A2:                               [AL, E]
intra-AS LSP X-A1: <--- [TL1, E]

IP pkt:       push <TL1, E, EL>
```

Here, ASBR A2 chooses to propagate Y's ability to process entropy labels, by "translating" Y's signaling of entropy label capability (say using LDP) to BGP; and A1 translate A2's BGP signaling to (say) RSVP-TE. The end-to-end tunnel (X to Y) will have entropy labels if

X chooses to insert them.

Figure 3: Inter-AS Tunnel LSP with Entropy Labels

```

                X ---- ... ---- A1 ----- A2 ---- ... ---- Y
intra-AS LSP A2-Y:                                <---- [TL0, E]
inter-AS LSP A1-A2:                                [AL, E]
intra-AS LSP X-A1: <--- [TL1, -]

IP pkt:                push <TL1> -->

```

Here, ASBR A1 decided that entropy labels are not to be used; thus, the end-to-end tunnel cannot have entropy labels, even though both X and Y may be capable of inserting and processing entropy labels.

Figure 4: Inter-AS Tunnel LSP with no Entropy Labels

9.2. LDP Pseudowires

[I-D.ietf-pwe3-fat-pw] describes the signaling and use of entropy labels in the context of RFC 4447 pseudowires, so this will not be described further here.

[RFC4762] specifies the use of LDP for signaling VPLS pseudowires. An egress VPLS PE that can process entropy labels can indicate this by adding the Entropy Label sub-TLV in the LDP message it sends to other PEs. An ELI is not required. An ingress PE must maintain state per egress PE as to whether it can process entropy labels.

```

                X ----- A --- ... --- B ----- Y
tunnel LSP L:   [TL, E] <--- ... <--- [TL0, E]
VPLS label:    <----- [VL, 0]

VPLS pkt:      push <TL, VL, EL> ----->

```

Figure 5: Entropy Labels with LDP VPLS

Note that although the underlying tunnel LSP signaling indicated the need for an ELI, VPLS packets don't need an ELI, and thus the label stack pushed by X do not have one.

[RFC4762] also describes the notion of "hierarchical VPLS" (H-VPLS). In H-VPLS, 'hub PEs' remove the label stack and process VPLS packets; thus, they must make their own decisions on the use of entropy labels, independent of other hub PEs or spoke PEs with which they exchange signaling. In the example below, spoke PEs X and Y and hub

PE B can process entropy labels, but hub PE A cannot.

```

      X ---- ... ---- A ---- ... ---- B ---- ... ---- Y
spoke PW1:                                <--- [SL1, 0]
hub-hub PW:                                <---- [HL, 0]
spoke PW2:                                <--- [SL2, -]

SPW2 pkt:      push <TL1, SL2>
H-H PW pkt:    push <TL2,HL,EL>
SPW1 pkt:      push <TL3,SL1,EL>

```

Figure 6: Entropy Labels with H-VPLS

9.3. BGP Applications

Section 9.1 described a BGP application for the creation of inter-AS tunnel LSPs. This section describes two other BGP applications, IP VPNs ([RFC4364]) and BGP VPLS ([RFC4761]). An egress PE for either of these applications indicates its ability to process entropy labels by adding the Entropy Label attribute to its BGP UPDATE message. Again, ingress PEs must maintain per-egress PE state regarding its ability to process entropy labels. In this section, both of these applications will be referred to as VPNs.

In the intra-AS case, PEs signal application labels and entropy label capability to each other, either directly, or via Route Reflectors (RRs). If RRs are used, they must not change the BGP NEXT_HOP attribute in the UPDATE messages; furthermore, they can simply pass on the Entropy Label attribute as is.

```

      X ----- A --- ... --- B ----- Y
tunnel LSP L:  [TL, E] <--- ... <--- [TL0, E]
BGP VPN label: <----- [VL, 0]

BGP VPN pkt:   push <TL, VL, EL> ----->

```

Figure 7: Entropy Labels with Intra-AS BGP apps

For BGP VPLS, the application label is at the bottom of stack, so no ELI is needed. For BGP IP VPNs, the application label is usually at the bottom of stack, so again no ELI is needed. However, in the case of Carrier's Carrier (CsC) VPNs, the BGP VPN label may not be at the bottom of stack. In this case, an ELI is necessary for CsC VPN packets with entropy labels to distinguish them from nested VPN packets. In the example below, the nested VPN signaling is not shown; the egress PE for the nested VPN (not shown) must signal

whether or not it can process egress labels, and the ingress nested VPN PE may insert an entropy label if so.

Three cases are shown: a plain BGP VPN packet, a CsC VPN packet originating from X, and a transit nested VPN packet originating from a nested VPN ingress PE (conceptually to the left of X). It is assumed that the nested VPN packet arrives at X with label stack <ZL, CVL> where ZL is the tunnel label (to be swapped with <TL, CL>) and CVL is the nested VPN label. Note that Y can use the same ELI for the tunnel LSP and the CsC VPN (and any other application that needs an ELI).

```

X ----- A --- ... --- B ----- Y
tunnel LSP L:      [TL,  E] <--- ... <--- [TL0, E]
BGP VPN label:    <----- [VL, 0]
BGP CsC VPN label: <----- [CL, E]

BGP VPN pkt:      push <TL, VL, EL> ----->
CsC VPN pkt:      push <TL, CL, E, EL> ----->
nested VPN pkt:   swap <ZL> with <TL, CL> ----->

```

Figure 8: Entropy Labels with CoC VPN

9.3.1. Inter-AS BGP VPNs

There are three commonly used options for inter-AS IP VPNs and BGP VPLS, known informally as "Option A", "Option B" and "Option C". This section describes how entropy labels can be used in these options.

9.3.1.1. Option A Inter-AS VPNs

In option A, an ASBR pops the full label stack of a VPN packet exiting an AS, processes the payload header (IP or Ethernet), and forwards the packet natively (i.e., as IP or Ethernet, but not as MPLS) to the peer ASBR. Thus, entropy label signaling and insertion are completely local to each AS. The inter-AS paths do not use entropy labels, as they do not use a label stack.

9.3.1.2. Option B Inter-AS VPNs

The ASBRs in option B inter-AS VPNs have a choice (usually determined by configuration) of whether to just swap labels (from within the AS to the neighbor AS or vice versa), or to pop the full label stack and process the packet natively. This choice occurs at each ASBR in each direction. In the case of native packet processing at an ASBR, entropy label signaling and insertion is local to each AS and to the

inter-AS paths (which, unlike option A, do have labeled packets).

In the case of simple label swapping at an ASBR, the ASBR can propagate received entropy label signaling onward. That is, if a PE signals to its ASBR that it can process entropy labels (via an Entropy Label attribute), the ASBR can propagate that attribute to its peer ASBR; if a peer ASBR signals that it can process entropy labels, the ASBR can propagate that to all PEs within its AS). Note that this is the case even though ASBRs change the BGP NEXT_HOP attribute to "self", because of clause B2 in Section 5.2.

9.3.1.3. Option C Inter-AS VPNs

In Option C inter-AS VPNs, the ASBRs are not involved in signaling; they do not have VPN state; they simply swap labels of inter-AS tunnels. Signaling is PE to PE, usually via Route Reflectors; however, if RRs are used, the RRs do not change the BGP NEXT_HOP attribute. Thus, entropy label signaling and insertion are on a PE-pair basis, and the intermediate routers, ASBRs and RRs do not play a role.

9.4. Multiple Applications

It has been mentioned earlier that an ingress PE must keep state per egress PE with regard to its ability to process entropy labels. An ingress PE must also keep state per application, as entropy label processing must be based on the application context in which a packet is received (and of course, the corresponding entropy label signaling).

In the example below, an egress LSR Y signals a tunnel LSP L, and is prepared to receive entropy labels on L, but requires an ELI. Furthermore, Y signals two pseudowires PW1 and PW2 with labels PL1 and PL2, respectively, and indicates that it can receive entropy labels for both pseudowires without the need of an ELI; and finally, Y signals a L3 VPN with label VL, but Y does not indicate that it can receive entropy labels for the L3 VPN. Ingress LSR X chooses to send native IP packets to Y over L with entropy labels, thus X must include the given ELI (yielding a label stack of <TL, ELI, EL>). X chooses to add entropy labels on PW1 packets to Y, with a label stack of <TL, PL1, EL>, but chooses not to do so for PW2 packets. X must not send entropy labels on L3 VPN packets to Y, i.e., the label stack must be <TL, VL>.

```

      X ----- A --- ... --- B ----- Y
tunnel LSP L:  [TL,  E] <--- ... <--- [TL0, E]
PW1 label:    <----- [PL1, 0]
PW2 label:    <----- [PL2, 0]
VPN label:    <----- [VL,  -]

IP pkt:       push <TL, ELI, EL> ----->
PW1 pkt:      push <TL, PL1, EL> ----->
PW2 pkt:      push <TL, PL2> ----->
VPN pkt:      push <TL, VL> ----->

```

Figure 9: Entropy Labels for Multiple Applications

10. Security Considerations

This document describes advertisement of the capability to support receipt of entropy-labels and an Entropy Label Indicator that an ingress LSR may apply to MPLS packets in order to allow transit LSRs to attain better load-balancing across LAG and/or ECMP paths in the network.

This document does not introduce new security vulnerabilities to LDP. Please refer to the Security Considerations section of LDP ([RFC5036]) for security mechanisms applicable to LDP.

Given that there is no end-user control over the values used for entropy labels, there is little risk of Entropy Label forgery which could cause uneven load-balancing in the network.

If Entropy Label Capability is not signaled from an egress PE to an ingress PE, due to, for example, malicious configuration activity on the egress PE, then the PE's will fall back to not using entropy labels for load-balancing traffic over LAG or ECMP paths which, in some cases, is no worse than the behavior observed in current production networks. That said, operators are recommended to monitor changes to PE configurations and, more importantly, the fairness of load distribution over equal-cost LAG or ECMP paths. If the fairness of load distribution over a set of paths changes that could indicate a misconfiguration, bug or other non-optimal behavior on their PE's and they should take corrective action.

Given that most applications already signal an Application Label, e.g.: IPVPNs, LDP VPLS, BGP VPLS, whose Bottom of Stack bit is being re-used to signal entropy label capability, there is little to no additional risk that traffic could be misdirected into an inappropriate IPVPN VRF or VPLS VSI at the egress PE.

In the context of downstream-signaled entropy labels that require the use of an Entropy Label Indicator (ELI), there should be little to no additional risk because the egress PE is solely responsible for allocating an ELI value and ensuring that ELI label value DOES NOT conflict with other MPLS labels it has previously allocated. On the other hand, for upstream-signaled entropy labels, e.g.: RSVP-TE point-to-point or point-to-multipoint LSP's or Multicast LDP (mLDP) point-to-multipoint or multipoint-to-multipoint LSP's, there is a risk that the head-end MPLS LER may choose an ELI value that is already in use by a downstream LSR or LER. In this case, it is the responsibility of the downstream LSR or LER to ensure that it MUST NOT accept signaling for an ELI value that conflicts with MPLS label(s) that are already in use.

11. IANA Considerations

11.1. LDP Entropy Label TLV

IANA is requested to allocate the next available value from the IETF Consensus range in the LDP TLV Type Name Space Registry as the "Entropy Label TLV".

11.2. BGP Entropy Label Attribute

IANA is requested to allocate the next available Path Attribute Type Code from the "BGP Path Attributes" registry as the "BGP Entropy Label Attribute".

11.3. Attribute Flags for LSP_Attributes Object

IANA is requested to allocate a new bit from the "Attribute Flags" sub-registry of the "RSVP TE Parameters" registry.

Bit	Name	Attribute	Attribute	RRO
No		Flags Path	Flags Resv	
TBD	Entropy Label LSP	Yes	Yes	No

11.4. Attributes TLV for LSP_Attributes Object

IANA is requested to allocate the next available value from the "Attributes TLV" sub-registry of the "RSVP TE Parameters" registry.

12. Acknowledgments

We wish to thank Ulrich Drafz for his contributions, as well as the entire 'hash label' team for their valuable comments and discussion.

13. References

13.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.

13.2. Informative References

- [I-D.ietf-pwe3-fat-pw] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow Aware Transport of Pseudowires over an MPLS PSN", draft-ietf-pwe3-fat-pw-05 (work in progress), October 2010.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379,

February 2006.

- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

Appendix A. Applicability of LDP Entropy Label sub-TLV

In the case of unlabeled IPv4 (Internet) traffic, the Best Current Practice is for an egress LSR to propagate eBGP learned routes within a SP's Autonomous System after resetting the BGP next-hop attribute to one of its Loopback IP addresses. That Loopback IP address is injected into the Service Provider's IGP and, concurrently, a label assigned to it via LDP. Thus, when an ingress LSR is performing a forwarding lookup for a BGP destination it recursively resolves the associated next-hop to a Loopback IP address and associated LDP label of the egress LSR.

Thus, in the context of unlabeled IPv4 traffic, the LDP Entropy Label sub-TLV will typically be applied only to the FEC for the Loopback IP address of the egress LSR and the egress LSR will not announce an entropy label capability for the eBGP learned route.

Authors' Addresses

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: jdrake@juniper.net

Shane Amante
Level 3 Communications, LLC
1025 Eldorado Blvd
Broomfield, CO 80021
US

Email: shane@level3.net

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp
Belgium

Email: wim.henderickx@alcatel-lucent.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075
US

Email: lucyyong@huawei.com

Network Working Group
Internet-Draft
Updates: 3209, 3473, 5420
(if approved)
Intended status: Standards Track
Expires: September 10, 2015

K. Kompella
Juniper Networks
M. Hellers
LINX
R. Singh
Juniper Networks
March 9, 2015

Multi-path Label Switched Paths Signaled Using RSVP-TE
draft-kompella-mpls-rsvp-ecmp-06.txt

Abstract

This document describes extensions to Resource ReSerVation Protocol - Traffic Engineering for the set up of multi-path Traffic Engineered Label Switched Paths (LSPs) in Multi Protocol Label Switching (MPLS) and Generalized MPLS networks, i.e., LSPs that conform to traffic engineering constraints, but follow multiple independent paths from source to destination.

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 September 10, 2015.

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. Terminology	3
1.2. Conventions used in this document	4
2. Theory of Operation	5
2.1. Multi-path Label Switched Paths	5
2.2. ECMP	6
2.3. Discussion	8
2.4. The Capabilities of TE-based Load Balancing	9
3. Operation of MLSPs	10
3.1. Signaling MLSPs	10
3.1.1. Indicating Equi-bandwidth (EB) nature	10
3.2. Label Allocation	10
3.3. Bandwidth Accounting	10
3.4. MLSP Data Plane Actions	11
4. Manageability	13
5. Security Considerations	14
6. Acknowledgments	15
7. IANA Considerations	16
8. References	17
8.1. Normative References	17
8.2. Informative References	17
Authors' Addresses	19

1. Introduction

In selecting a protocol for setting up and signaling "tunnel" Labeled Switched Paths (LSPs) in Multi Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks, one first chooses whether one wants Equal Cost Multi-Path (ECMP) load balancing or Traffic Engineering (TE). For the former, one uses the Label Distribution Protocol (LDP) ([RFC5036]); for the latter, the Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE) ([RFC3209]). [Two other criteria, the need for fast protection and the desire for less configuration, are no longer the deciding factors they used to be, thanks to "IP fast reroute" ([RFC5286]) and "RSVP-TE automesh" ([RFC4972])].

This document describes how one can set up a tunnel LSP that has both ECMP and TE characteristics using RSVP-TE. The techniques described in this document can be used to create a "Multipath LSP" (MLSP) to a destination, that consists of several "sub-LSPs", each potentially taking a different path through the network to the destination. The techniques can also be used to create a single MLSP to multiple equivalent destinations (such as equidistant BGP nexthops announcing a common set of reachable addresses), such that each destination is served by one or more sub-LSPs.

There are several alternatives to choose from when considering MLSPs. One is whether the ingress Label Switching Router (LSR) computes (or otherwise obtains) the full path for each sub-LSP, or whether LSRs along the various paths can compute paths further downstream (using techniques such as "loose hop expansion", as in [RFC5152]). Another is whether the various paths that make up the MLSP have equal cost (or distance) from ingress to egress (i.e., ECMP), whether they may have differing costs. Finally, one can choose whether to terminate a multi-path LSP on a single egress or on several equivalent egresses. For now, the first of each of these alternatives is assumed; future work can explore other choices.

1.1. Terminology

The term Multipath LSP, or MLSP, will be used to denote the (logical) container LSP from an ingress LSR to one or more egress LSR(s). An MLSP is the unit of configuration and management.

An MLSP consists of one or more "sub-LSPs". A sub-LSP consists of a single path from the ingress of the MLSPs to one of its egresses. A sub-LSP is the unit of signaling of an MLSP. An Explicit Route Object (ERO) will be used to define the path of a sub-LSP.

The "downstream links" of an MLSP Z at LSR X is the union of the

downstream links of all sub-LSPs of Z traversing X. Similarly, the "upstream links" of an MLSP Z at LSR X is the union of upstream links of all sub-LSPs of Z traversing X.

The agent that takes the configuration parameters of a tunnel and computes the corresponding paths is called the Path Computation Agent (PCA). The PCA is responsible for acquiring the tunnel configuration, computing the paths of the sub-LSPs, and, if the PCA is not co-located with the ingress, informing the ingress about the tunnel and the EROs for the sub-LSPs.

1.2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Theory of Operation

2.1. Multi-path Label Switched Paths

An MLSP is configured with various constraints associated with TE LSPs, such as destination LSR(s), bandwidth (on a per-class basis, if desired), link colors, Shared Risk Link Groups, etc. [Auto-mesh techniques ([RFC4972]) can be used to reduce configuration; this is not described further here.] In addition, parameters specifically related to MLSPs, such as how many (or the maximum number of) sub-LSPs to create, whether traffic should be split equally across sub-LSPs or not, etc. may also be specified. This configuration lives on the PCA, which is responsible for computing the paths (i.e., the EROs) for the various sub-LSPs. The PCA informs the ingress LSR about the MLSP and the constituent sub-LSPs, including EROs and bandwidths.

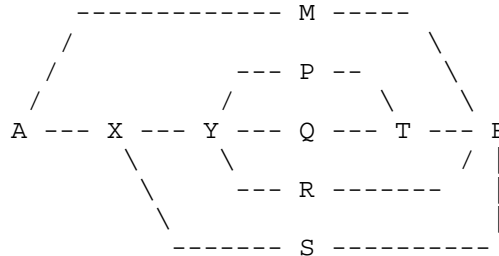
The PCA uses the configuration parameters to decide how many sub-LSPs to compute for this MLSP, what paths they should take, and how much bandwidth each sub-LSP is responsible for. Each sub-LSP MUST meet all the constraints of the MLSP (except bandwidth). The bandwidths (per-class, if applicable) of all the sub-LSPs MUST add up to the bandwidth of the MLSP. A Path Computation Element ([RFC4655]) that is multi-path LSP-aware may be used as the PCA.

Having computed (or otherwise obtained) the paths of all the sub-LSPs, the ingress A then signals the MLSP by signaling all the individual sub-LSPs across the MPLS/GMPLS network. To do this, the ingress first picks an MLSP ID, a 16-bit number that is unique in the context of the ingress. This ID is used in an ASSOCIATION object that is placed in each sub-LSP to let all transit LSRs know that the sub-LSPs belong to the same MLSP.

If multiple sub-LSPs of the same MLSP pass through LSR Y, and Y has downstream links YP, YQ and YR for the various sub-LSPs, then Y has to load balance incoming traffic for the MLSP across the three downstream links in proportion to the sum of the bandwidths of the sub-LSPs going to each downstream (see Figure 1).

One must distinguish carefully between the signaled bandwidth of a sub-LSP, a static value capturing the expected or maximum traffic on the sub-LSP, and the instantaneous traffic received on a sub-LSP, a constantly varying quantity. Suppose there are three sub-LSPs traversing Y, with bandwidths 10Gbps, 20Gbps and 30Gbps, going to P, Q and R respectively. Suppose further Y receives some traffic over each of these sub-LSPs. Y must balance this received traffic over the three downstream links YP, YQ and YR in the ratio 1:2:3.

2.2. ECMP



An example network illustrating ECMP. Assume that paths AMB, AXYP TB, AXYQTB, AXYRB and AXSB all have the same path length (cost).

Figure 1: Example Network Topology

In an IP or LDP network, incoming traffic arriving at A headed for B will be split equally between M and X at A. Similarly, traffic for B arriving at Y will be split equally among P, Q and R. If the traffic arriving at A for B is 120Gbps, then the AMB path will carry 60Gbps, the paths AXYP TB, AXYQTB and AXYRB will each carry 10Gbps, and the AXSB path will carry 30Gbps. We'll call this "IP-style" load balancing.

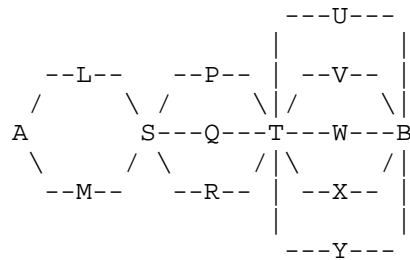
Note: all load balancing is subject to the overriding requirement of mapping the same "flow" to the same downstream. (What constitutes a "flow" is beyond the scope of this document.) This requirement takes precedence over all attempts to balance traffic among downstreams. Thus, the statements above (e.g., "the AMB path will carry 60Gbps") are to be interpreted as ideal targets, not hard requirements, of load balancing.

One can simulate the IP or LDP ECMP behavior with TE-based ECMP by creating an MLSP with five sub-LSPs S1 through S5 taking paths AMB, AXYP TB, AXYQTB, AXYRB and AXSB, with bandwidths 60Gbps, 10Gbps, 10Gbps, 10Gbps and 30Gbps, respectively.

With such an arrangement, the MB link carries 60Gbps while the RB link carries just 10Gbps. If one wishes instead to carry equal amounts of traffic on the links incoming to B, then one could arrange the sub-LSPs S1 to S5 to have bandwidths 30Gbps, 15Gbps, 15Gbps, 30Gbps and 30Gbps, respectively. In this case, the bandwidth on each of the four links going to B is 30Gbps, illustrating some of the capabilities of TE-based ECMP.

Staying with this example, A has one sub-LSP of bandwidth 30Gbps to M and four sub-LSPs of total bandwidth 90Gbps through X. Thus, A should

load balance traffic in the ratio 1:3 between the AM and the AX links. Similarly, X has three sub-LSPs of total bandwidth 60Gbps to Y and one sub-LSP of bandwidth 30Gbps to S, so X should load balance traffic 2:1 between Y and S. Y has a sub-LSP of bandwidth 15Gbps to each of P and Q and one sub-LSP of bandwidth 30Gbps to R, so Y should load balance traffic 1:1:2 among P, Q and R, respectively. Thus, in general, TE-based ECMP does not assume equal distribution of traffic among downstream LSRs, unlike IP- or LDP-style ECMP.



Another example network illustrating 30 ECMP paths between A and B.

Figure 2: Another Network Topology

In Figure 2, there are potentially $2 \times 3 \times 5 = 30$ ECMP paths between A and B. With IP or LDP, exploiting all these paths is straightforward, and doesn't need a lot of state. With an MLSP as seen so far, this would require 30 sub-LSPs to achieve equivalent load balancing. This suggests that a different approach is needed to efficiently achieve IP-style load balancing with TE LSPs. To this end, we introduce the notion of "equi-bandwidth" (EB) sub-LSPs and EB MLSPs. A sub-LSP is equi-bandwidth if its "E" bit is set (see Section 3.1.1). An MLSP is equi-bandwidth if all of its sub-LSPs are equi-bandwidth.

If a set of EB sub-LSPs of the same MLSP traverse an LSR S, say to downstream links SP, SQ and SR, then S MUST attempt to load balance traffic received on these EB sub-LSPs equally among the links SP, SQ and SR, independent of how many sub-LSPs go over each of these links. Furthermore, S MUST redistribute traffic received from each of its upstream LSRs, and SHOULD redistribute all traffic received from upstream as a whole. One can do the former by signaling the same label to each of its upstream LSRs; one can do the latter by signaling the same label to all upstream LSRs (see Section 3.2). For example, in Figure 2, if L sends 12Gbps of traffic to S and M sends 18Gbps to S, S can redistribute L's traffic by sending 4Gbps to each of P, Q and R; and can similarly send 6Gbps of M's traffic to each of P, Q and R. Alternatively, S can load balance the aggregate 30Gbps of traffic received from L and M to each of P, Q and R, thus sending 10Gbps to each. EB sub-LSPs have an added benefit of not requiring

unequal load balancing across links, which may pose problems for some hardware.

Given the notion of EB sub-LSPs and EB MLSPs, A can signal an EB MLSP Z comprised of five EB sub-LSPs E1 through E5 with the following paths: ALSPTUB, AMSQTVB, ALSRTWB, AMSPTXB and ALSQTYB (respectively). Then, A has two downstream links for the five sub-LSPs, AL and AM, between which A will load balance equally. Similarly, S has three downstream links, SP, SQ and SR; and T has five downstreams, TU, TV, TW, TX and TY. Thus the load balancing behavior of the MLSP will replicate IP load balancing. The state required for an EB MLSP to achieve IP-style load balancing is somewhat greater than for LDP LSPs, but significantly less than that for multiple "regular" TE LSPs, or for a non-EB MLSP.

2.3. Discussion

Some of the power of TE-based ECMP was illustrated in the above examples. Another is ability to request that all sub-LSPs avoid links colored red. If in the example network in Figure 1, the QT link is colored red but all other links are not, then there are four ECMP paths that satisfy these constraints, and the traffic distribution among them will naturally be different than it would without the link color constraint.

One can also ask whether an MLSP with sub-LSPs is any better than N "regular" LSPs from the same ingress to the same egress. Here are some benefits of an MLSP:

1. With an MLSP, there is a single entity to provision, manage and monitor, versus N separate entities in the case of LSPs. A consequence of this is that with an MLSP, changes in topology can be dealt with easily and autonomously by the ingress LSR, by adding, changing or removing sub-LSPs to rebalance traffic, while maintaining the same TE constraints. With individual LSPs, such changes would require changes in configuration, and thus are harder to automate.
2. An ingress LSR, knowing that an MLSP is for load balancing, can decide on an optimum number of sub-LSPs, and place them appropriately across the network to optimize load balancing. On the other hand, an ingress LSR asked to create N independent LSPs will do so without regard to whether N is a good number of equal cost paths, and, more importantly, may place several of the N LSPs on the same path, defeating the purpose of load balancing.
3. The EB sub-LSP mechanism will, in many cases, result in far fewer sub-LSPs than independent LSPs and thus less control plane state.

4. Finally, an MLSP will usually have less data plane state than N independent LSPs: whenever multiple sub-LSPs traverse a link, a single label will be used for all of them, whereas if multiple LSPs traverse a link, each will need a separate label.

2.4. The Capabilities of TE-based Load Balancing

Definition: Let $G=(V, E)$ be a directed graph (or network), and let A and B in V be two nodes in G . Let T be the traffic arriving at A destined for B . T is said to be "IP-style" load balanced if for every node X on a shortest path from A to B , the portion of T arriving at X is split equally among all nodes Y_i that are adjacent to X and are on a shortest path from X to B .

Theorem: An MLSP can accurately mimic IP-style load balancing between any two nodes in any network.

Proof: left to the reader.

Corollary: MLSPs provide a strictly more powerful load balancing mechanism than IP-style load balancing.

3. Operation of MLSPs

3.1. Signaling MLSPs

Sub-LSPs of an MLSP are tied together using ASSOCIATION objects. ASSOCIATION objects have a new Association Type for MLSPs (TBD). The Association ID is chosen by the ingress of the MLSP; the Association Source is the loopback address of the ingress of the MLSP. All sub-LSPs containing an ASSOCIATION object with a given Association Source and Type belong to the same MLSP.

3.1.1. Indicating Equi-bandwidth (EB) nature

A sub-LSP is considered equi-bandwidth if its Path message carries the optional LSP_ATTRIBUTES object ([RFC5420]) with an EBC (equi-bandwidth capability) flag in the Attribute Flags TLV. The bit number for the EBC flag is yet to be assigned by IANA.

3.2. Label Allocation

A LSR S that receives Path messages for several sub-LSPs of the same MLSP from the same upstream LSR SHOULD allocate the same label for all the sub-LSPs. This simplifies load balancing for the aggregate traffic on those sub-LSPs. If the sub-LSPs are EB sub-LSPs, then S SHOULD allocate the same label for all EB sub-LSPs of the same MLSP that pass through S, regardless of which upstream LSR they come from. This allows S to load balance the aggregate traffic received on the MLSP, as all the MLSP traffic arrives at S with the same label. However, an LSR that can achieve the load balancing requirements independent of label allocation strategies is free to do so.

3.3. Bandwidth Accounting

Since MLSPs are traffic engineered, there needs to be strict bandwidth accounting, or admission control, on every link that an MLSP traverses. For non-EB sub-LSPs, this is straightforward, and analogous to regular TE LSPs. However, for EB sub-LSPs, two new procedures are needed, one for signaling bandwidth, and the other for admission control. First, for a given MLSP Z, an LSR X MUST ensure (via signaling) that the total incoming bandwidth of EB sub-LSPs of MLSP Z is divided equally among all the downstream links of X which at least one of the EB sub-LSPs traverses. Second, LSR X MUST ensure that, for each upstream link of X, there is sufficient bandwidth to accommodate all EB sub-LSPs of MLSP Z that traverse that link.

Let's take the example of Figure 2, with MLSP Z having five EB sub-LSPs E1 to E5, and say that MLSP Z is configured with a bandwidth of 30Gbps. Here are some of the steps involved.

1. LSR A, being the ingress, has no upstream links. A has two downstream links, AL and AM. Three EB sub-LSPs of MLSP Z traverse AL, and two traverse AM. A MUST signal a total of 15Gbps for the sub-LSPs to L, and a total of 15Gbps for the sub-LSPs to M. The required bandwidth may be divided up among the sub-LSPs to L (similarly, to M) in any manner so long as the total is 15Gbps. For example, A can signal sub-LSP E1 with 15Gbps, and sub-LSPs E3 and E5 with 0 bandwidth.
2. LSR L has one upstream link AL with three EB sub-LSPs with a total bandwidth of 15Gbps. L MUST ensure that 15Gbps is available for the AL link. If this bandwidth is not available, L MUST send a PathErr on ALL of the EB sub-LSPs on the AL link. Let's assume that the AL link has sufficient bandwidth.
3. Next, it is up to L to decide how to divide the incoming 15Gbps among the three downstream EB sub-LSPs to S. Say L signals sub-LSP E1 with 15Gbps, and the others with 0 bandwidth.
4. LSR S has two upstream links: LS with three EB sub-LSPs with a total bandwidth of 15Gbps, and MS with two EB sub-LSPs with a total bandwidth of 15Gbps. S MUST ensure that 15Gbps is available for each of the LS and MS links. S has thus a total incoming bandwidth of 30Gbps on MLSP Z. S has to divide this equally among its downstream links SP, SQ and SR, yielding 10Gbps each. S MUST ensure that the total bandwidth requested on the SP link for sub-LSPs E1 and E4 is 10Gbps. S may choose to signal these sub-LSPs with 5Gbps each. Similarly for the SQ and SR links.

There are two important points to note here. One is that the bandwidth reservation (TSpec) for a given EB sub-LSP can (and usually will) change hop-by-hop. The second is that as new EB sub-LSPs are signaled for an MLSP, the bandwidth reservations for existing EB sub-LSPs belonging to the same MLSP may have to be updated. To minimize these updates, it is RECOMMENDED that the first EB sub-LSP on a link be signaled with the total required bandwidth (as far as is known), and later sub-LSPs on the same link be signaled with 0 bandwidth.

3.4. MLSP Data Plane Actions

Traffic intended to be sent over an MLSP is determined at the ingress LSR by means outside the scope of this document, and at transit LSRs by the label(s) assigned by the transit LSR to its upstream LSRs. In the case of non-EB sub-LSPs, this traffic is load balanced across downstream links in the ratio of the bandwidths of the sub-LSPs that comprise the MLSP. In the case of EB sub-LSPs, the traffic belonging to an MLSP from an upstream LSR (or better still, the aggregate

traffic for the MLSP from all upstream LSRs) is load balanced equally among all downstream links.

As noted above, the overriding concern is that flows are mapped to the same downstream link (except when the MLSP or some constituent sub-LSPs are changing); this is typically done by hashing fields that define a flow, and mapping hash results to different downstream LSRs. Hash-based load balancing typically assumes that the numbers of flows is sufficiently large and the bandwidth per flow is reasonably well-balanced so that the results of hashing yields reasonable traffic distribution.

Entropy labels ([RFC6790] and [RFC6391]) can be used to improve load balancing at intermediate nodes.

4. Manageability

TBD

5. Security Considerations

This document introduces no new security concerns in the setup and signaling of LSPs using RSVP-TE, or in the use of the RSVP protocol. [RFC2205] specifies the message integrity mechanisms for RSVP signaling. These mechanisms apply to RSVP-TE signaling of MLSPs described in this document, and are highly recommended pending newer integrity mechanisms for RSVP.

6. Acknowledgments

The author would like to thank the Routing Protocol group at Juniper Networks for their questions, comments and encouragement for this proposal. While many participated, special thanks go to Yakov Rekhter, John Drake and Rahul Aggarwal. Many thanks too to John for suggesting the use of ASSOCIATION objects.

7. IANA Considerations

IANA is requested to assign the following:

A new Association Type for MLSP. This Association Type is to be used for ASSOCIATION objects with C-Type 1 (IPv4 Source) and 2 (IPv6 Source).

A new flag in the Attribute Flags TLV in the LSP_ATTRIBUTES object ([RFC5420]: a bit number for the EBC (equi-bandwidth capability) to indicate that a specific sub-LSP is an equi-bandwidth sub-LSP.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangar, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.

8.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4972] Vasseur, JP., Leroux, JL., Yasukawa, S., Previdi, S., Psenak, P., and P. Mabbey, "Routing Extensions for Discovery of Multiprotocol (MPLS) Label Switch Router (LSR) Traffic Engineering (TE) Mesh Membership", RFC 4972, July 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

Authors' Addresses

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti.kompella@gmail.com

Mike Hellers
LINX

Email: mikeh@linx.net

Ravi Singh
Juniper Networks

Email: ravis@juniper.net

MPLS Working Group
Internet Draft
Intended status: Informational
Expires: November 2011

R. Ram
D. Cohn
Orckit-Corrigent

M. Daikoku
KDDI

M. Yuxia
Y. Jian
ZTE Corp.

A. D'Alessandro
Telecom Italia

May 31, 2011

SD detection and protection triggering in MPLS-TP
draft-rkhd-mpls-tp-sd-03.txt

Abstract

This document describes guidelines for Signal Degrade (SD) fault condition detection at an arbitrary transport path (LSP or PW) and the usage of MPLS-TP fault management [3] for triggering protection switching as defined in the MPLS-TP survivability framework [2].

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire in November 2011.

Copyright Notice

Copyright (c) 2011 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.

Table of Contents

1. Introduction	3
2. Conventions used in this document	3
3. Signal Degrade and MPLS-TP protection switching	4
4. SD detection method	4
4.1. Guidelines for SD detection	4
4.2. Examples for SD detection methods	6
5. Transmission of link degradation fault indication	6
5.1. Lower layer Bit Error transmission	7
6. Handling of link degradation fault indication	7
7. Security Considerations	7
8. IANA Considerations	7
9. Acknowledgments	7
10. References	7
10.1. Normative References	7
10.2. Informative References	8

1. Introduction

Telecommunication carriers and network operators expect to replace aged TDM Services (e.g. legacy VPN services) provided by legacy TDM equipment by new VPN services provided by MPLS-TP equipment.

From a service level agreement (SLA) point of view, service quality and availability degradation are not acceptable, even after migration to MPLS-TP equipment.

In addition, from an operational point of view, comparable performance monitoring features to those provided by TDM networks are expected from MPLS-TP networks. For example, OAM maintenance points should be the same after TDM to MPLS-TP migration, as SLA revision is typically NOT feasible for telecommunication carriers and network operators.

MPLS-TP transport path (i.e. LSP,PW) resiliency actions such as protection switching can be triggered by fault conditions and external manual commands. Fault conditions include Signal Failure (SF) and Signal Degrade (SD). The SD condition could be detected at an intermediate link, based on lower layer indications or other sub-layer techniques.

Since the transport path protection switching is not necessarily managed by the transport entity that detects the SD condition, an indication of the link SD condition must be sent over the transport paths that traverse the affected link.

This document describes guidelines for SD detection by lower layers indication, and a mechanism for relaying the degraded transport path condition to the network element handling the protection switching at the appropriate transport path level.

2. Conventions used in this document

BER: Bit Error Rate

LSP: Label Switched Path

LSR: Label Switching Router

MEP: Maintenance End Point

MPLS: Multi-Protocol Label Switching

MPLS-TP: MPLS Transport Profile

OAM: Operations, Administration and Maintenance

OTN: Optical Transport Network

PCS: Physical Coding Sublayer

SF: Signal Failure

SD: Signal Degrade

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 [1].

3. Signal Degrade and MPLS-TP protection switching

Network survivability, as defined in [2], is the ability of a network to recover traffic delivery following failure or degradation of network resources. [5] defines an LSP protection mechanism and state machine that handles SF, SD and operator manual commands.

4. SD detection method

4.1. Guidelines for SD detection

Signal degrade is a transport path condition in which the expected quality of transport service delivery is not provided. The signal degrade condition can be used by operators to detect different types of failures, especially those with slow externalization such as optical device aging (e.g. photo detector and laser diode in line amplifier, transponder or SFP), transmission medium external impairment (e.g. temperature or pressure fluctuation, fiber elongation), and time-variable optical impairments in fiber (e.g. chromatic dispersion, polarization mode dispersion).

Signal degrade condition in a transport path is derived from bit error detection in the traversed links.

Bit errors in a link are caused by the following phenomena:

1. Physical conditions such as bad electrical connections, low received optical power, dispersion effects.

2. Non-physical conditions such as network congestion, CPU overload, selective packet discard, packet processing error.

The common basis for the guidelines set forth in this section is that the SD condition SHOULD reflect only physical error conditions in the traversed links, without any influence from non-physical conditions.

The following conditions SHOULD be met by the signal degrade condition detection mechanism:

- o Method for determining signal degrade MUST NOT affect the services transmitted over the transport path (e.g. add delay or jitter to real-time traffic)
- o Criterion for determining signal degrade MUST be agnostic to the length of transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to the transmission rate of transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to the type of service carried by the transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to the traffic class of transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to drop-precedence marking of transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to congestion
- o Criterion for determining signal degrade SHOULD be able to detect low error levels (e.g. BER of $10E-8$)
- o Criterion for determining signal degrade SHOULD have low misdetection probability
- o Criterion for determining signal degrade SHOULD have low false alarm probability
- o Criterion for determining signal degrade SHOULD be agnostic to number of transport paths (LSPs and PWs) transported over the transmission link
- o Signal degrade conditions MUST be monitored by the lowest server layer or sub-layer that is not terminated between monitoring points

- o Method for determining signal degrade SHOULD NOT require transmission of additional packets
- o Method for determining signal degrade SHOULD allow to localize links that contribute to signal degrade
- o Method for determining signal degrade MUST be able to exit signal degrade condition when error rate returns to normal condition
- o Method for determining signal degrade condition MUST be scalable

4.2. Examples for SD detection methods

- o A Server MEP [4] related to SONET or SDH sub-layers can determine SD condition based on error indication from parity information in the path overhead.
- o A Server MEP related to OTN sub-layer can determine SD condition based on error indications from Forward-Error-Correction functionality inherent in encapsulation.
- o A Server MEP related to 10GE PCS sub-layer can determine SD condition based on rate of errored 66-bit block headers. (a.k.a. symbol errors)
- o A Server MEP related to 1GE PCS sub-layer can determine SD condition based on rate of 10-bit code violations dispersion errors.

As specified in section 4.1, these examples assume that the layer carrying the information used for SD detection is not terminated by non-MPLS-TP-LSR entities (e.g. media converter).

5. Transmission of link degradation fault indication

When SD condition is detected, a link degradation fault indication [3] SHOULD be transmitted over affected transport paths, in the downstream direction from the detection point. The link degradation indication will be transmitted immediately following the detection and periodically until the SD condition is removed. The messages will be terminated and handled by the downstream client MEP.

The encapsulation and mechanism defined in [3] is suitable for transmission of link degradation fault indication. It is RECOMMENDED that [3] will include this definition in future work.

5.1. Lower layer Bit Error transmission

There are scenarios where the lower layer bit error rate in each of the links traversed by the transport path is below the SD threshold, while the accumulated end-to-end BER on the LSP is above the threshold. This is possible in lower layer technologies where errored information is dropped, so errors in one link will not be detected by LSRs downstream of this link. An example of such a situation is when an LSP is carried over multiple Ethernet links, and each link drops errored Ethernet frames.

To enable SD detection in such scenarios, LSRs MAY optionally include the measured BER in the link degradation fault indication message. The client MEP may then receive multiple link degradation fault indication messages from different LSRs. When this occurs, the client MEP SHOULD compare the sum of the received BER values with the SD threshold to decide on the LSP SD condition.

6. Handling of link degradation fault indication

LSR behavior upon receiving link degradation fault indication is out of the scope of this document.

SD condition processing and prioritization for protection triggering is out of the scope of this document.

SD clear condition processing and prioritization for protection triggering is out of the scope of this document.

7. Security Considerations

To be added in a future version of the document.

8. IANA Considerations

<N/A>

9. Acknowledgments

The editors gratefully acknowledge the contributions of Amir Halperin and Shachar Katz.

10. References

10.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

10.2. Informative References

- [2] Sprecher,N., and Farrel,A., "Multiprotocol Label Switching Transport Profile Survivability Framework", draft-ietf-mpls-tp-survive-fwk-06(work in progress), June 2010
- [3] Swallow,G., Fulignoli,A., Vigoureux,M., Boutros,S., and Ward,D., "MPLS Fault Management OAM", draft-ietf-mpls-tp-fault-04 (work in progress), April 2011
- [4] Busi,I. and Allan,D., "MPLS-TP OAM Framework", draft-ietf-mpls-tp-oam-framework-11 (work in progress), February 2011
- [5] Bryant,S., Osborne,E., Weingarten,Y., Sprecher,N., Fulignoli,A., "MPLS-TP Linear Protection", draft-ietf-mpls-tp-linear-protection-06 (work in progress), March 2011

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Rafi Ram
Orckit-Corrigent
126 Yigal Alon St.
Tel Aviv
Israel

Email: rafir@orckit.com

Daniel Cohn
Orckit-Corrigent
126 Yigal Alon St.
Tel Aviv
Israel

Email: danielc@orckit.com

Masahiro Daikoku
KDDI
3-10-10, Iidabashi, Chiyoda-ku,
Tokyo
Japan

Email: ms-daikoku@kddi.com

Ma Yuxia
ZTE Corp.
China

Email: ma.yuxia@zte.com.cn

Yang Jian
ZTE Corp.
China

Email: yang.jian90@zte.com.cn

Alessandro D'Alessandro
Telecom Italia
Italy

Email: alessandro.dalessandro@telecomitalia.it

Contributors

Amir Halperin

Shachar Katz

MPLS
Internet-Draft
Intended status: Informational
Expires: April 5, 2013

C. Villamizar, Ed.
Outer Cape Cod Network
Consulting
October 2, 2012

Use of Multipath with MPLS-TP and MPLS
draft-villamizar-mpls-tp-multipath-03

Abstract

Many MPLS implementations have supported multipath techniques and many MPLS deployments have used multipath techniques, particularly in very high bandwidth applications, such as provider IP/MPLS core networks. MPLS-TP has strongly discouraged the use of multipath techniques. Some degradation of MPLS-TP OAM performance cannot be avoided when operating over many types of multipath implementations.

Using MPLS Entropy label, MPLS can LSP can be carried over multipath links while also providing a fully MPLS-TP compliant server layer for MPLS-TP LSP. This document describes the means of supporting MPLS as a server layer for MPLS-TP. The use of MPLS-TP LSP as a server layer for MPLS LSP is also discussed.

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 5, 2013.

Copyright Notice

Copyright (c) 2012 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. Definitions	3
3. MPLS as a Server Layer for MPLS-TP	5
4. MPLS-TP as a Server Layer for MPLS	7
5. IANA Considerations	8
6. Security Considerations	8
7. References	8
7.1. Normative References	8
7.2. Informative References	8
Author's Address	9

1. Introduction

Today the requirement to handle large aggregations of traffic, can be handled by a number of techniques which we will collectively call multipath. Multipath applied to parallel links between the same set of nodes includes Ethernet Link Aggregation [IEEE-802.1AX], link bundling [RFC4201], or other aggregation techniques some of which may be vendor specific. Multipath applied to diverse paths rather than parallel links includes Equal Cost MultiPath (ECMP) as applied to OSPF, ISIS, or BGP, and equal cost LSP. Some vendors support load split across equal cost MPLS LSP where the load is split proportionally to the reserved bandwidth of the set of LSP.

RFC 5654 requirement 33 requires the capability to carry a client MPLS-TP or MPLS layer over a server MPLS-TP or MPLS layer [RFC5654]. This is possible in all cases with one exception. When an MPLS LSP exceeds the capacity of any single component link it may be carried by a network using multipath techniques, but may not be carried by an MPLS-TP LSP due to the inherent MPLS-TP capacity limitation imposed by MPLS-TP OAM packet ordering constraints.

The term composite link is more general than terms such as link aggregation (which is specific to Ethernet) or ECMP (which implies equal cost paths within a routing protocol). The use of the term composite link here is consistent with the broad definition in [ITU-T.G.800]. Multipath is very similar to composite link as defined by ITU, but specifically excludes inverse multiplexing.

2. Definitions

Multipath

The term multipath includes all techniques in which

1. Traffic can take more than one path from one node to a destination.
2. Individual packets take one path only. Packets are not subdivided and reassembled at the receiving end.
3. Packets are not resequenced at the receiving end.
4. The paths may be:
 - a. parallel links between two nodes, or
 - b. may be specific paths across a network to a destination node, or

- c. may be links or paths to an intermediate node used to reach a common destination.

Link Bundle

Link bundling is a multipath technique specific to MPLS [RFC4201]. Link bundling supports two modes of operations. Either an LSP can be placed on one component link of a link bundle, or an LSP can be load split across all members of the bundle. There is no signaling defined which allows a per LSP preference regarding load split, therefore whether to load split is generally configured per bundle and applied to all LSP across the bundle.

Link Aggregation

The term "link aggregation" generally refers to Ethernet Link Aggregation [IEEE-802.1AX] as defined by the IEEE. Ethernet Link Aggregation defines a Link Aggregation Control Protocol (LACP) which coordinates inclusion of LAG members in the LAG.

Link Aggregation Group (LAG)

A group of physical Ethernet interfaces that are treated as a logical link when using Ethernet Link Aggregation is referred to as a Link Aggregation Group (LAG).

Equal Cost Multipath (ECMP)

Equal Cost Multipath (ECMP) is a specific form of multipath in which the costs of the links or paths must be equal in a given routing protocol. The load may be split equally across all available links (or available paths), or the load may be split proportionally to the capacity of each link (or path).

Loop Free Alternate Paths

"Loop-free alternate paths" (LFA) are defined in RFC 5714, Section 5.2 [RFC5714] as follows. "Such a path exists when a direct neighbor of the router adjacent to the failure has a path to the destination that can be guaranteed not to traverse the failure." Further detail can be found in [RFC5286]. LFA as defined for IPFRR can be used to load balance by relaxing the equal cost criteria of ECMP, though IPFRR defined LFA for use in selecting protection paths. When used with IP, proportional split is generally not used. LFA use in load balancing is implemented by some vendors though it may be rare or non-existent in deployments.

Composite Link

The term Composite Link had been a registered trademark of Avici Systems, but was abandoned in 2007. The term composite link is now defined by the ITU in [ITU-T.G.800]. The ITU definition

includes multipath as defined here, plus inverse multiplexing which is explicitly excluded from the definition of multipath.

Inverse Multiplexing

Inverse multiplexing either transmits whole packets and resequences the packets at the receiving end or subdivides packets and reassembles the packets at the receiving end. Inverse multiplexing requires that all packets be handled by a common egress packet processing element and is therefore not useful for very high bandwidth applications.

Component Link

The ITU definition of composite link in [ITU-T.G.800] and the IETF definition of link bundling in [RFC4201] both refer to an individual link in the composite link or link bundle as a component link. The term component link is applicable to all multipath.

LAG Member

Ethernet Link Aggregation as defined in [IEEE-802.1AX] refers to an individual link in a LAG as a LAG member. A LAG member is a component link. An Ethernet LAG is a composite link. IEEE does not use the terms composite link or component link.

load split

Load split, load balance, or load distribution refers to subdividing traffic over a set of component links such that load is fairly evenly distributed over the set of component links and certain packet ordering requirements are met. Some existing techniques better achieve these objectives than others.

A small set of requirements are discussed. These requirements make use of keywords such as MUST and SHOULD as described in [RFC2119].

3. MPLS as a Server Layer for MPLS-TP

MPLS LSP may be used as a server layer for MPLS-TP LSP as long as all MPLS-TP requirements are met, including the requirement that packets within an MPLS-TP LSP are not reordered, including both payload and OAM packets.

Supporting MPLS-TP LSP over a fully MPLS-TP conformant MPLS LSP server layer where the MPLS LSP are making use of multipath, requires special treatment of the MPLS-TP LSP such that those LSP only are not subject to the multipath load splitting. This implies the following brief set of requirements.

- MP#1 It MUST be possible to identify MPLS-TP LSP.
- MP#2 It MUST be possible to completely exclude MPLS-TP LSP from the multipath hash and load split.
- MP#3 It SHOULD be possible to insure that an MPLS-TP LSP will not be moved to another component link as a result of a composite link load rebalancing operation.
- MP#4 Where an RSVP-TE control plane is used, it MUST be possible for an ingress LSR which is setting up an MPLS-TP or MPLS LSP to determine at CSPF time whether a link or MPLS PSC LSP within the topology can support the MPLS-TP requirements of the LSP.

There is currently no signaling mechanism defined to support requirement MP#1. In the absence of a signaling extension, MPLS-TP can be identified through some form of configuration, such as configuration which provides an MPLS-TP compatible server layer to all LSP arriving on a specific interface or originating from a specific set of ingress LSR. Alternately an MPLS-TP LSP can be created with an Entropy Label Indicator (ELI) and entropy label (EL) below the MPLS-TP label [I-D.ietf-mpls-entropy-label].

Some hardware which exists today can support requirement MP#2. Signaling in the absence of MPLS Entropy Label can make use of link bundling with a specific component for MPLS-TP LSP and link bundling with the all-zeros component for MPLS LSP. This prevents MPLS-TP LSP from being carried within MPLS LSP but does allow the co-existence of MPLS-TP and very large MPLS LSP.

MPLS-TP LSP can be carried as client LSP within an MPLS server LSP if an Entropy Label Indicator (ELI) and entropy label (EL) is added after the server layer LSP label(s) in the label stack, just above the MPLS-TP LSP label entry [I-D.ietf-mpls-entropy-label]. This allows MPLS-TP LSP to be carried as client LSP within MPLS LSP and satisfies requirement MP#2 but requires that MPLS LSR be able to identify MPLS-TP LSP (requirement MP#1).

MPLS-TP traffic can be protected from a degraded performance due to an imperfect load split if the MPLS-TP traffic is given queuing priority (using strict priority and policing or shaping at ingress or locally or weighted queuing locally). This can be accomplished using the Traffic Class field and Diffserv treatment of traffic [RFC5462][RFC2475]. In the event of congestion due to load imbalance, other traffic will suffer as long as there is a minority of MPLS-TP traffic.

If MPLS-TP LSP are carried within MPLS LSP and ELI and EL are used,

requirement MP#2 is satisfied, but without a signaling extension, requirement MP#3 is not satisfied if there is a need to rebalance the load on any composite link carrying the MPLS server LSP. Load rebalance is generally needed only when congestion occurs, therefore restricting MPLS-TP to be carried only over MPLS LSP that are known to traverse only links which are expected to be uncongested can satisfy requirement MP#3.

Requirement MP#4 can be supported using administrative attributes. Administrative attributes are defined in [RFC3209]. Some configuration is required to support this.

4. MPLS-TP as a Server Layer for MPLS

Carrying MPLS LSP which are larger than a component link over a MPLS-TP server layer requires that the large MPLS client layer LSP be accommodated by multiple MPLS-TP server layer LSPs. MPLS multipath can be used in the client layer MPLS.

Creating multiple MPLS-TP server layer LSP places a greater ILM scaling burden on the LSR. High bandwidth MPLS cores with a smaller amount of nodes have the greatest tendency to require LSP in excess of component links, therefore the reduction in number of nodes offsets the impact of increasing the number of server layer LSP in parallel. Today, only in cases where deployed LSR ILM are small would this be an issue.

The most significant disadvantage of MPLS-TP as a Server Layer for MPLS is that the use MPLS-TP server layer LSP reduces the efficiency of carrying the MPLS client layer. The service which provides by far the largest offered load in provider networks is Internet, for which the LSP capacity reservations are predictions of expected load. Many of these MPLS LSP may be smaller than component link capacity. Using MPLS-TP as a server layer results in bin packing problems for these smaller LSP. For those LSP that are larger than component link capacity, their capacity are not increments of convenient capacity increments such as 10Gb/s. Using MPLS-TP as an underlying server layer greatly reduces the ability of the client layer MPLS LSP to share capacity. For example, when one MPLS LSP is underutilizing its predicted capacity, the fixed allocation of MPLS-TP to component links may not allow another LSP to exceed its predicted capacity. Using MPLS-TP as a server layer may result in less efficient use of resources may result in a less cost effective network.

No additional requirements beyond MPLS-TP as it is now currently defined are required to support MPLS-TP as a Server Layer for MPLS. It is therefore viable but has some undesirable characteristics

discussed above.

5. IANA Considerations

This memo includes no request to IANA.

6. Security Considerations

This document specifies requirements with discussion of framework for solutions using existing MPLS and MPLS-TP mechanisms. The requirements and framework are related to the coexistence of MPLS/GMPLS (without MPLS-TP) when used over a packet network, MPLS-TP, and multipath. The combination of MPLS, MPLS-TP, and multipath does not introduce any new security threats. The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [I-D.ietf-mpls-tp-security-framework].

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

7.2. Informative References

[I-D.ietf-mpls-entropy-label]
Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-ietf-mpls-entropy-label-06 (work in progress), September 2012.

[I-D.ietf-mpls-tp-security-framework]
Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS-TP Security Framework", draft-ietf-mpls-tp-security-framework-04 (work in progress), July 2012.

[IEEE-802.1AX]
IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 2006, <<http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf>>.

[ITU-T.G.800]

ITU-T, "Unified functional architecture of transport networks", 2007, <<http://www.itu.int/rec/T-REC-G/recommendation.asp?parent=T-REC-G.800>>.

- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

Author's Address

Curtis Villamizar (editor)
Outer Cape Cod Network Consulting
Email: curtis@ocnc.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 18, 2011

M. Xiao, Ed.
L. Jin
B. Wu
J. Yang
ZTE Corporation
October 15, 2010

Throughput Estimation for MPLS based Transport Networks
draft-xiao-mpls-tp-throughput-estimation-01

Abstract

An important Operation, Administration and Maintenance requirement of the MPLS Transport Profile (MPLS-TP) is the ability to estimate the throughput (i.e. bandwidth) for an MPLS-TP connection which could be an MPLS-TP PW, LSP or Section. This document specifies OAM packets and protocol mechanisms to facilitate the efficient and precise measurement of throughput.

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 18, 2011.

Copyright Notice

Copyright (c) 2010 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. Conventions	3
1.2. Abbreviations	3
2. Overview	4
2.1. Two-way Throughput Measurement	4
2.2. One-way Throughput Measurement	5
2.3. Unidirectional Connections	5
3. Packet Format	5
3.1. Throughput Measurement Indication Packet Format	5
3.2. Throughput Measurement Test Packet Format	10
4. Throughput Measurement Procedures	10
4.1. Transmitting a Throughput Measurement Start Request	10
4.2. Receiving a Throughput Measurement Start Request	11
4.3. Transmitting a Throughput Measurement Start Reply	11
4.4. Receiving a Throughput Measurement Start Reply	11
4.5. Sending and Receiving Test Traffic	12
4.6. Transmitting a Throughput Measurement Stop Request	12
4.7. Receiving a Throughput Measurement Stop Request	12
4.8. Transmitting a Throughput Measurement Stop Reply	13
4.9. Receiving a Throughput Measurement Stop Reply	13
4.10. Consequent Actions and Searching Algorithm	13
5. Throughput Measurement Time	15
6. Open Issue	15
7. IANA Considerations	15
8. Security Considerations	15
9. Acknowledgements	16
10. References	16
10.1. Normative References	16
10.2. Informative References	16
Authors' Addresses	17

1. Introduction

As defined in [RFC5860], the MPLS-TP OAM toolset MUST provide a function to enable conducting diagnostic tests on a PW, LSP or Section, this function SHOULD be performed on-demand and one example of such diagnostic test consists in estimating the bandwidth of e.g., an LSP.

To make this requirement clearer and provide more details, this sub-function of diagnostic tests is specified as "throughput estimation" in [I-D.ietf-mpls-tp-oam-framework], throughput estimation is an on-demand out-of-service function, that allows verifying the bandwidth/throughput of an MPLS-TP transport path (LSP or PW) before it is put in-service. Throughput estimation is performed between MEPs and can be performed in one-way or two-way mode.

This document specifies the OAM packets and procedures for both one-way and two-way throughput estimation/measurement.

1.1. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

1.2. Abbreviations

CRC: Cyclic Redundancy Check

G-ACh: Generic Associated Channel

DUT: Device Under Test

LSP: Label Switched Path

MEG: Maintenance Entity Group

MEP: Maintenance Entity Group End Point

MPLS-TP: MPLS Transport Profile

NMS: Network Management System

OAM: Operations, Administration and Maintenance

PHB: Per-hop Behavior

PRBS: Pseudo-Random Bit Sequence

PW: PseudoWire

TLV: Type Length Value

2. Overview

In [RFC1242], the throughput is specified as a performance metric for network interconnection device, and it's defined by "the maximum rate at which none of the offered frames are dropped by the device". In MPLS-TP context the concept of throughput is not just for a particular device, but extended to apply to an MPLS-TP connection which could be a PW, LSP or Section.

In [RFC2544], corresponding to [RFC1242], the throughput measurement procedures are specified as "send a specific number of frames at a specific rate through the DUT and then count the frames that are transmitted by the DUT. If the count of offered frames is equal to the count of received frames, the fewer frames are received than were transmitted, the rate of the offered stream is reduced and the test is rerun. The throughput is the fastest rate at which the count of test frames transmitted by the DUT is equal to the number of test frames sent to it by the test equipment". But in current practical throughput measurement scenario, usually the throughput is measured by test equipment using the more efficient and precise binary search algorithm.

It should also be noted that for different test packet size, or test packet pattern, or test packet PHB, or expected measurement resolution, or even sending duration of test traffic, different result of throughput measurement may be obtained, so all these parameters need to be configurable for throughput measurement.

2.1. Two-way Throughput Measurement

For a bidirectional MPLS-TP connection, two-way throughput measurement needs to be supported. Two-way throughput should include both the throughput for the forward direction of the connection and the throughput for the reverse direction of the connection. In order to simplify the implementation and facilitate the results collection, all computational overhead and procedures control will be taken by the initiator MEP of throughput measurement, and the peer MEP will act just as a responder. Also note that both the initiator MEP and the peer MEP need to send test traffic for two-way throughput measurement.

It is worth noting that there is another optional definition of two-way throughput estimation, in which only the initiator MEP needs to

send test traffic and the peer MEP will loop back all received test packets. But note that in this case only the minimum of available throughput of the two directions can be achieved, so this optional definition of two-way throughput estimation is not recommended in this draft.

2.2. One-way Throughput Measurement

For a bidirectional MPLS-TP connection, one-way throughput measurement also needs to be supported. One-way throughput only indicates the throughput for the forward direction of the connection. Similar to two-way throughput measurement, the initiator MEP controls the whole process of one-way throughput measurement and the peer MEP will act just as a responder. Also note that only the initiator MEP needs to send test traffic for one-way throughput measurement.

2.3. Unidirectional Connections

For a unidirectional MPLS-TP connection (such as a unidirectional LSP), only one-way throughput measurement needs to be supported. If it's a unidirectional connection with return path, the procedures of one-way throughput measurement for bidirectional connection still apply. Else if it's a unidirectional connection without return path, the procedures of one-way throughput measurement are not as automatic as that for bidirectional connection, and manual provision of test parameters is needed for every run of sending test traffic. Besides, in this case the peer MEP instead of the initiator MEP will act as calculator for the packet loss of every run.

3. Packet Format

For throughput measurement the specific packets sent by the MEP can be divided into indication packets and test packets. The throughput measurement indication packets flow over the Generic Associated Channel Channel (G-ACh) [RFC5586] of an MPLS-TP connection and perform signaling between the initiator MEP and the peer MEP, while the throughput measurement test packets compose the test traffic which intends to emulate the real user traffic.

3.1. Throughput Measurement Indication Packet Format

The format of a throughput measurement indication packet is shown below.

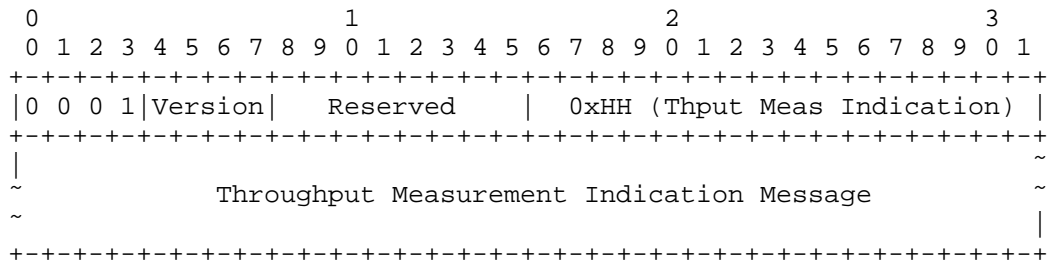


Figure 1: Throughput Measurement Indication Packet

The Version and Reserved field are always set to 0.

The Thput Meas Indication Channel Type is 0xHH (to be assigned by IANA).

The format of a throughput measurement indication message is shown below.

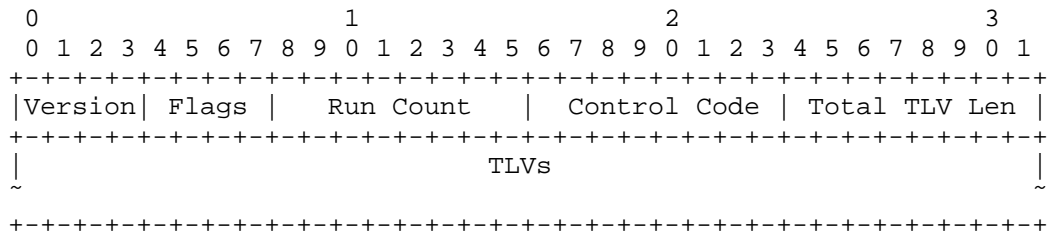


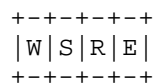
Figure 2: Throughput Measurement Indication Message

Version

The Version Number is currently set to 0.

Flags

Each bit indicates a message control flag. Three flags are defined and listed from left to right as follow:



W-flag: This Flag represents the operational mode which could be One-way mode or Two-way mode. Set to 0 for a One-way throughput measurement; Set to 1 for a Two-way throughput measurement.

S-flag: This Flag represents the message type which could be Start type or Stop type. Set to 0 for a Start message; Set to 1 for a Stop message.

R-flag: This Flag represents the message direction which could be Forward direction (i.e. Request) or Reverse direction (i.e. Reply). Set to 0 for a Request message; Set to 1 for a Reply message.

E bit (the fourth bit): Reserved for future use and set to 0.

Run Count

The Run Count is set to the number of all run times in one throughput measurement process and it starts from 1.

Control Code

According to the value of R-flag, the Control Code is set as follow.

For a Request:

0x0: Request (in-band reply requested). Indicates that this request has been sent over a bidirectional connection and the reply is expected over the same connection.

0x1: Request (out-of-band reply requested). Indicates that the reply is expected over an out-of-band path.

0x2: Request (no reply requested). Indicates that no reply is expected.

For a Reply:

0x0: Success. Indicates that the operation succeeded.

0x1: Error. Indicates that the operation failed.

Total TLV Length

The total TLV length is the total of all included TLVs.

TLVs

According to the values of W-flag, S-flag and R-flag, the TLVs are defined as follow.

For Start Request/Reply message in One-way throughput measurement:

No TLVs are defined at this time.

For Start Request/Reply message in Two-way throughput measurement:

One TLV is defined as follow.

0										1										2										3									
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	2	3	4	5	6	7	8	9
Type = 0										Length = 10																													
										Sending Rate																													
Sending Duration										Packet Size																													
Packet Pattern										PHB										Reserved																			

All the values in this TLV are test parameters for the peer MEP to send test traffic.

Sending Rate

The Sending Rate in Mbps is set to the provisioned initial sending rate of test traffic for the first run, and set to the calculated sending rate of test traffic for the rerun.

Sending Duration

The Sending Duration in seconds is set to the provisioned sending interval of test traffic for every run.

Packet Size

The Packet Size in octets is set to the provisioned throughput measurement test packet size.

Packet Pattern

The Packet Pattern is set to the provisioned throughput measurement test packet pattern. According to [ITU-T Y.1731], four pattern types of throughput measurement test packets pattern types are defined as below:

0x00: Null (all-zeros) signal without CRC-32

0x01: Null (all-zeros) signal with CRC-32

0x02: PRBS ($2^{31}-1$) without CRC-32

0x03: PRBS ($2^{31}-1$) with CRC-32

0x04~0xFF: Reserved for future standardization

PHB

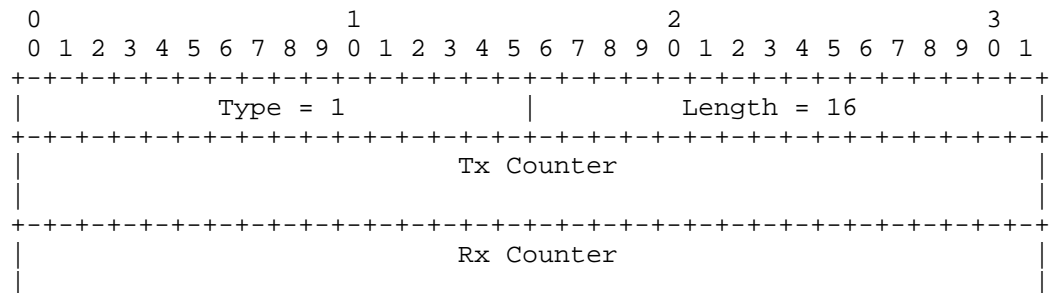
The PHB is set to the provisioned throughput measurement test packet PHB.

Reserved

Reserved bits for future use and always set to 0.

For Stop Request/Reply message in One-way/Two-way throughput measurement:

One TLV is defined as follow.



trigger the peer MEP to start counting test packets. Specifically if one-way throughput measurement is performed on a unidirectional MPLS-TP connection without return path, the initiator MEP also should start sending test traffic a while (such as 1 second) after transmitting the Start Request. For two-way throughput measurement, except for the same intention as one-way throughput measurement, this message is also intended to convey necessary test parameters to the peer MEP and trigger the test traffic sending at the peer MEP, and note that for rerun only Run Count and Sending Rate in the message need to be changed while other parameters retain initial values. Furthermore, for both one-way and two-way measurement, the initiator MEP should start counting test packets as soon as it transmits this message.

Specifically if the connection is unidirectional, then the Control Code in the message must not be set to 0x0 (in-band reply requested), moreover if no return path exists the Control Code in the message must be set to 0x2 (no reply requested).

4.2. Receiving a Throughput Measurement Start Request

Upon the reception of a throughput measurement Start Request, the peer MEP must inspect this message at first, if no unexpected field or value is found then the peer MEP should start counting test packets. In addition, if the received W-flag indicates that this is a two-way throughput measurement, then the peer MEP also should start sending test traffic.

Specifically if the received W-flag indicates that this is a one-way throughput measurement, and the received Control Code is set to 0x2 (no reply requested) which means the connection is unidirectional without return path, then the peer MEP won't transmit Start Reply.

4.3. Transmitting a Throughput Measurement Start Reply

When the Control Code in a received Start Request is set to 0x0 (in-band reply requested) or 0x1 (out-of-band reply requested), the peer MEP must transmit a throughput measurement Start Reply to the initiator MEP. The Control Code in Start Reply Message should be set to 0x0 to reflect the successful operation at the peer MEP, or on the contrary set to 0x1 to reflect the failed operation at the peer MEP. Except the R-flag and Control Code field, other fields of Start Reply Message will be copied from the received Start Request Message.

4.4. Receiving a Throughput Measurement Start Reply

Upon the reception of a throughput measurement Start Reply, the initiator MEP must inspect this message at first, if no unexpected

field or value is found, and the received Control Code indicates successful operation at the peer MEP, then the initiator MEP should start sending test traffic. If there is no any throughput measurement Start Reply received after a while (such as 1 second), then specific error should be returned at the initiator MEP and no test traffic will be sent from the initiator MEP.

4.5. Sending and Receiving Test Traffic

From above procedures it can be seen that for two-way throughput measurement the pair of MEPs will send test traffic asynchronously, and the peer MEP will start/stop sending test traffic some earlier than the initiator MEP, but the asynchronism has no side-effect on the measurement result because both MEPs shall start counting test packets before they receive any test traffic.

Also note that when the initiator MEP sends test traffic the test parameters are all derived from the provisioned test parameters for the first run, and for rerun only the sending rate is changed and derived from the local calculation. When the peer MEP sends test traffic, the test parameters are all derived from the received Start Request Message.

4.6. Transmitting a Throughput Measurement Stop Request

For every run, after the initiator MEP finished sending test traffic, it will transmit a throughput measurement Stop Request to the peer MEP. This message is intended to inform the peer MEP about the stop of test traffic sending, and also trigger the peer MEP to stop counting test packets and feed back the counters.

Specifically if the connection is unidirectional, then the Control Code in the message must not be set to 0x0 (in-band reply requested), moreover if no return path exists the Control Code in the message must be set to 0x2 (no reply requested).

4.7. Receiving a Throughput Measurement Stop Request

Upon the reception of a throughput measurement Stop Request, the peer MEP must inspect this message at first, if no unexpected field or value is found then the peer MEP should stop counting test packets.

Specifically if the received W-flag indicates that this is a one-way throughput measurement, and the received Control Code is set to 0x2 (no reply requested) which means the connection is unidirectional without return path, then the peer MEP won't transmit Stop Reply and it will use the received Tx Counter to calculate test packet loss directly.

4.8. Transmitting a Throughput Measurement Stop Reply

When the Control Code in a received Stop Request is set to 0x0 (in-band reply requested) or 0x1 (out-of-band reply requested), the peer MEP must transmit a throughput measurement Stop Reply to the initiator MEP. The Control Code in Stop Reply Message should be set to 0x0 to reflect the successful operation at the peer MEP, or on the contrary set to 0x1 to reflect the failed operation at the peer MEP. Furthermore, the Stop Reply is transmitted also to confirm that the peer MEP has stopped sending test traffic for this run. The Tx Counter and Rx Counter are set to the test packet counting values at the peer MEP.

4.9. Receiving a Throughput Measurement Stop Reply

Upon the reception of a throughput measurement Stop Reply, the initiator MEP must inspect this message at first, if no unexpected field or value is found, and the received Control Code indicates successful operation at the peer MEP, then the initiator MEP should stop counting test packets and start calculating the test packet loss. Suppose the Tx Counter and Rx Counter for the initiator MEP are TxP1 and RxP1, and for the peer MEP are TxP2 and RxP2.

For two-way throughput measurement, the calculation formulas are as follow:

$$\text{Packet Loss (forward)} = \text{TxP1} - \text{RxP2}$$
$$\text{Packet Loss (reverse)} = \text{TxP2} - \text{RxP1}$$

For one-way throughput measurement, the calculation formula is as follow:

$$\text{Packet Loss (one-way)} = \text{TxP1} - \text{RxP2}$$

If there is no any throughput measurement Stop Reply received after a while (such as 1 second), then specific error should be returned at the initiator MEP and no consequent action will happen.

4.10. Consequent Actions and Searching Algorithm

Procedures for one run of test traffic sending and test packet loss calculation have been described above in details, but usually iterative reruns of the procedures are needed for a throughput measurement. Whether the rerun is needed or not is based on the calculated test packet loss and whether the expected measurement resolution is met. For one-way throughput measurement, if calculated Packet Loss (one-way) is equal to zero and the expected measurement

resolution is met, then rerun is not needed (i.e. the one-way throughput measurement finished) and the current sending rate is the measured one-way throughput, otherwise the one-way throughput measurement proceeds. For two-way throughput measurement, if calculated forward Packet Loss and reverse Packet Loss are both equal to zero and the expected measurement resolution for both forward and reverse directions is met, then rerun is not needed (i.e. the two-way throughput measurement finished) and the current sending rate for forward/reverse direction is the measured forward/reverse throughput, otherwise the two-way throughput measurement proceeds, and in this case the sending rates for rerun should be calculated for forward direction and reverse direction respectively.

The simple and efficient binary search algorithm is RECOMMENDED to calculate the sending rate for the next run, which is the only changed test parameter compared with this run. How the binary search works, if packet loss is found for this run, it searches downwards for a lower rate which is halfway rate between the rate of this run and the known highest rate at which no packet loss is found; if no packet loss is found but the expected measurement resolution is not met for this run, it searches upwards for a higher rate which is halfway rate between the rate of this run and the known lowest rate at which packet loss is found; the measurement searches among higher and lower rates on the analogy of this, until it finds the rate at which no test packet is lost and expected measurement resolution is met, and this rate is the measured throughput. How to judge whether the expected measurement resolution is met or not, if the rate difference between the two consecutive runs (i.e. this run and the previous run), expressed as a percentage, is smaller than or equal to the specified measurement resolution, it's known as that the expected measurement resolution is met, otherwise it's not met.

For example, suppose to measure the throughput of a connection whose actual throughput is 70Mbps, the provisioned initial sending rate is 100Mbps and the specified measurement resolution is 0.1. Note that the initial sending rate should be higher than the actual throughput, otherwise the binary search is not applicable, and so it's often set to the maximum theoretical throughput of the measured connection. For the first run, packet loss is found, so for the second run, the sending rate will be calculated as $(100+0)/2 = 50\text{Mbps}$, no packet loss is found, then the resolution will be calculated as $(100-50)/50 = 1$, which is bigger than 0.1, the expected measurement resolution is not met, so for the third run, the sending rate will be calculated as $(100+50)/2 = 75\text{Mbps}$, packet loss is found, so for the fourth run, the sending rate will be calculated as $(50+75)/2 = 62.5\text{Mbps}$, no packet loss is found, then the resolution will be calculated as $(75-62.5)/62.5 = 0.2$, which is bigger than 0.1, the expected measurement resolution is not met, so for the fifth run, the sending rate will be

calculated as $(75+62.5)/2 = 68.75\text{Mbps}$, no packet loss is found, then the resolution will be calculated as $(68.75-62.5)/68.75 = 0.09$, which is smaller than 0.1, the expected measurement resolution is met, so the measurement finished and the rate 68.75Mbps is the measured throughput.

Other algorithms than the binary search algorithm could also be used to search throughput in practice, e.g. increasing or decreasing the sending rate in a fixed step from a specified initial sending rate until the test packet loss appears or disappears.

5. Throughput Measurement Time

The throughput measurement time is about the product of sending duration for one run and number of all run times. The sending duration for one run is provisioned before the throughput measurement starts, and the number of all run times is related to several factors, which include the provisioned initial sending rate, the applied searching algorithm and the specified expected measurement resolution. It's obvious that longer sending duration is provisioned, then longer throughput measurement time is needed, but it should be noted that longer sending duration can result in more precise measured throughput, so there should be a balance between them. Also obviously the expectations for shorter throughput measurement time and higher throughput measurement resolution are mutually exclusive, so the balance between them is needed too.

6. Open Issue

Wouldn't it be better to have a threshold on the acceptable frame loss rate and not require absolutely no packet loss?

[Editor's note: As the authors know in practice when the throughput is measured by test devices, one threshold on the acceptable frame loss rate is configurable, but in [RFC1242] and [RFC2544] the throughput is defined as that way no packet loss permitted.]

7. IANA Considerations

To be added in a later version of this document.

8. Security Considerations

To be added in a later version of this document.

9. Acknowledgements

The authors would like to thank Huub (Huawei), Curtis (Infinera) and Ayal (celtro) for their valuable comments on this draft.

10. References

10.1. Normative References

- [I-D.ietf-mpls-tp-li-lb]
Boutros, S., Sivabalan, S., Swallow, G., Bryant, S., and C. Pignataro, "MPLS Transport Profile Lock Instruct and Loopback Functions", draft-ietf-mpls-tp-li-lb-00 (work in progress), September 2010.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.

10.2. Informative References

- [I-D.ietf-mpls-tp-oam-framework]
Allan, D., Busi, I., Niven-Jenkins, B., Fulignoli, A., Hernandez-Valencia, E., Levrau, L., Sestito, V., Sprecher, N., Helvoort, H., Vigoureux, M., Weingarten, Y., and R. Winter, "Operations, Administration and Maintenance Framework for MPLS- based Transport Networks", draft-ietf-mpls-tp-oam-framework-09 (work in progress), October 2010.
- [ITU-T Y.1731]
International Telecommunications Union - Telecommunication Standardization, "OAM functions and mechanisms for Ethernet based networks", ITU-T Y.1731, February 2008.
- [RFC1242] Bradner, S., "Benchmarking terminology for network interconnection devices", RFC 1242, July 1991.
- [RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, March 1999.

Authors' Addresses

Min Xiao (editor)
ZTE Corporation

Email: xiao.min2@zte.com.cn

LiZhong Jin
ZTE Corporation

Email: lizhong.jin@zte.com.cn

Bo Wu
ZTE Corporation

Email: wu.bo@zte.com.cn

Jian Yang
ZTE Corporation

Email: yang_jian@zte.com.cn

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 28, 2011

F. Zhang, Ed.
L. Jin
B. Wu
ZTE Corporation
October 25, 2010

The Analysis of MPLS-TP Path Segment Monitoring
draft-zhang-mpls-tp-path-segment-monitoring-01

Abstract

This specification analyzes the different schemes to realize path segment monitoring in MPLS-TP network.

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 28, 2011.

Copyright Notice

Copyright (c) 2010 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. Conventions used in this document	3
3. Path Segment Monitoring Analysis	3
3.1. MBB	3
3.2. Local Rerouting	4
3.3. TTL TLV	5
3.3.1. The scaling Analysis	6
4. IANA Considerations	6
5. Security Considerations	6
6. Acknowledgement	6
7. Normative references	7
Authors' Addresses	7

1. Introduction

In order to monitor, protect and manage a portion (i.e. segment or concatenated segment) of a transport path, a path segment is defined between the edges of the portion of the LSP that needs to be monitored, protected or managed. If this path segment is created as a hierarchical LSP, it is called SPME (Sub-Path Maintenance Element).

SPMEs are usually instantiated when the transport path is created by either the management plane or control plane for proactive monitoring. However, pre-design and pre-configuration of all the considered patterns of SPME are not sometimes preferable in real operation due to the burden of design works, a number of header consumptions, bandwidth consumption and so on, as described in section 3.8 of [I-D.ietf-mpls-tp-oam-framework].

There are different schemes to configure SPMEs after the transport path has been created, and two network objectives SHOULD be met:

1. The monitoring and maintenance of existing transport paths has to be conducted in service without traffic disruption.
2. The monitored or managed transport path condition has to be exactly the same irrespective of any configurations necessary for maintenance.

Here we will discuss the the advantages and disadvantages of different potential schemes.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119.

3. Path Segment Monitoring Analysis

3.1. MBB

The make-before-break (MBB) procedures which are supported by MPLS allow the creation of a SPME on existing LSPs in-service without traffic disruption, as described in [RFC5921]. An SPME can be defined corresponding to one or more end-to-end LSPs at first, then new end-to-end LSPs that are tunneled within the SPME can be set up, which may require coordination across administrative boundaries, finally traffic of the existing LSPs is switched over to the new end-

to-end tunneled LSPs. The old end-to-end LSPs can then be torn down. See the figure below, copied from [RFC5921]:

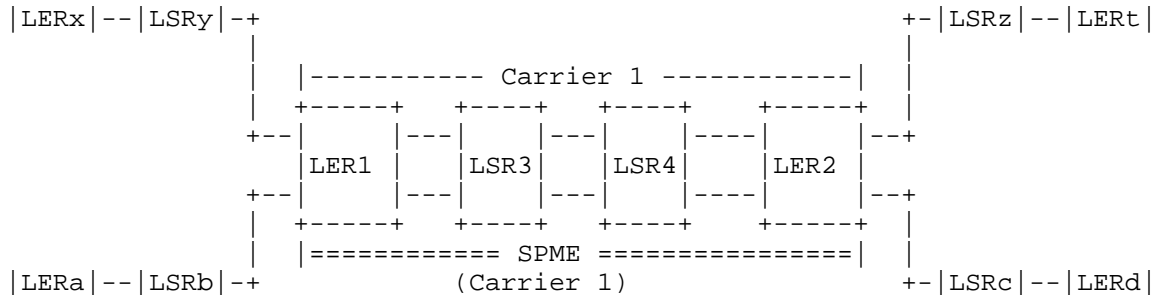


Figure 1: SPME for a set of transport path segments

In the MBB schemes, LER1 needs to inform the old LSPs's ingress nodes (for example, LERx and LERa) that a SPME has been setup to monitor the segment between LER1 and LER2, so that LERx/LERa can instantiate the new LSPs. However, the coordination schemes across administrative boundaries are not explicitly described in [RFC5921].

[RFC4736] gives the RSVP-TE extension to realize reoptimization of MPLS TE Loosely Routed LSP. It is said that when a mid-point LSR whose next hop is a loose hop or an abstract node can locally trigger a path re-evaluation when a configurable timer expires, some specific events occur (e.g., link-up event), or the user explicitly requests it. If a preferable path is found, the LSR sends an RSVP PathErr to the head-end LSR (Error code 25 (Notify), Error sub-code=6 ("preferable path exists")). Although SPME can be seen as a new link, the ingress nodes do know that they need to be triggered to establish new LSPs. In order to differentiate the cases between SPME and reoptimization, the new value "SPME up" is suggested to be assigned.

As we can see, network objective (1) can be met, but network objective (2) can not be met due to the new assignment of MPLS labels.

3.2. Local Rerouting

A bidirectional LSP1(LERx-LSRy-LER1-LSR3-LSR4-LER2-LSRz-LERt) exists between LERx and LERt, the forwarding label values along LERx->LERt direction are Lyx-Lly-L31-L43-L24-Lz2-Ltz, and the forwarding label values along LERt->LERx direction are Lxy-Ly1-L13-L34-L42-L2z-Lzt. Assuming that SPME1 (LER1-LSR3-LSR4- LER2) is established to monitor this LSP, in order to restrict the operation in the scope of Carrier

1, local rerouting technology described in [RFC4090] can be used here.

LER1 uses the label L24 as the inner label and pushes it into SPME1, LER2 uses the label L13 as the inner label and pushes it into SPME1. But LER1 (LER2) needs to learn L24 (L13), which can be learned by the following procedures:

When SPME1 is up, LER1 pushes LSP1's Path message into SPME1, the next hop is changed from LSR3 to LER2, upstream label unchanged (L13 is allocated to LER2). Similarly, LER2 pushes LSP1's Resv message into SPME1, the next hop is changed from LSR4 to LER1, and Label unchanged (L24 is allocated to LER1). After LER1 (LER2) has learned the inner label value L24 (L13), it can push the user traffic into SPME1.

If LSP1 is unidirectional, LER1 pushes LSP1's Path message into SPME1, the next hop is changed from LSR3 to LER2. But for Resv message, the next of LER2 is changed from LSR4 to LER1, and it needs to be transmitted hop by hop.

Local rerouting is more optimized compared to MBB, for the new assigned labels just exist in the scope of Carrier 1, but it still can not fully math the requirements of network objective (2).

3.3. TTL TLV

In order to totally meet the requirements of network objective (2), the schemes based on non-label stack are needed.

TTL TLV, as one of the optional ACH TLV objects, is defined in [I-D.boutros-mpls-lsp-ping-ttl-tlv], which is used to inform the receiver how many hops away the originator is on the path of the MS-PW or Bidirectional LSP. It can be used to realize path segment monitoring also, see the figure below.

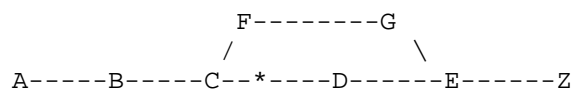


Figure 2: TTL TLV for PSM

The path segment PS1 (C-D-E) of LSP1 (A-B-C-D-E-Z) needs to be provisioned. Node C, as the MEP node of PS1, sends OAM message (like CC/CV, PM loss/dely, etc.) to node D, the TTL TLV MUST be inserted. TTL value is set to the hop counts from E to C, here it is 2 (if LSP1 is an associated bidirectional LSP, the hops form E to C maybe not be

2). In this way, node E can use the hops carried in TTL TLV to response the OAM message.

The TTL values can be configured by NMS, or learned by control plane. [I-D.ietf-mpls-tp-identifiers] describes the MEP-ID of Pseudowire Segments, and the MEP configuration of path Segments can be defined similarly. That is to say, the MEP_ID of path segment can be formed by a combination of a LSP MEP_ID and the identification of the local node, such as "Src-Global_ID::Src-Node_ID::Src-Tunnel_Num::LSP_Num::PS-Global_ID::PS-Node_ID".

3.3.1. The scaling Analysis

Assuming there is another path segment PS2 that exists between node B and E, proactive and on-demand OAM messages are running between B and E also. Just like PS1, the TTL TLV MUST be inserted, and the value is the hop counts from E to B (here it is 3). At some time, a defect happens between node C and D, the customer traffic would be switched from PS1 to the backup path(C---F---G---E). However, node B may not know that node C has switched all the traffic to the backup path, and in this case, node E can not receive the OAM message sent by node B and may make wrong decision.

In conclusion, TTL TLV scheme can meet both the two network objectives. But it can not be used if two or more path segments are nested.

4. IANA Considerations

A new error sub-code values for the RSVP PathErr Notify message (Error code=25) is required in this document:

Error sub-code=TBD by IANA: "SPME up".

5. Security Considerations

TBD.

6. Acknowledgement

The authors would like to thank Hui Su for the discussion, thank Alexander Vainshtein, Kannan KV Sampath, Nurit Sprecher, Yoshinori Koike for their valuable comments.

7. Normative references

- [I-D.boutros-mpls-lsp-ping-ttl-tlv]
Manral, V., Boutros, S., Sivabalan, S., Saxena, S., and G. Swallow, "Definition of Time-to-Live TLV for LSP-Ping Mechanisms", draft-boutros-mpls-lsp-ping-ttl-tlv-01 (work in progress), June 2010.
- [I-D.ietf-mpls-tp-identifiers]
Bocci, M. and G. Swallow, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-02 (work in progress), July 2010.
- [I-D.ietf-mpls-tp-oam-framework]
Allan, D., Busi, I., Niven-Jenkins, B., Fulignoli, A., Hernandez-Valencia, E., Levrau, L., Sestito, V., Sprecher, N., Helvoort, H., Vigoureux, M., Weingarten, Y., and R. Winter, "Operations, Administration and Maintenance Framework for MPLS- based Transport Networks", draft-ietf-mpls-tp-oam-framework-09 (work in progress), October 2010.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4736] Vasseur, JP., Ikejiri, Y., and R. Zhang, "Reoptimization of Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Loosely Routed Label Switched Path (LSP)", RFC 4736, November 2006.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

Authors' Addresses

Fei Zhang (editor)
ZTE Corporation
4F,RD Building 2,Zijinghua Road
Yuhuatai District,Nanjing 210012
P.R.China

Phone: +86 025 52877612
Email: zhang.fei3@zte.com.cn

LZ Jin
ZTE Corporation
889, Bibo Road, Zijinghua Road
Pudong District, Shanghai 201203
P.R.China

Phone: +86 021 68896273
Email: lizhong.jin@zte.com.cn

Bo Wu
ZTE Corporation
4F,RD Building 2,Zijinghua Road
Yuhuatai District,Nanjing 210012
P.R.China

Phone: +86 025 52877276
Email: wu.bo@zte.com.cn

MPLS Working Group
Internet-Draft
Intended status: Informational
Expires: September 4, 2014

H. van Helvoort, Ed.
Huawei Technologies
J. Ryoo, Ed.
ETRI
H. Zhang
Huawei Technologies
F. Huang
Philips
H. Li
China Mobile
A. D'Alessandro
Telecom Italia
March 3, 2014

Pre-standard Linear Protection Switching in MPLS-TP
draft-zulr-mpls-tp-linear-protection-switching-12.txt

Abstract

The IETF Standards Track solution for MPLS Transport Profile (MPLS-TP) Linear Protection is provided in RFC 6378, draft-ietf-mpls-psc-updates and draft-ietf-mpls-tp-psc-itu.

This document describes the pre-standard implementation of MPLS-TP Linear Protection that has been deployed by several network operators using equipment from multiple vendors. At the time of publication these pre-standard implementations were still in operation carrying live traffic.

The specified mechanism supports 1+1 unidirectional/bidirectional protection switching and 1:1 bidirectional protection switching. It is purely supported by MPLS-TP data plane, and can work without any control plane.

[Editor's note] The followings are to be included in "Status of Memo":

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

This is a contribution to the RFC Series, independently of any other RFC stream. The RFC Editor has chosen to publish this document at its discretion and makes no statement about its value for implementation or deployment. Documents approved for publication by the RFC Editor are not a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

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

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 September 4, 2014.

Copyright Notice

Copyright (c) 2014 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. Conventions Used in This Document	4
3. Acronyms	5
4. Linear protection switching overview	5
4.1. Protection architecture types	5
4.1.1. 1+1 architecture	6
4.1.2. 1:1 architecture	6
4.1.3. 1:n architecture	6
4.2. Protection switching type	6
4.3. Protection operation type	7

5.	Protection switching trigger conditions	7
5.1.	Fault conditions	7
5.2.	External commands	8
5.2.1.	End-to-end commands	8
5.2.2.	Local commands	9
6.	Protection switching schemes	9
6.1.	1+1 unidirectional protection switching	9
6.2.	1+1 bidirectional protection switching	10
6.3.	1:1 bidirectional protection switching	12
7.	APS protocol	13
7.1.	APS PDU format	13
7.2.	APS transmission	16
7.3.	Hold-off timer	16
7.4.	WTR timer	17
7.5.	Command acceptance and retention	18
7.6.	Exercise operation	18
8.	Protection switching logic	18
8.1.	Principle of operation	18
8.2.	Equal priority requests	21
8.3.	Signal degrade of the protection transport entity	22
9.	Protection switching state transition table	22
10.	Security considerations	23
11.	IANA considerations	24
12.	Acknowledgements	24
13.	References	24
13.1.	Normative References	24
13.2.	Informative References	25
Appendix A.	Operation examples of APS protocol	25
Authors' Addresses	31

1. Introduction

The IETF Standards Track solution for MPLS Transport Profile (MPLS-TP) Linear Protection is provided in RFC 6378 [RFC6378], draft-ietf-mpls-psc-updates [I-D.ietf-mpls-psc-updates] and draft-ietf-mpls-tp-psc-itu [I-D.ietf-mpls-tp-psc-itu].

This document describes the pre-standard implementation of MPLS-TP Linear Protection that has been deployed by several network operators using equipment from multiple vendors. At the time of publication these pre-standard implementations were still in operation carrying live traffic.

This implementation was considered in the MPLS WG, however a different path was chosen.

This document may be useful in the future if a vendor or operator is trying to interwork with a different vendor or operator who has

deployed the pre-standard implementation, and it provides a permanent record of the pre-standard implementation. It is also worth noting that the experience gained during deployment of the implementations of this document was used to refine [I-D.ietf-mpls-tp-psc-itu].

MPLS-TP is defined as the transport profile of MPLS technology to allow its deployment in transport networks. A typical feature of a transport network is that it can provide fast protection switching for end-to-end or segments. The protection switching time is generally required to be less than 50ms to meet the strict requirements of services such as voice, private line, etc.

The goal of a linear protection switching mechanism is to satisfy the requirement of fast protection switching for an MPLS-TP network. Linear protection switching means that, for one or more working transport entities (working paths), there is one protection transport entity (protection path), which is disjoint from any of working transport entities, ready to take over the service transmission when a working transport entity has failed.

This document specifies a 1+1 unidirectional protection switching mechanism for unidirectional transport entity (either point-to-point or point-to-multipoint) as well as a bidirectional point-to-point transport entity, and a 1+1/1:1 bidirectional protection switching mechanism for point-to-point bidirectional transport entity. Since bidirectional protection switching needs the coordination of the two endpoints of the transport entity, this document also specifies the Automatic Protection Switching (APS) protocol which is used for this purpose.

The linear protection mechanism described in this document is applicable to both Label Switched Paths (LSPs) and Pseudowires (PWs).

The APS protocol specified in this document is based on the same principles and behavior of the APS protocol designed for Synchronous Optical Network (SONET) [T1.105.01]/Synchronous Digital Hierarchy (SDH) [G.841], Optical Transport Network (OTN) [G.873.1] and Carrier Class Ethernet [G.8031] and provides commonality with the established operation models utilized in transport network technologies (e.g., SDH/SONET, OTN and Ethernet).

2. Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. Acronyms

This document uses the following acronyms:

APS	Automatic Protection Switching
DNR	Do not Revert
EXER	Exercise
G-ACh	Generic Associated Channel
FS	Forced Switch
LO	Lockout of Protection
LSP	Label Switched Path
MPLS-TP	MPLS Transport Profile
MS	Manual Switch
MS-P	Manual Switch to Protection transport entity
MS-W	Manual Switch to Working transport entity
NR	No Request
OAM	Operations, Administration, and Maintenance
OTN	Optical Transport Network
PDU	Protocol Data Unit
PW	Pseudowire
RR	Reverse Request
SD	Signal Degrad
SD-P	Signal Degrad on Protection transport entity
SD-W	Signal Degrad on Working transport entity
SDH	Synchronous Digital Hierarchy
SF	Signal Fail
SF-P	Signal Fail on Protection transport entity
SF-W	Signal Fail on Working transport entity
SONET	Synchronous Optical Network
WTR	Wait to Restore

4. Linear protection switching overview

To guarantee the protection switching time for a working transport entity, its protection transport entity is always pre-configured before the failure occurs. Normally, traffic will be transmitted and received on the working transport entity. Switching to the protection transport entity is usually triggered by link or node failure, external commands, etc. Note that external commands are often used in transport networks by operators, and they are very useful in cases of service adjustment, path maintenance, etc.

4.1. Protection architecture types

4.1.1. 1+1 architecture

In the 1+1 architecture, the protection transport entity is associated with a working transport entity. The normal traffic is permanently bridged onto both the working transport entity and the protection transport entity at the source endpoint of the protected domain. The normal traffic on working and protection transport entities is transmitted simultaneously to the destination sink endpoint of the protected domain, where a selection between the working and protection transport entity is made based on predetermined criteria, such as signal fail and signal degrade indications.

4.1.2. 1:1 architecture

In the 1:1 architecture, the protection transport entity is associated with a working transport entity. When the working transport entity is determined to be impaired, the normal traffic MUST be transferred from the working to the protection transport entity at both the source and sink endpoints of the protected domain. The selection between the working and protection transport entities is made based on predetermined criteria, such as signal fail and signal degrade indications from the working or protection transport entity.

The bridge at the source endpoint can be realized in two ways: it is either a selector bridge or a broadcast bridge. With a selector bridge the normal traffic is connected either to the working transport entity or the protection transport entity. With a broadcast bridge the normal traffic is permanently connected to the working transport entity, and in case a protection switch is active also to the protection transport entity. The broadcast bridge is recommended to be used in revertive mode only.

4.1.3. 1:n architecture

Details for the 1:n protection switching architecture are out of scope of this document and will be provided in a different document in the future.

It is worth noting that the APS protocol defined here is capable of supporting 1:n operations.

4.2. Protection switching type

The linear protection switching types can be a unidirectional switching type or a bidirectional switching type.

- o Unidirectional switching type: Only the affected direction of working transport entity is switched to protection transport entity; the selectors at each endpoint operate independently. This switching type is recommended to be used for 1+1 protection in this document.
- o Bidirectional switching type: Both directions of working transport entity, including the affected direction and the unaffected direction, are switched to protection transport entity. For bidirectional switching, the APS protocol is required to coordinate the two endpoints so that both have the same bridge and selector settings, even for a unidirectional failure. This type is applicable for 1+1 and 1:1 protection.

4.3. Protection operation type

The linear protection operation types can be a non-revertive operation type or a revertive operation type.

- o Non-revertive operation: The normal traffic will not be switched back to the working transport entity even after a protection switching cause has cleared. This is generally accomplished by replacing the previous switch request with a "Do not Revert (DNR)" request, which has a low priority.
- o Revertive operation: The normal traffic is restored to the working transport entity after the condition(s) causing the protection switching have cleared. In the case of clearing a command (e.g., Forced Switch), this happens immediately. In the case of clearing of a defect, this generally happens after the expiry of a "Wait to Restore (WTR)" timer, which is used to avoid chattering of selectors in the case of intermittent defects.

5. Protection switching trigger conditions

5.1. Fault conditions

Fault conditions mean the requests generated by the local Operations, Administration, and Maintenance (OAM) function.

- o Signal Failure (SF): If an endpoint detects a failure by an OAM function or other mechanism, it will submit a local signal failure (local SF) to APS module to request a protection switch. The local SF could be on the working transport entity (Signal Fail on Working transport entity (SF-W)) or the protection transport entity (Signal Fail on Protection transport entity (SF-P)).

- o Signal Degrade (SD): If an endpoint detects signal degradation by an OAM function or other mechanism, it will submit a local signal degrade (local SD) to the APS module to request a protection switching. The local SD could be on the working transport entity (Signal Degrade on Working transport entity (SD-W)) or the protection transport entity (Signal Degrade on Protection transport entity (SD-P)).

5.2. External commands

The external command issues an appropriate external request to the protection process.

5.2.1. End-to-end commands

These commands are applied to both local and remote nodes. When the APS protocol is present, these commands, except the Clear command, are signaled to the far end of the connection. In bidirectional switching, these commands affect the bridge and selector at both ends.

- o Lockout of Protection (LO): This command is used to provide the operator a tool for temporarily disabling access to the protection transport entity.
- o Manual switch (MS): This command is used to provide the operator a tool for temporarily switching normal traffic to the working transport entity (Manual Switch to Working transport entity (MS-W)) or to the protection transport entity (Manual Switch to Protection transport entity (MS-P)), unless a higher priority switch request (i.e., LO, FS, or SF) is in effect.
- o Forced switch (FS): This command is used to provide the operator a tool for temporarily switching normal traffic from working transport entity to protection transport entity, unless a higher priority switch request (i.e., LO or SF-P) is in effect.
- o Exercise (EXER): Exercise is a command to test if the APS communication is operating correctly. The EXER command SHALL NOT affect the state of the protection selector and bridge.
- o Clear: This command between management and local protection process is not a request sent by APS to other endpoints. It is used to clear the active near end external command or WTR state.

5.2.2. Local commands

These commands apply only to the near end (local node) of the protection group. Even when an APS protocol is supported, they are not signaled to the far end.

- o Freeze: This command freezes the state of the protection group. Until the freeze is cleared, additional near end commands are rejected and condition changes and received APS information are ignored. When the Freeze command is cleared, the state of the protection group is recomputed based on the condition and received APS information.

Because the freeze is local, if the freeze is issued at one end only, a failure of protocol can occur as the other end is open to accept any operator command or fault condition.

- o Clear Freeze: This command clears the local freeze.

6. Protection switching schemes

6.1. 1+1 unidirectional protection switching

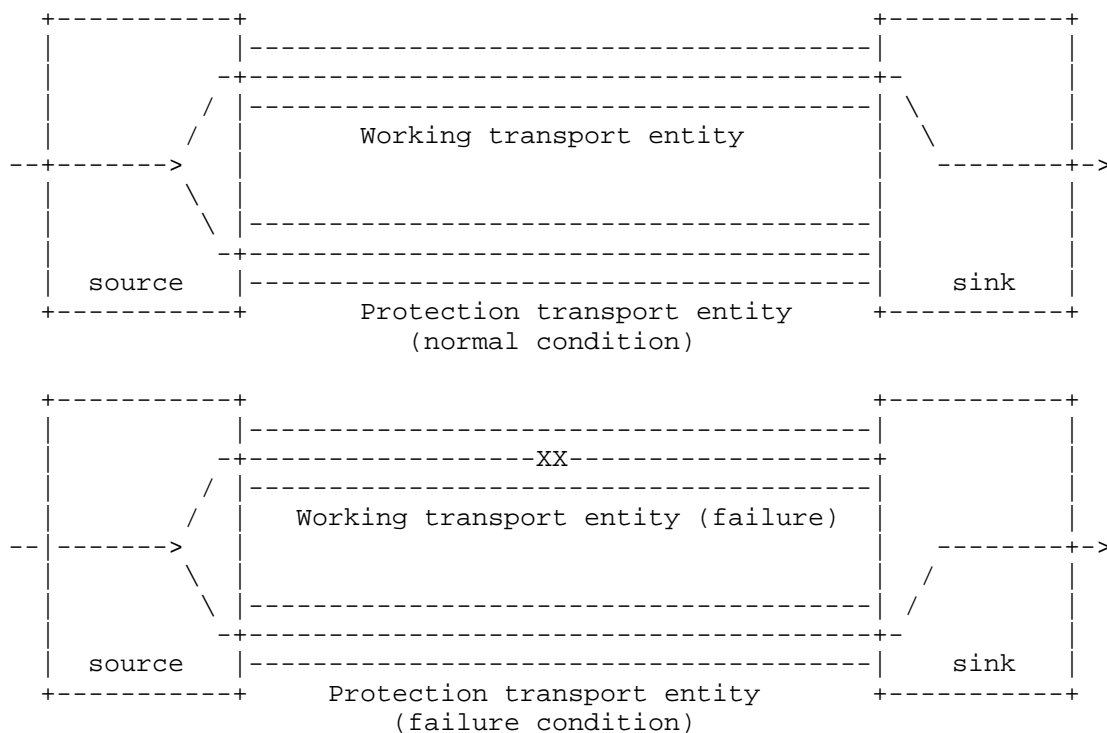


Figure 1: 1+1 unidirectional linear protection switching

1+1 unidirectional protection switching is the simplest protection switching mechanism. The normal traffic is permanently bridged on both the working and protection transport entities at the source endpoint of the protected domain. In the normal condition, the sink endpoint receives traffic from the working transport entity. If the sink endpoint detects a failure on the working transport entity, it will switch to receive traffic from the protection transport entity. 1+1 unidirectional protection switching is recommended to be used for unidirectional transport.

Note that 1+1 unidirectional protection switching does not use the APS coordination protocol since it only perform protection switching based on the local request.

6.2. 1+1 bidirectional protection switching

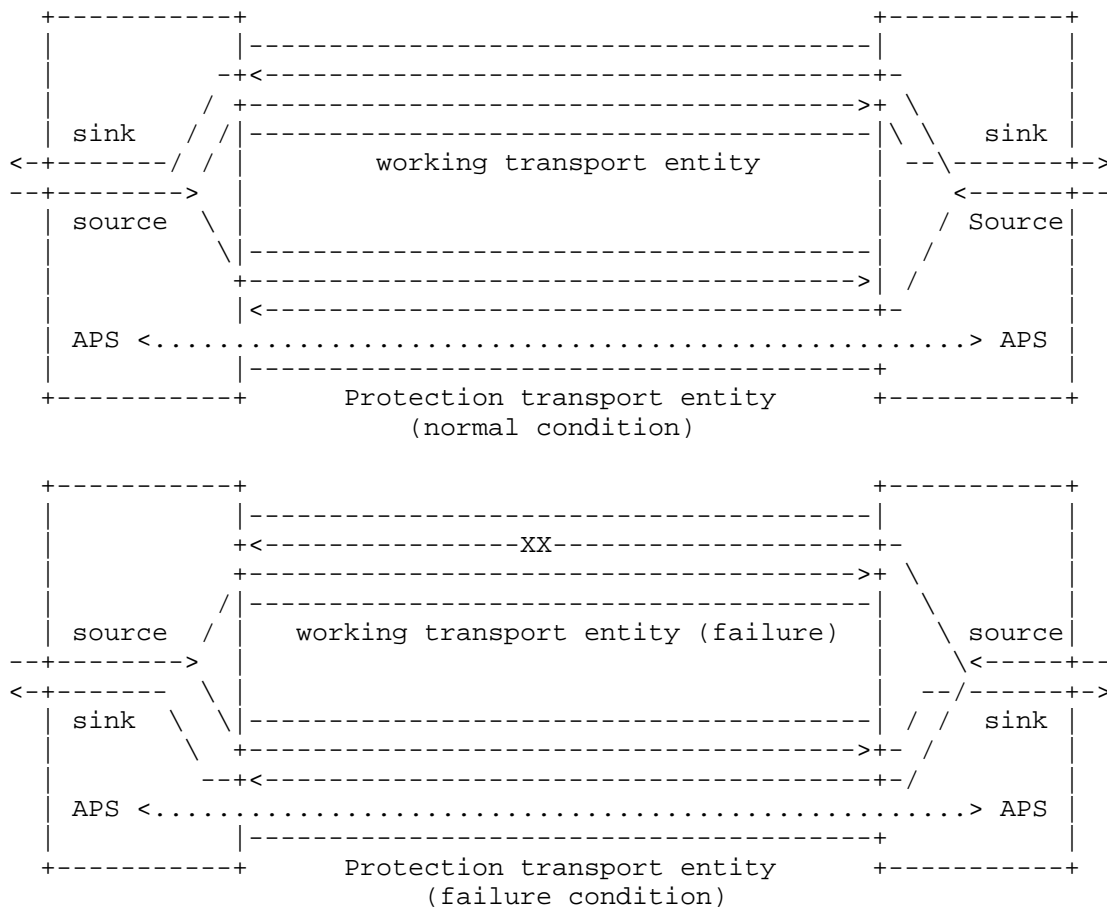


Figure 2: 1+1 bidirectional linear protection switching

In 1+1 bidirectional protection switching, for each direction, the normal traffic is permanently bridged on both the working and protection transport entities at the source endpoint of the protected domain. In the normal condition, for each direction, the sink endpoint receives traffic from the working transport entity.

If the sink endpoint detects a failure on the working transport entity, it will switch to receive traffic from the protection transport entity. It will also send an APS message to inform the sink endpoint on the other direction to switch to receive traffic from the protection transport entity.

The APS mechanism is necessary to coordinate the two endpoints of transport entity and implement 1+1 bidirectional protection switching even for a unidirectional failure.

6.3. 1:1 bidirectional protection switching

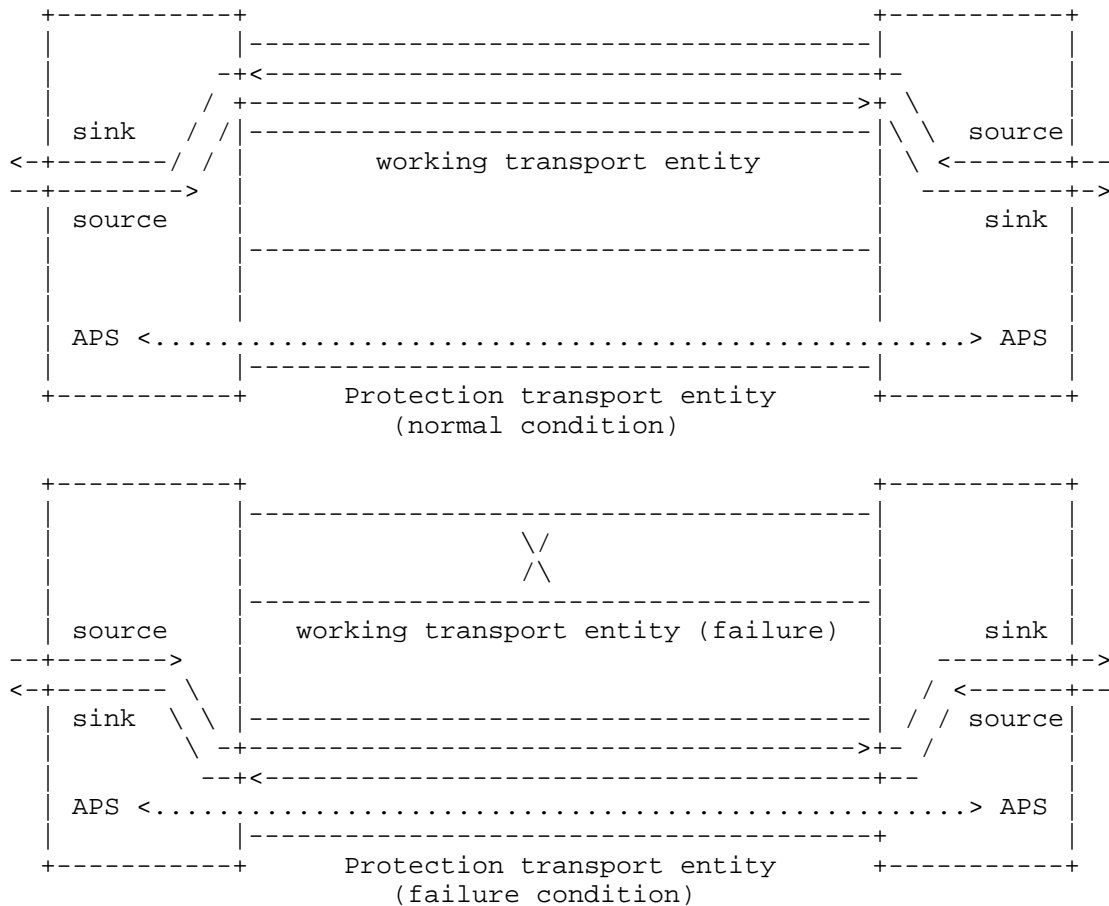


Figure 3: 1:1 bidirectional linear protection switching

In 1:1 bidirectional protection switching, for each direction, the source endpoint sends traffic on either the working transport entity or the protection transport entity. The sink endpoint receives the traffic from the transport entity where the source endpoint sends on.

In the normal condition, for each direction, the source endpoint and sink endpoint send and receive traffic from the working transport entity.

If the sink endpoint detects a failure on the working transport entity, it will switch to send and receive traffic from the protection transport entity. It will also send an APS message to inform the sink endpoint on another direction to switch to send and receive traffic from the protection transport entity.

The APS mechanism is necessary to coordinate the two endpoints of the transport entity and implement 1:1 bidirectional protection switching even for a unidirectional failure.

7. APS protocol

This APS protocol is based upon the APS protocol defined in Clause 11 of [G.8031]. See that reference for further definition of the PDU fields and protocol details beyond the description in this document.

7.1. APS PDU format

APS packets MUST be sent over a Generic Associated Channel (G-ACh) as defined in RFC 5586 [RFC5586].

The format of APS Protocol Data Unit (PDU) is specified in Figure 4 below.

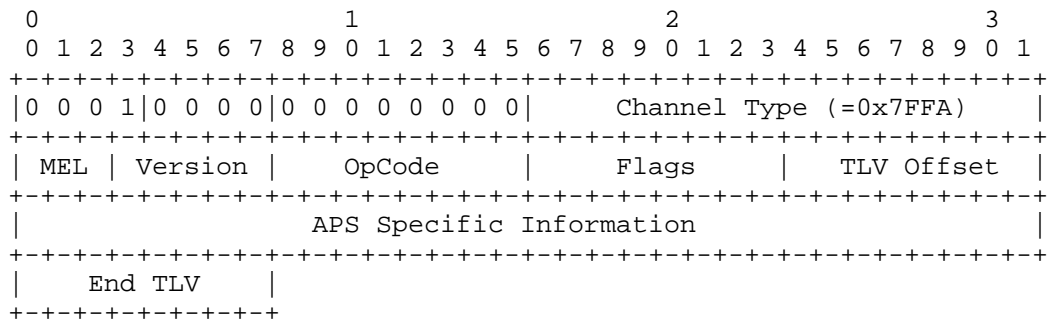


Figure 4: APS PDU format

The following values MUST be used for APS PDU:

- o Channel Type: The Channel Type MUST be configurable by the implementation. During deployment the local system administrator provisioned the value 0x7FFA. This is a code point value in the range of experimental Channel Types as described in RFC 5586 section 10.
- o MEL (Maintenance Entity group Level): The MEL value to set and check MUST be configurable. The DEFAULT value MUST be "111".

With co-routed bidirectional transport paths, the configured MEL MUST be the same in both directions.

- o Version: 0x00
- o OpCode: 0x27 (=0d39)
- o Flags: 0x00
- o TLV Offset: 4
- o End TLV: 0x00

The format of the APS-specific information is defined in Figure 5.

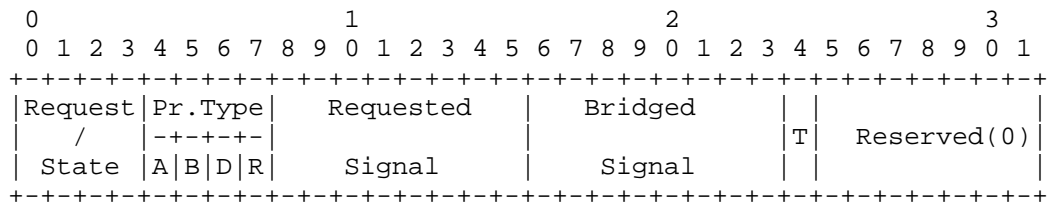


Figure 5: APS specific information format

All bits defined as "Reserved" MUST be transmitted as 0 and ignored on reception.

- o Request/State:

The four bits indicate the protection switching request type. See Figure 6 for the code of each request/state type.

In case that there are multiple protection switching requests, only the protection switching request with the highest priority MUST be processed.

Request/State	code/priority
Lockout of Protection (LO)	1111 (highest)
Signal Fail on Protection (SF-P)	1110
Forced Switch (FS)	1101
Signal Fail on Working (SF-W)	1011
Signal Degrade (SD)	1001
Manual Switch (MS)	0111
Wait to Restore (WTR)	0101
Exercise (EXER)	0100
Reverse Request (RR)	0010
Do Not Revert (DNR)	0001
No Request (NR)	0000 (lowest)

Figure 6: Protection switching request code/priority

- o Protection type (Pr.Type):

The four bits are used to specify the protection type.

A: reserved (set by default to 1)
 B: 0 - 1+1 (permanent bridge)
 1 - 1:1 (no permanent bridge)
 D: 0 - Unidirectional switching
 1 - Bidirectional switching
 R: 0 - Non-revertive operation
 1 - Revertive operation

- o Requested Signal:

This byte is used to indicate the traffic that the near end requests to be carried over the protection entity.

value = 0: Null traffic
 value = 1: Normal traffic 1
 value = 2~255: Reserved

- o Bridged Signal:

This byte is used to indicate the traffic that is bridged onto the protection entity.

value = 0: Null traffic

value = 1: Normal traffic 1

value = 2~255: Reserved

- o Bridge Type (T):

This bit is used to further specify the type of non-permanent bridge for 1:1 protection switching.

value = 0: Selector bridge

value = 1: Broadcast bridge

- o Reserved:

This field MUST be set to zero.

7.2. APS transmission

The APS message MUST be transported on the protection transport entity by encapsulation with the protection transport entity label (the label of the LSP used to transport protection traffic). If an endpoint receives APS-specific information from the working transport entity, it MUST ignore this information, and MUST report the Failure of Protocol defect (see Section 8.1) to the operator.

A new APS packet MUST be transmitted immediately when a change in the transmitted status occurs. The first three APS packets MUST be transmitted as fast as possible only if the APS information to be transmitted has been changed so that fast protection switching is possible even if one or two APS packets are lost or corrupted. The interval of the first three APS packets SHOULD be 3.3ms. APS packets after the first three MUST be transmitted with the interval of 5 seconds.

If no valid APS-specific information is received, the last valid received information remains applicable.

7.3. Hold-off timer

In order to coordinate timing of protection switches at multiple layers, a hold-off timer MAY be required. The purpose is to allow a server layer protection switch to have a chance to fix the problem before switching at a client layer.

Each selector SHOULD have a provisioned hold-off timer. The suggested range of the hold-off timer is 0 to 10 seconds in steps of 100 ms (accuracy of +/-5 ms).

When a new defect or more severe defect occurs (new SF or SD) on the active transport entity (the transport entity that currently carries and selects traffic), this event will not be reported immediately to protection switching if the provisioned hold-off timer value is non-zero. Instead, the hold-off timer SHALL be started. When the hold-off timer expires, it SHALL be checked whether a defect still exists on the transport entity that started the timer. If it does, that defect SHALL be reported to protection switching. The defect need not be the same one that started the timer.

This hold-off timer mechanism SHALL be applied for both working and protection transport entities.

7.4. WTR timer

In revertive mode of operation, to prevent frequent operation of the protection switch due to an intermittent defect, a failed working transport entity MUST become fault-free. After the failed working transport entity meets this criterion, a fixed period of time SHALL elapse before a normal traffic signal uses it again. This period, called a WTR period, MAY be configured by the operator in 1 minute steps between 5 and 12 minutes; the default value is 5 minutes. An SF or SD condition will override the WTR. To activate the WTR timer appropriately, even when both ends concurrently detect clearance of SF-W and SD-W, when the local state transits from SF-W or SD-W to No Request (NR) with the requested signal number 1, the previous local state, SF-W or SD-W, MUST be memorized. If both the local state and far-end state are NR with the requested signal number 1, the local state transits to WTR only when the previous local state is SF-W or SD-W. Otherwise, the local state transits to NR with the requested signal number 0.

In revertive mode of operation, when the protection is no longer requested, i.e., the failed working transport entity is no longer in SF or SD condition (and assuming no other requesting transport entities), a local WTR state will be activated. Since this state becomes the highest in priority, it is indicated on the APS signal, and maintains the normal traffic signal from the previously failed working transport entity on the protection transport entity. This state SHALL normally time out and become a NR state. The WTR timer deactivates earlier when any request of higher priority request pre-empt this state.

7.5. Command acceptance and retention

The commands Clear, LO, FS, MS, and EXER are accepted or rejected in the context of previous commands, the condition of the working and protection entities in the protection group, and (in bidirectional switching only) the APS information received.

The Clear command MUST be only valid if a near end LO, FS, MS, or EXER command is in effect, or if a WTR state is present at the near end and rejected otherwise. This command will remove the near-end command or WTR state, allowing the next lower-priority condition or (in bidirectional switching) APS request to be asserted.

Other commands MUST be rejected unless they are higher priority than the previously existing command, condition, or (in bidirectional switching) APS request. If a new command is accepted, any previous, lower-priority command that is overridden MUST be forgotten. If a higher priority command overrides a lower-priority condition or (in bidirectional switching) APS request, that other request will be reasserted if it still exists at the time the command is cleared. If a command is overridden by a condition or (in bidirectional switching) APS request, that command MUST be forgotten.

7.6. Exercise operation

Exercise is a command to test if the APS communication is operating correctly. It is lower priority than any "real" switch request. It is only valid in bidirectional switching, since this is the only place where you can get a meaningful test by looking for a response.

The Exercise command SHALL issue the command with the same requested and bridged signal numbers of the NR, Reverse Request (RR) or DNR request that it replaces. The valid response will be an RR with the corresponding requested and bridged signal numbers. When Exercise commands are input at both ends, an EXER, instead of RR, MUST be transmitted from both ends. The standard response to DNR MUST be DNR rather than NR. When the exercise command is cleared, it MUST be replaced with NR or RR if the requested signal number is 0, and DNR or RR if the requested signal number is 1.

8. Protection switching logic

8.1. Principle of operation

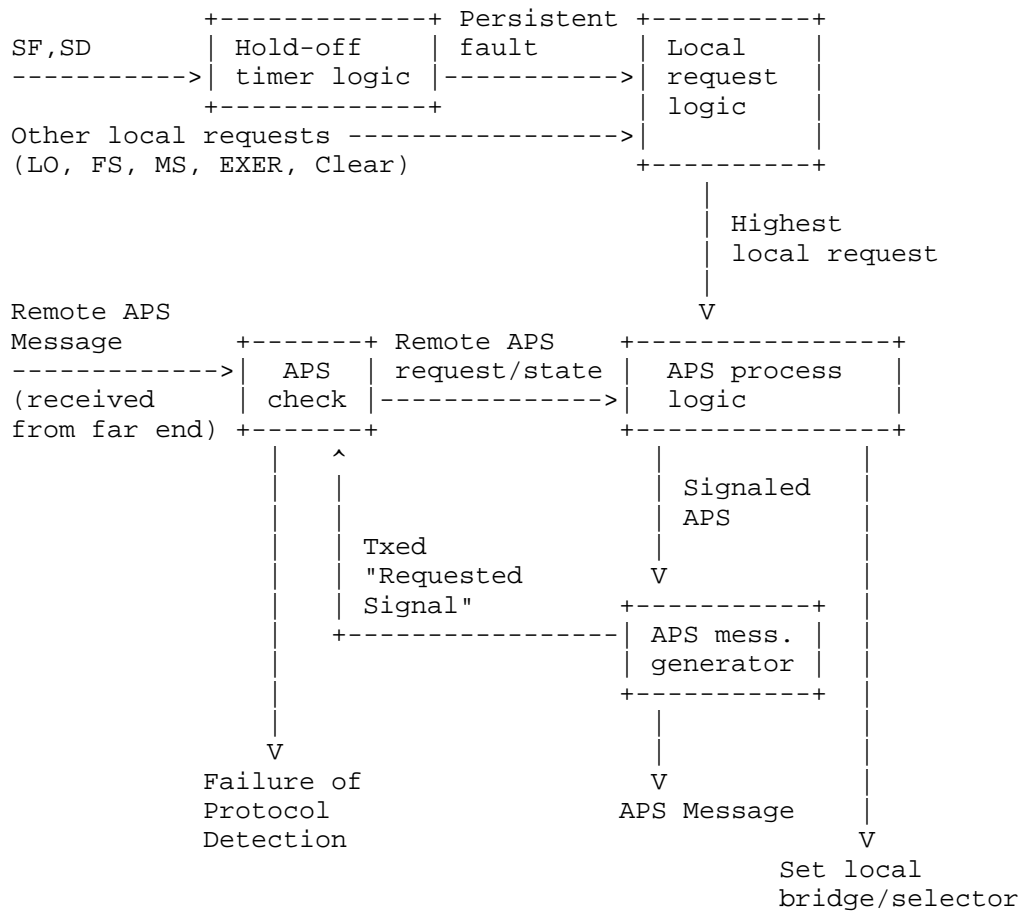


Figure 7: Protection Switching Logic

Figure 7 describes the protection switching logic.

One or more local protection switching requests may be active. The "local request logic" determines which of these requests is highest using the order of priority given in Figure 6. This highest local request information SHALL be passed on to the "APS process logic". Note that an accepted Clear command, clearance of SF or SD or expiration of WTR timer SHALL NOT be processed by the local request logic, but SHALL be considered as the highest local request and submitted to the APS process logic for processing.

The remote APS message is received from the far end and is subjected to the validity check and mismatch detection in "APS check". Failure of Protocol situations are as follows:

- o The "B" field mismatch due to incompatible provisioning;
- o The reception of APS message from the working entity due to working/protection configuration mismatch;
- o No match in sent "Requested Signal" and received "Requested Signal" for more than 50 ms;
- o No APS message is received on the protection transport entity during at least 3.5 times the long APS interval (e.g. at least 17.5 seconds) and there is no defect on the protection transport entity.

Provided the "B" field matches:

- o If "D" bit mismatches, the bidirectional side will fall back to unidirectional switching.
- o If the "R" bit mismatches, one side will clear switches to WTR and the other will clear to DNR. The two sides will interwork and the traffic is protected.
- o If the "T" bit mismatches, the side using a broadcast bridge will fall back to using a selector bridge.

The APS message with invalid information MUST be ignored, and the last valid received information remains applicable.

The linear protection switching algorithm SHALL commence immediately every time one of the input signals changes, i.e., when the status of any local request changes, or when a different APS specific information is received from the far end. The consequent actions of the algorithm are also initiated immediately, i.e., change the local bridge/selector position (if necessary), transmit a new APS specific information (if necessary), or detect the failure of protocol defect if the protection switching is not completed within 50 ms.

The state transition is calculated in the "APS process logic" based on the highest local request, the request of the last received "Request/State" information, and state transition tables defined in Section 9, as follows:

- o If the highest local request is Clear, clearance of SF or SD, or expiration of WTR, a state transition is calculated first based on the highest local request and state machine table for local requests to obtain an intermediate state. This intermediate state is the final state in case of clearance of SF-P otherwise, starting at this intermediate state, the last received far end

request and the state machine table for far end requests are used to calculate the final state.

- o If the highest local request is neither Clear, nor clearance of SF or of SD, nor expiration of WTR, the APS process logic compares the highest local request with the request of the last received "Request/State" information based on Figure 6.
 1. If the highest local request has higher or equal priority, it is used with the state transition table for local requests defined in Section 9 to determine the final state; otherwise
 2. The request of the last received "Request/State" information is used with the state transition table for far end requests defined in Section 9 to determine the final state.

The "APS message generator" generates APS specific information with the signaled APS information for the final state from the state transition calculation (with coding as described in Figure 5).

8.2. Equal priority requests

In general, once a switch has been completed due to a request, it will not be overridden by another request of the same priority (first-come, first-served policy). Equal priority requests from both sides of a bidirectional protection group are both considered valid, as follows:

- o If the local state is NR, with the requested signal number 1, and the far-end state is NR, with the requested signal number 0, the local state transits to NR with the requested signal number 0. This applies to the case when the remote request for switching to the protection transport entity has been cleared.
- o If both the local and far-end states are NR, with the requested signal number 1, the local state transits to the appropriate new state (DNR state for non-revertive mode and WTR state for revertive mode). This applies to the case when the old request has been cleared at both ends.
- o If both the local and far-end states are RR, with the same requested signal number, both ends transit to the appropriate new state according to the requested signal number. This applies to the case of concurrent deactivation of EXER from both ends.
- o In other cases, no state transition occurs, even if equal priority requests are activated from both ends. Note that if MSs are issued simultaneously to both working and protection transport

entities, either as local or far-end requests, the MS to the working transport entity is considered as having higher priority than the MS to the protection transport entity.

8.3. Signal degrade of the protection transport entity

Signal degrade on protection transport entity has the same priority as signal degrade on working transport entity. As a result, if an SD condition affects both transport entities, the first SD detected MUST NOT be overridden by the second SD detected. If the SD is detected simultaneously, either as local or far-end requests on both working and protection transport entities, then the SD on the standby transport entity MUST be considered as having higher priority than the SD on the active transport entity, and the normal traffic signal continues to be selected from the active transport entity (i.e., no unnecessary protection switching is performed).

In the preceding sentence, "simultaneously" relates to the occurrence of SD on both the active and standby transport entities at input to the protection switching process at the same time, or as long as a SD request has not been acknowledged by the remote end in bidirectional protection switching.

9. Protection switching state transition table

In this section, state transition tables for the following protection switching configurations are described.

- o 1:1 bidirectional (revertive mode, non-revertive mode);
- o 1+1 bidirectional (revertive mode, non-revertive mode);
- o 1+1 unidirectional (revertive mode, non-revertive mode).

Note that any other global or local request which is not described in state transition tables does not trigger any state transition.

The states specified in the state transition tables can be described as follows:

- o NR: NR is the state entered by the local priority under all conditions where no local protection switching requests (including WTR and DNR) are active. NR can also indicate that the highest local request is overridden by the far end request, whose priority is higher than the highest local request. Normal traffic signal is selected from the corresponding transport entity.

- o LO, SF-P, SD-P: The access by the normal traffic to the protection transport entity is NOT allowed in this state. The normal traffic is carried by the working transport entity, regardless of the fault/degrade condition possibly present (due to the highest priority of the switching triggers leading to this state).
- o FS, SF-W, SD-W, MS-W, MS-P: A switching trigger, NOT resulting in the protection transport entity unavailability is present. The normal traffic is selected either from the corresponding working transport entity or from the protection transport entity, according to the behavior of the specific switching trigger.
- o WTR: In revertive operation, after the clearing of an SF-W or SD-W, maintains normal traffic as selected from the protection transport entity until the WTR timer expires or another request with higher priority, including Clear command, is received. This is used to prevent frequent operation of the selector in the case of intermittent failures.
- o DNR: In non-revertive operation, this is used to maintain a normal traffic to be selected from the protection transport entity.
- o EXER: Exercise of the APS protocol.
- o RR: The near end will enter and signal Reverse Request only in response to an EXER from the far end.

[State transition tables are shown at the end of the PDF form of this document.]

10. Security considerations

MPLS-TP is a subset of MPLS and so builds upon many of the aspects of the security model of MPLS. MPLS networks make the assumption that it is very hard to inject traffic into a network and equally hard to cause traffic to be directed outside the network. The control-plane protocols utilize hop-by-hop security and assume a "chain-of-trust" model such that end-to-end control-plane security is not used. For more information on the generic aspects of MPLS security, see RFC 5920 [RFC5920].

This document describes a protocol carried in the G-ACh [RFC5586], and so is dependent on the security of the G-ACh, itself. The G-ACh is a generalization of the Associated Channel defined in [RFC4385]. Thus, this document relies heavily on the security mechanisms provided for the Associated Channel and described in those two documents.

11. IANA considerations

There are no IANA actions requested.

12. Acknowledgements

The authors would like to thank Hao Long, Vincenzo Sestito, Italo Busi, Igor Umansky, and Andy Malis for their input to and review of the current document.

13. References

13.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [T1.105.01] American National Standards Institute, "Synchronous Optical Network (SONET) - Automatic Protection Switching", ANSI 0900105.01:2000 (R2010), 2000.
- [G.841] International Telecommunications Union, "Types and characteristics of SDH network protection architectures", ITU-T Recommendation G.841, October 1998.
- [G.873.1] International Telecommunications Union, "Optical Transport Network (OTN): Linear protection", ITU-T Recommendation G.873.1, July 2011.
- [G.8031] International Telecommunications Union, "Ethernet Linear Protection Switching", ITU-T Recommendation G.8031/Y.1342, June 2011.

13.2. Informative References

[RFC6378] Weingarten, Y., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, October 2011.

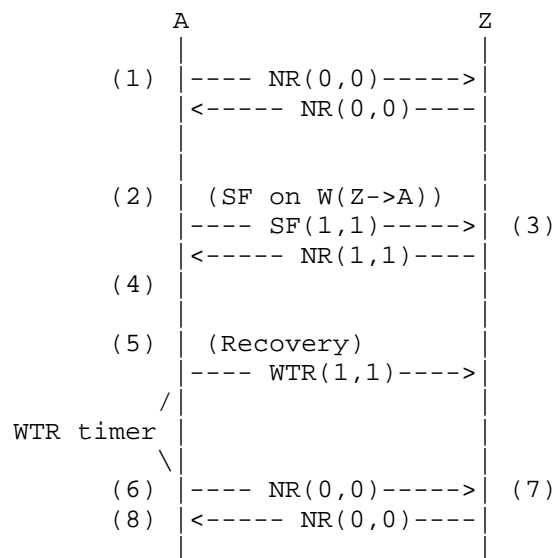
[I-D.ietf-mpls-psc-updates]
Osborne, E., "Updates to PSC", draft-ietf-mpls-psc-updates-01 (work in progress), January 2014.

[I-D.ietf-mpls-tp-psc-itu]
Ryoo, J., Gray, E., van Helvoort, H., D'Alessandro, A., Cheung, T., and E. Osborne, "MPLS Transport Profile (MPLS-TP) Linear Protection to Match the Operational Expectations of SDH, OTN and Ethernet Transport Network Operators", draft-ietf-mpls-tp-psc-itu-02 (work in progress), February 2014.

Appendix A. Operation examples of APS protocol

The sequence diagrams shown in this section are only a few examples of the APS operations. The first APS message which differs from the previous APS message is shown. The operation of hold-off timer is omitted. The fields whose values are changed during APS packet exchange are shown in the APS packet exchange. They are Request/State, requested traffic, and bridged traffic. For an example, SF(0,1) represents an APS packet with the following field values: Request/State = SF, Requested Signal = 0, and Bridged Signal = 1. The values of the other fields remain unchanged from the initial configuration. The signal numbers 0 and 1 refer to null signal and normal traffic signal, respectively. W(A->Z) and P(A->Z) indicate the working and protection paths in the direction of A to Z, respectively.

Example 1. 1:1 bidirectional protection switching (revertive mode) - Unidirectional SF case



(1) The protected domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Signal Fail occurs on the working entity in the Z to A direction. Selector and bridge of node A select protection entity. Node A generates SF(1,1) message.

(3) Upon receiving SF(1,1), node Z sets selector and bridge to protection entity. As there is no local request in node Z, node Z generates NR(1,1) message.

(4) Node A confirms that the far end is also selecting protection entity.

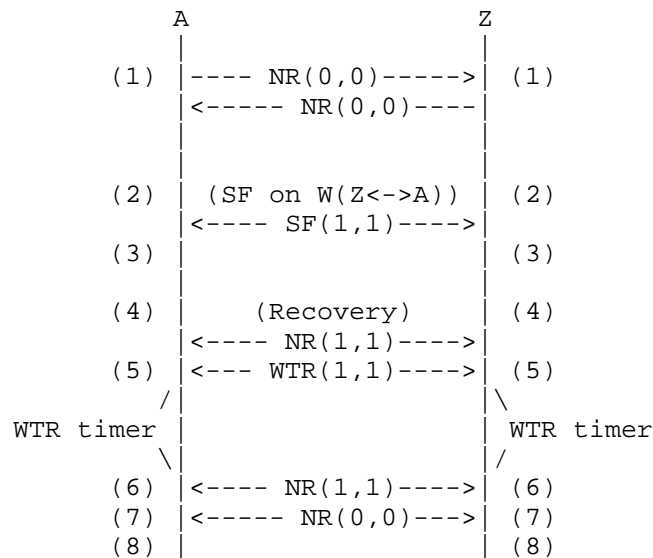
(5) Node A detects clearing of SF condition, starts the WTR timer, and sends WTR(1,1) message.

(6) At expiration of the WTR timer, node A sets selector and bridge to working entity and sends NR(0,0) message.

(7) Node Z is notified that the far end request has been cleared, and sets selector and bridge to working entity.

(8) It is confirmed that the far end is also selecting working entity.

Example 2. 1:1 bidirectional protection switching (revertive mode) - Bidirectional SF case



(1) The protected domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Nodes A and Z detect local Signal Fail conditions on the working entity, set selector and bridge to protection entity, and generate SF(1,1) messages.

(3) Upon receiving SF(1,1), each node confirms that the far end is also selecting protection entity.

(4) Each node detects clearing of SF condition, and sends NR(1,1) message as the last received APS message was SF.

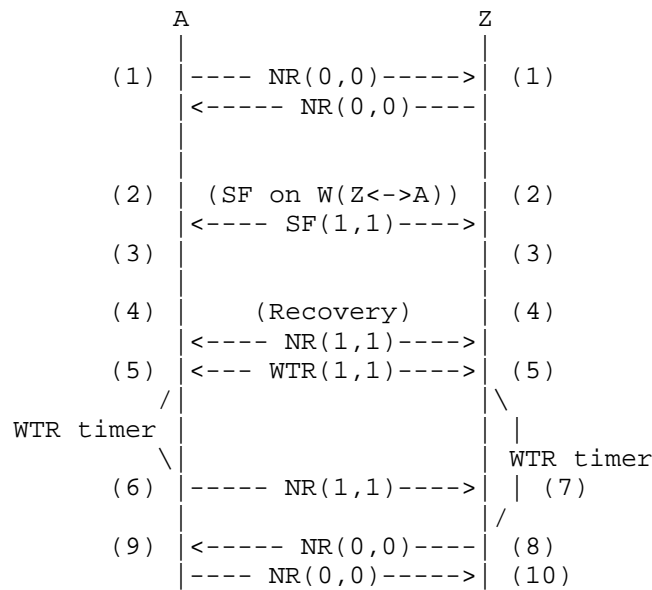
(5) Upon receiving NR(1,1), each node starts the WTR timer and sends WTR(1,1).

(6) At expiration of the WTR timer, each node sends NR(1,1) as the last received APS message was WTR.

(7) Upon receiving NR(1,1), each node sets selector and bridge to working entity and sends NR(0,0) message.

(8) It is confirmed that the far end is also selecting working entity.

Example 3. 1:1 bidirectional protection switching (revertive mode) - Bidirectional SF case - Inconsistent WTR timers



(1) The protected domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Nodes A and Z detect local Signal Fail conditions on the working entity, set selector and bridge to protection entity, and generate SF(1,1) messages.

(3) Upon receiving SF(1,1), each node confirms that the far end is also selecting protection entity.

(4) Each node detects clearing of SF condition, and sends NR(1,1) message as the last received APS message was SF.

(5) Upon receiving NR(1,1), each node starts the WTR timer and sends WTR(1,1).

(6) At expiration of the WTR timer in node A, node A sends NR(1,1) as the last received APS message was WTR.

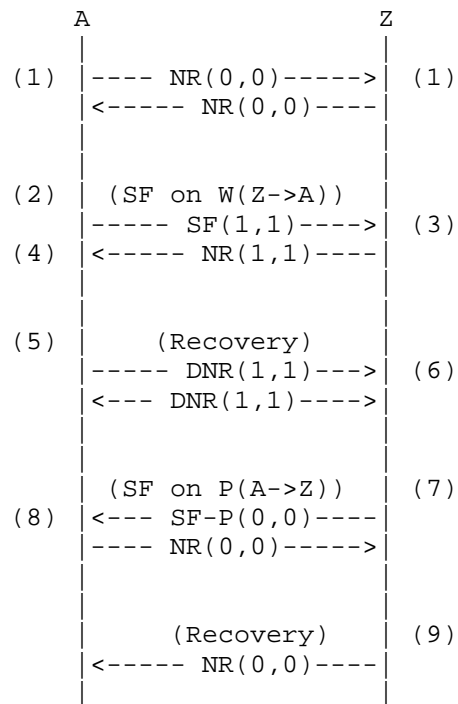
(7) At node Z, the received NR(1,1) is ignored as the local WTR has a higher priority.

(8) At expiration of the WTR timer in node Z, node Z sets selector and bridge to working entity, and sends NR(0,0) message.

(9) Upon receiving NR(0,0), node A sets selector and bridge to working entity and sends NR(0,0) message.

(10) It is confirmed that the far end is also selecting working entity.

Example 4. 1:1 bidirectional protection switching (non-revertive mode) - Unidirectional SF on working followed by unidirectional SF on protection



(1) The protected domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Signal Fail occurs on the working entity in the Z to A direction. Selector and bridge of node A select the protection entity. Node A generates SF(1,1) message.

(3) Upon receiving SF(1,1), node Z sets selector and bridge to protection entity. As there is no local request in node Z, node Z generates NR(1,1) message.

(4) Node A confirms that the far end is also selecting protection entity.

(5) Node A detects clearing of SF condition, and sends DNR(1,1) message.

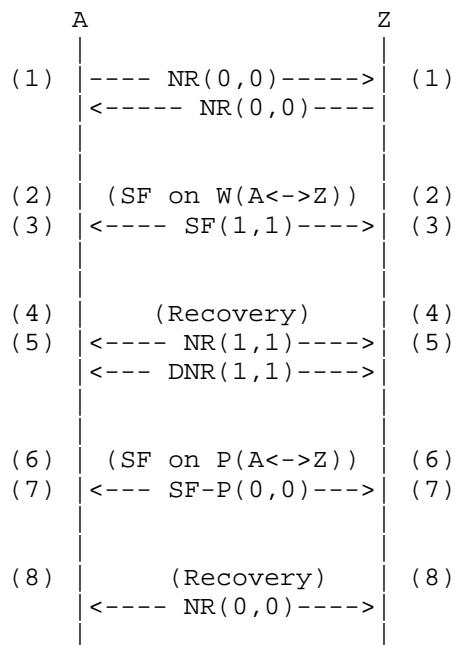
(6) Upon receiving DNR(1,1), node Z also generates DNR(1,1) message.

(7) Signal Fail occurs on the protection entity in the A to Z direction. Selector and bridge of node Z select the working entity. Node Z generates SF-P(0,0) message.

(8) Upon receiving SF-P(0,0), node A sets selector and bridge to working entity, and generates NR(0,0) message.

(9) Node Z detects clearing of SF condition, and sends NR(0,0) message.

Exmaple 5. 1:1 bidirectional protection switching (non-revertive mode) - Bidirectional SF on working followed by bidirectional SF on protection



(1) The protected domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Nodes A and Z detect local Signal Fail conditions on the working entity, set selector and bridge to protection entity, and generate SF(1,1) messages.

(3) Upon receiving SF(1,1), each node confirms that the far end is also selecting protection entity.

(4) Each node detects clearing of SF condition, and sends NR(1,1) message as the last received APS message was SF.

(5) Upon receiving NR(1,1), each node sends DNR(1,1).

(6) Signal Fail occurs on the protection entity in both directions. Selector and bridge of each node selects the working entity. Each node generates SF-P(0,0) message.

(7) Upon receiving SF-P(0,0), each node confirms that the far end is also selecting working entity

(8) Each node detects clearing of SF condition, and sends NR(0,0) message.

Authors' Addresses

Huub van Helvoort (editor)
Huawei Technologies

Email: huub.van.helvoort@huawei.com

Jeong-dong Ryoo (editor)
ETRI

Email: ryoo@etri.re.kr

Haiyan Zhang
Huawei Technologies

Email: zhanghaiyan@huawei.com

Feng Huang
Philips

Email: feng.huang@philips.com

Han Li
China Mobile

Email: lihan@chinamobile.com

Alessandro D'Alessandro
Telecom Italia

Email: alessandro.dalessandro@telecomitalia.it