Path MTU Discovery in Bridged Network

Authors: **Hesham ElBakoury**

Version: 1.0

HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

GOAL

Discuss different ideas to discover the Path MTU in Bridged Ethernet Network with the intent of embarking on a protocol to be used in IEEE 1904.2

Agenda

- **Overview of IP Path MTU Protocols.**
- **Discuss three ideas:**
  - Using IEEE 1904.2
  - Using IEEE 802.1ag /Y.1732 (CFM) Linktrace with Enhancement.
    - **An overview of CFM is provided.**
  - Using Probes
    - **LLC TEST Command is used as an example**
- **Discuss if any of these ideas or others can be used for a baseline proposal for PMTU Discovery in IEEE 1904.2**

IP Path MTU discovery

- **Path MTU discovery is described in**
  - RFC 1191 for IPv4
  - RFC 1981 for IPV6
  - Both RFCs use IP data packets to detect Path MTU.
- **Hosts dynamically discover minimum MTU of path**
- **Algorithm:**
  - Initialize MTU to MTU for first hop
  - Send datagrams with Don't Fragment bit set
  - If ICMP "pkt too big" msg, decrease MTU
- **What happens if path changes?**
  - Periodically (>5mins, or >1min after previous increase), increase MTU
- **Some routers will return MTU size of next hop**
- **RFC 4821 Uses in-band probes to estimate path MTU.**

1 Using IEEE 802.1AB

IEEE 802.3 Maximum Frame Size TLV

Differences in maximum frame size can result in loss of frames if a sending station transmits frames larger than the advertised maximum frame size supported by the receiving station. The Maximum Frame Size TLV can be used to detect mis-configurations or incompatibility between two stations with different maximum supported frame sizes

F.4 Maximum Frame Size TLV

The Maximum Frame Size TLV indicates the maximum frame size capability of the implemented MAC and PHY. Figure F.3 shows the format of this TLV.

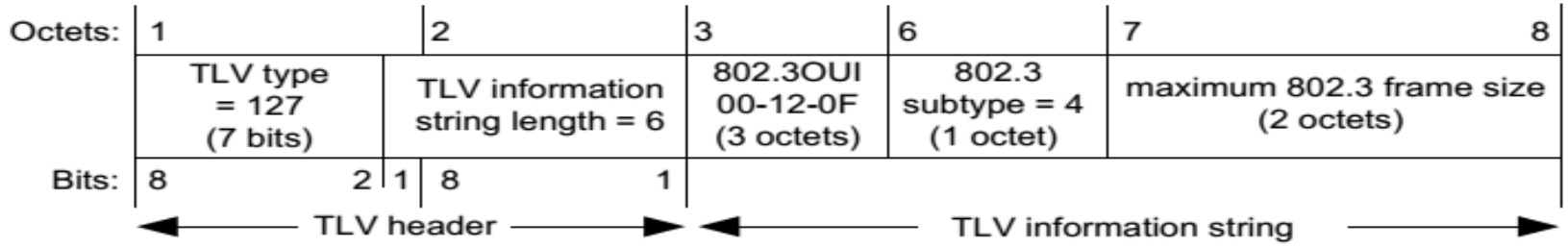


Figure F.3—Maximum Frame Size TLV format

F.4.1 maximum frame size

The maximum frame size field shall contain an integer value indicating the maximum supported frame size in octets as determined by the following:

a) If the MAC/PHY supports only the basic MAC frame format as defined in 3.1.1 of IEEE Std 802.3-2008, the maximum frame size field shall be set to 1518.

b) If the MAC/PHY supports an extension of the basic MAC frame format for Tagged MAC frames as defined 3.5 of IEEE Std 802.3-2008, the maximum frame size field shall be set to 1522.

c) If the MAC/PHY supports an extension of the MAC frame format that is different from either of the above, the maximum frame size field shall be set to the maximum value supported.

Using IEEE 802.1AB For MTU Discovery

- **Centralized Management System can use IEEE 802.1AB to build the L2 Topology.**
- **Assuming that speed mismatch on any hob is detected and fixed, Management system can find out from IEEE 802.1AB MIB the max frame size of each hob on any path. From that it can calculate the path MTU.**
- **While it is possible to send an LLDP frame to a specific destination, it wouldn't directly tell the smallest max frame size unless we use a trial and error approach.**
  - A central view provides more accurate results.

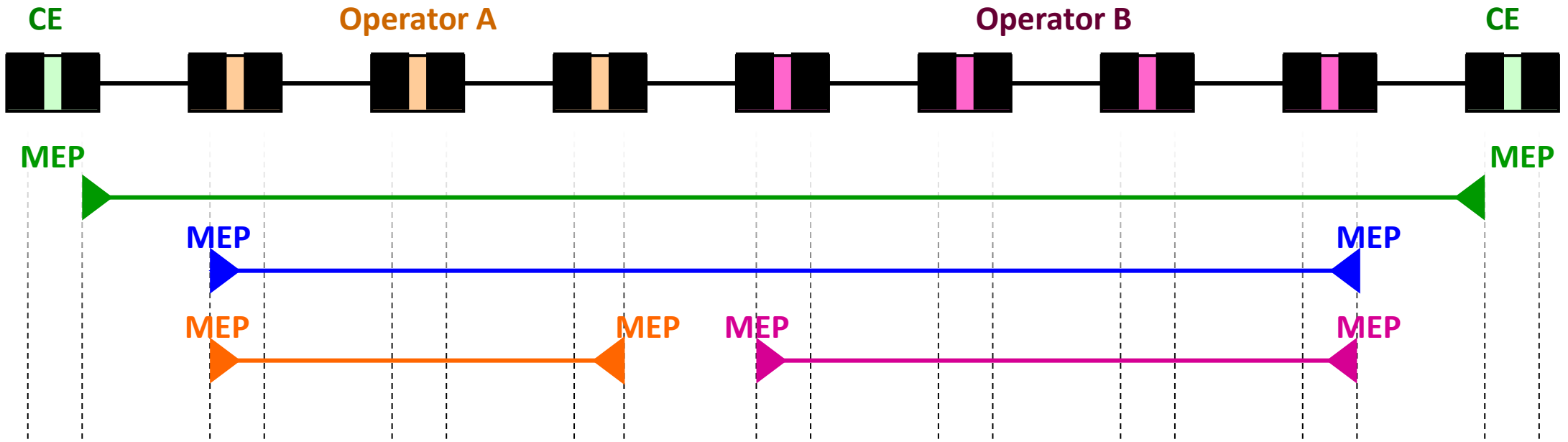
2 Using CFM Linktrace (IEEE 802.1ag)

Connectivity Fault Management (CFM) Overview

- **Family of protocols that provides capabilities to detect, verify, isolate and report end-to-end Ethernet connectivity faults**
- **Employs regular Ethernet frames that travel in-band with the customer traffic**
  - Devices that cannot interpret CFM Messages forward them as normal data frames
- **CFM frames are distinguishable by Ether-Type (0x8902) and two Multicast MAC addresses (for multicast messages)**
- **Standardized by IEEE in late 2007**
  - IEEE std. 802.1ag-2007

CFM Concepts

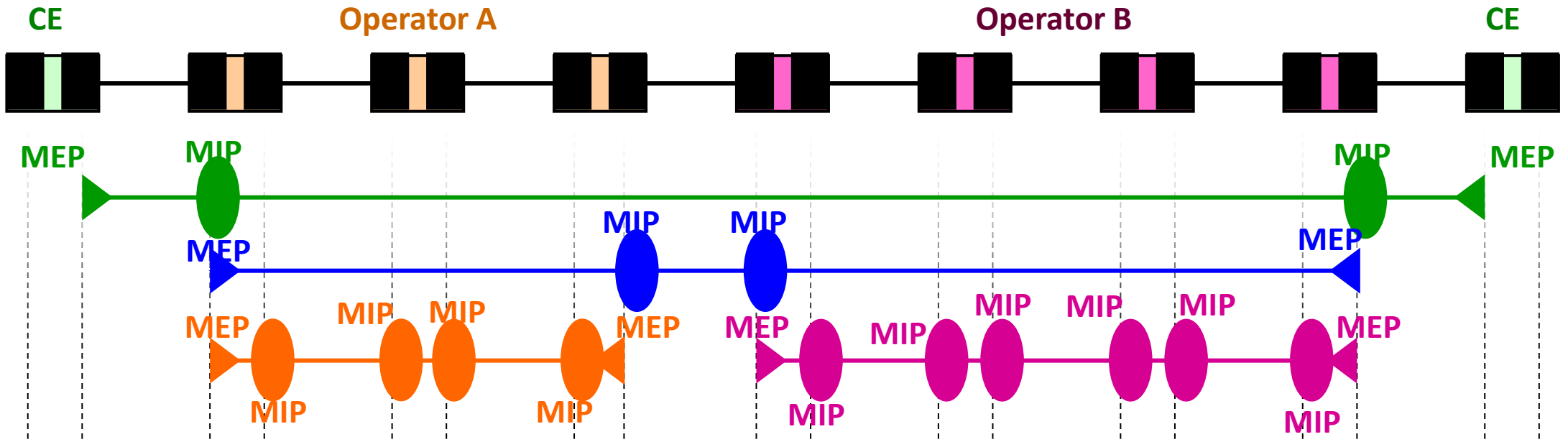
Maintenance Point (MP)—MEP



- Maintenance Association End Point (MEP)
- Define the boundaries of a Maintenance Domain
- Support the detection of connectivity failures between any pair of MEPs
- Can initiate and respond to CFM PDUs

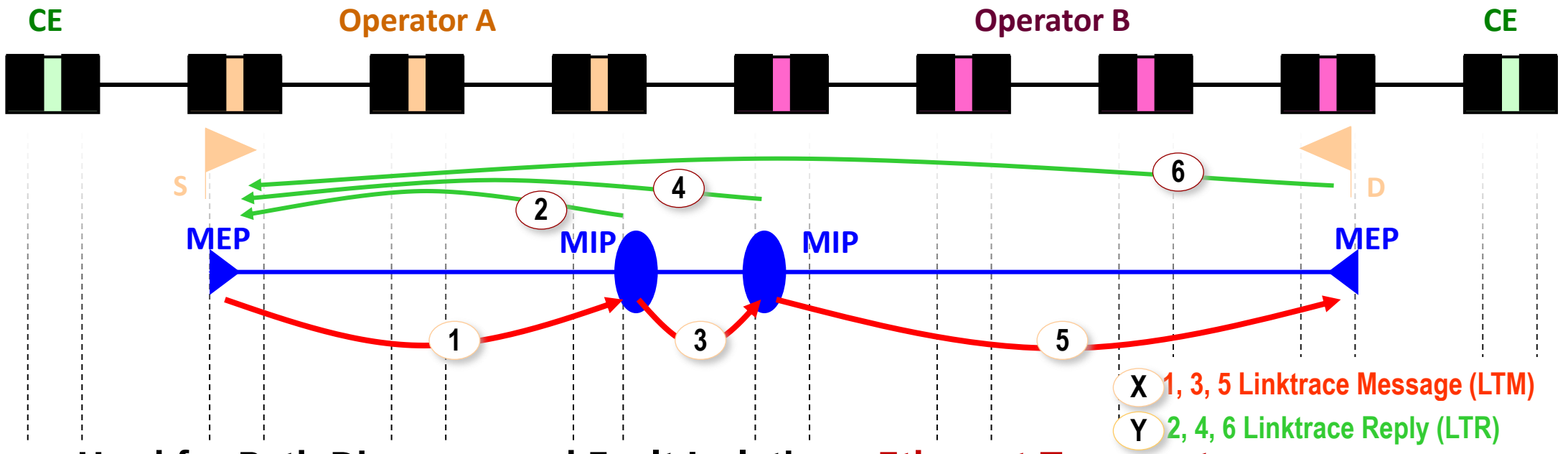
CFM Concepts

Maintenance Point (MP)—MIP



- **Maintenance Domain Intermediate Point (MIP)**
- **Support the discovery of paths among MEPs and location of faults along those paths**
- **Can be associated per MD and VLAN / EVC (manually or automatically created)**
- **Can add, check and respond to received CFM PDUs**

CFM Linktrace Protocol



- Used for Path Discovery and Fault Isolation—Ethernet Traceroute
- MEP can transmit a multicast message (LTM) in order to discover the MPs and path to a MIP or MEP in the same MA
- Each MIP along the path and the terminating MP return a unicast LTR to originating MEP.
- The LTR message can be enhanced to contain the MTU of the egress interface.

Linktrace MAC DA

- **LTM frames are generated with a multicast DA that both MEPs and MIPs listen to.**
- **A multicast DA is used instead of unicast DA for LTM frames since in current bridges the MIPs would not be able to intercept a frame with a unicast DA which was not their own address.**
  - The limitation is that current ports do not look at the EtherType before looking at the DA.
- **The MAC address of the LT destination MEP/MIP is in the TargetMAC field of the LTM.**
- **LTR frames are always generated with unicast Das.**

3 Using LLC TEST

LLC TEST

- **The TEST function provides a facility to conduct loopback tests of the LLC to LLC transmission path. The initiation of the TEST function may be caused by administration or management entity within the data link layer.**
  - Successful completion of the test consists of sending a TEST command PDU with a particular information field provided by this administration or management entity to the designated destination LLC address and receiving, in return, the identical information field in a TEST response PDU.
  - Implementation of the TEST command PDU is optional but every LLC must be able to respond to a received TEST command PDU with a TEST response PDU. The length of the information field is variable from 0 to the largest size specified that each LLC on this local area network must support for normal data transfer.

LLC TEST

- **It shall also be possible to send even larger information fields with the following interpretations. If the receiving LLC can successfully receive and return the larger information field, it will do so.**
  - If it cannot receive the entire information field but the MAC can detect a satisfactory FCS, the LLC shall discard the portion of the information field received, and may return a TEST response PDU with no information field.
  - If the MAC cannot properly compute the FCS for the overlength information fields, the LLC shall discard the portion of the information field received, and shall give no response.
  - Any TEST command PDU received in error shall be discarded and no response PDU sent. In the event of failure, it shall be the responsibility of the administration or management entity that initiated the TEST function to determine any future actions.

Using TEST Command as a Probe

- **The general strategy is to find an appropriate Path MTU by probing the path with progressively larger frames until it reaches the first-hop MTU.**
  - A simple strategy might be to do a binary search halving the probe size range with each probe.
  - Raise the probe size in smaller increments
  - It may be appropriate to probe at certain common or expected MTU sizes (1518, 1522, 2000).
- **The presence of other losses near the loss of the probe may indicate that the probe was lost due to congestion or path problems rather than due to an MTU limitation.**
  - At this point, it is particularly appropriate to re-probe
- **The idea is that the isolated loss of a probe frame is treated as an indication of an MTU limit, and not as a congestion or path problem indicator.**

Thank You