Metro Ethernet Connection Management

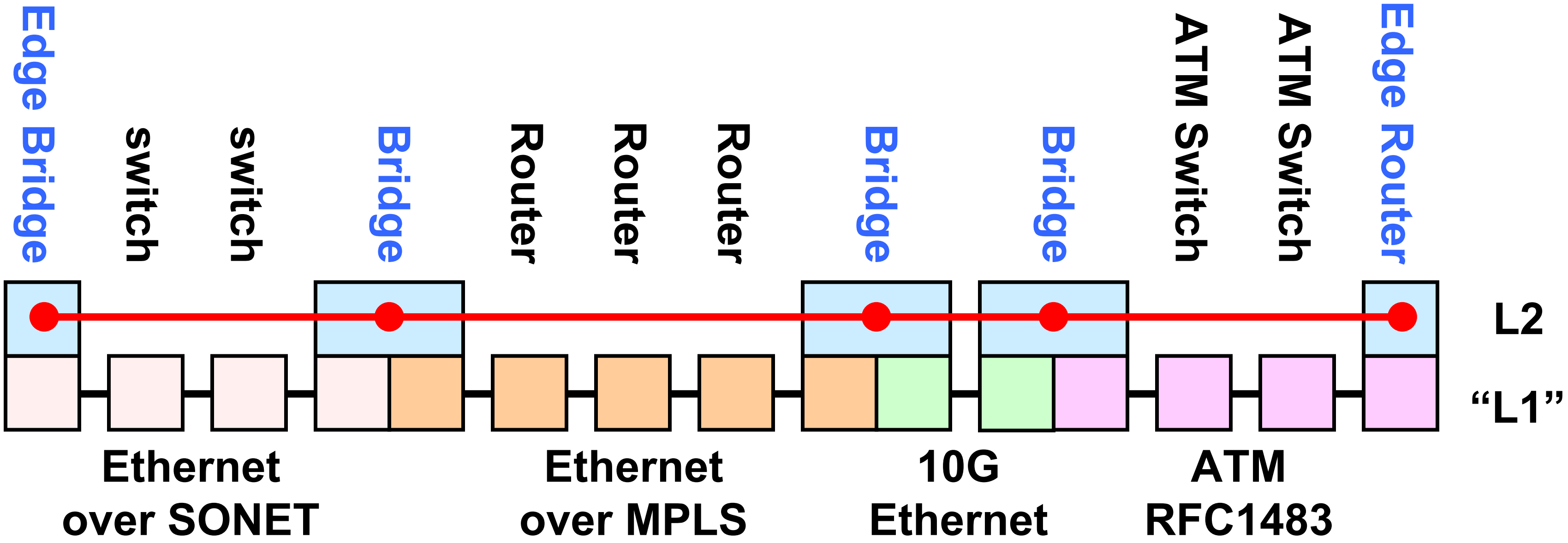
Submitted to IEEE 802.1 Interim meeting Vancouver, BC, January 12-15, 2004

Norman Finn

Cisco.com

Overview

What is Metro Ethernet Connection Management?



- **CM: Standard Ethernet frames, distinguished from ordinary data frames only by destination MAC address and/or EtherType, and seen, relayed, and/or terminated by Provider Bridges.**

Why do Service Providers want Connection Management?

- Debugging Layer 2 networks is not easy in the Enterprise space.
- Debugging networks of Layer 2 networks promises to be even harder.
- Especially when the component networks belong to different organizations and/or different companies.
- Experience in providing circuit-based services provides a very useful model for managing and debugging nets of nets.

What is Ethernet Connection Management?

- Connection Management is “E2E OAM”.
- Metro E2E OAM is Ethernet frames, not MPLS, ATM, or SONET frames, cells, or sub-Ethernet control information.
- Only bridges see OAM because bridges are the only active relay functions that exist.
- MPLS, ATM, and SONET OAM are important for debugging Ethernet “wires”, but are not always end-to-end.

(Terminology issue) There is no such thing as a "Metro Ethernet Edge Router"

- If interoperability is to be achieved among the various technologies used for Metro Ethernet, the IEEE 802 LAN architecture must be observed.
- There are only two kinds of active relay elements in IEEE 802: Bridges and Hubs (Repeaters).
- A "Metro Ethernet Edge Router" is a Bridge that does not need to run spanning tree, has only two ports per VLAN enabled (one trunk and one local Ethernet), and uses Pseudowires for trunks.
- (This notion saves a lot of verbiage in this document. Whether a box is a router or a bridge depends on which function you are looking at.)

Maintenance and other OAM issues not discussed

- Minimal discussion of Provider-to-Customer (single-link) IEEE 802.3ah OAM.
- No explanation of MPLS, ATM, or other OAM.
- No explanation of other techniques such as periodic confirmation of network topology and configuration, SNMP-based “traceroute”, or Layer 3 functions such as Ping.
- No explanation of other protocols such as Ethernet Line Management Interface, BPDUs, etc.

Two standards bodies are defining end-to-end OAM

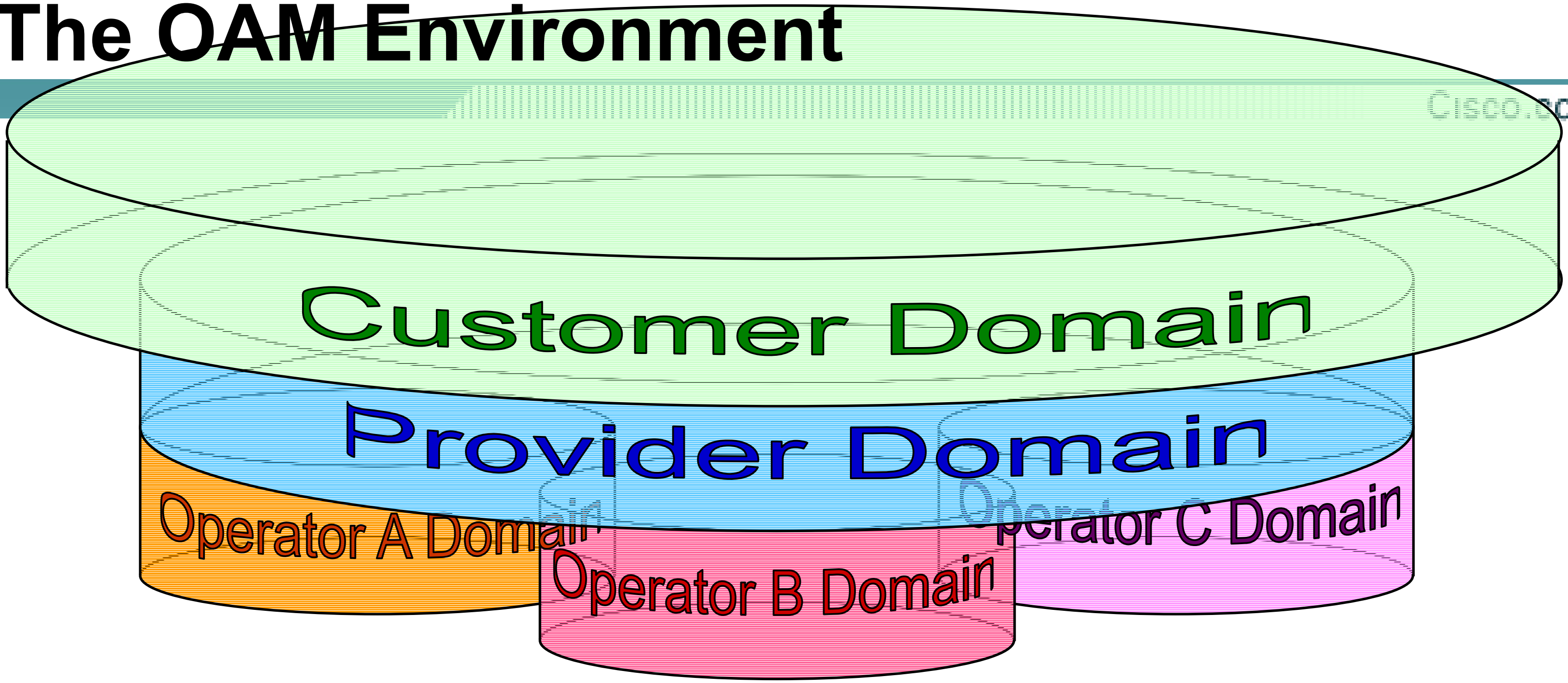
- IEEE 802.1ad defines Q-in-Q “Provider Bridges”. IEEE 802.1ag defines end-to-end Ethernet OAM for those Bridges.
- ITU-T Question 3, Study Group 13, is defining end-to-end Ethernet OAM for both circuit-switched equipment (e.g. Ether-over-SONET) and packet-switched equipment (802.1ad Bridges).
- Both 802.1 and Q.3/13 share common membership and are cooperating fully.

  802.1 should define low-level aspects tied closely to bridging technology.

  Q.3/13 should define high-level aspects tied to the service models.

Domains,
Maintenance Levels,
and
Flow Points

The OAM Environment



- Customer contracts with Provider for end-to-end service.
- Provider contracts with Operator(s) to provide equipment and networks.
- Provider and Operator(s) may or may not be the same company or same division.

OAM Flow Points and Domains

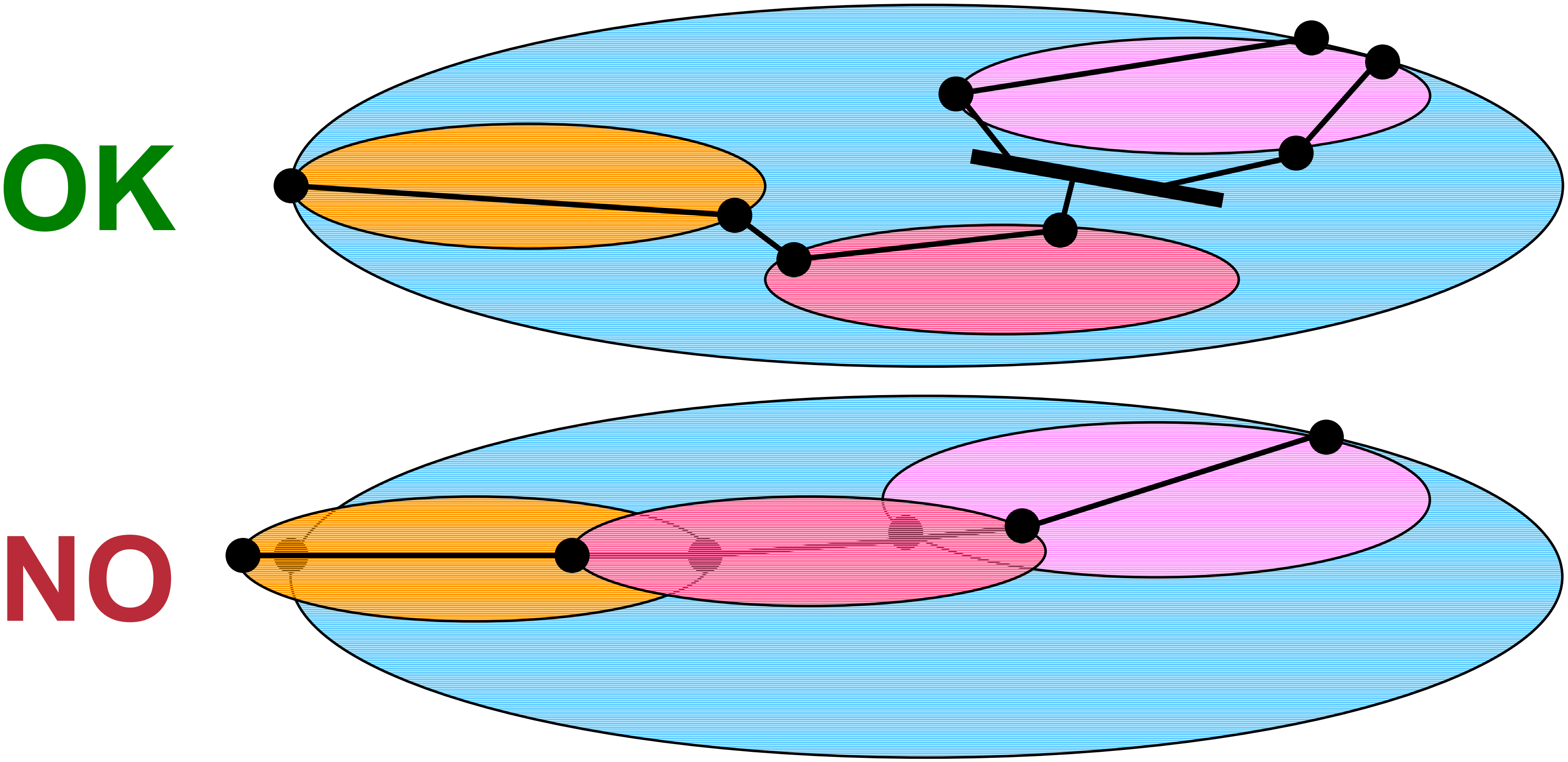
- Domains are defined in terms of "flow points", which are "MACs" to IEEE 802, and "interfaces" or "ports" to others.
- A flow point (FP) at the edge of a Domain is called a "Maintenance Point" or MP.

  System administrators use MPs to initiate and monitor OAM activity and report the results.

- A flow point inside a Domain, and visible to an MP, is called a "Loopback Point".

  Loopback Points (LPs) passively receive and respond to OAM packets initiated by MPs.

OAM Domains

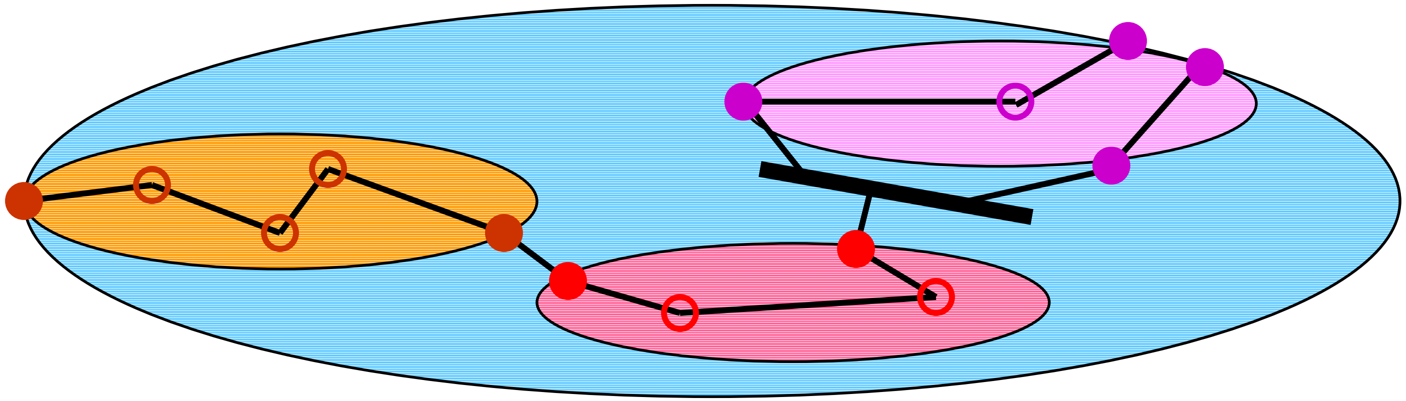


OK

NO

- **OAM Domains may nest or touch, but must never intersect.**

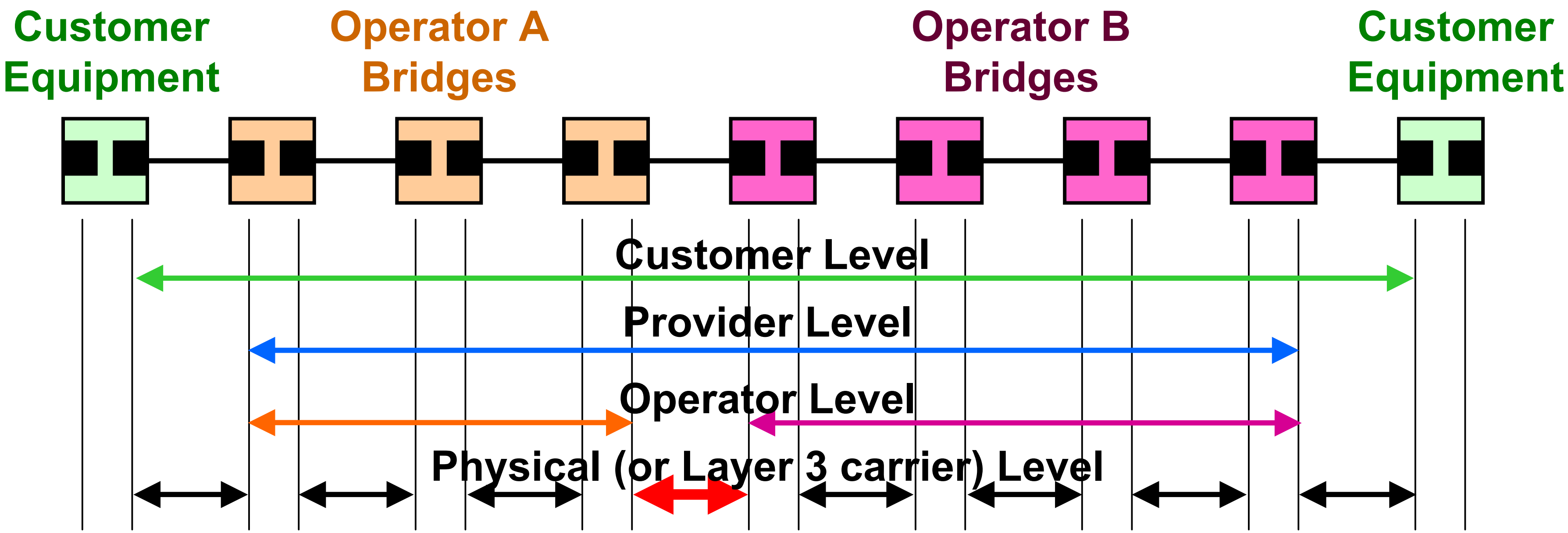
OAM Domains



● ● ● Maintenance Points ○ ○ ○ Loopback Points

- Maintenance Points are always at the **edges** of Domains.
- Loopback Points are always **within** domains.

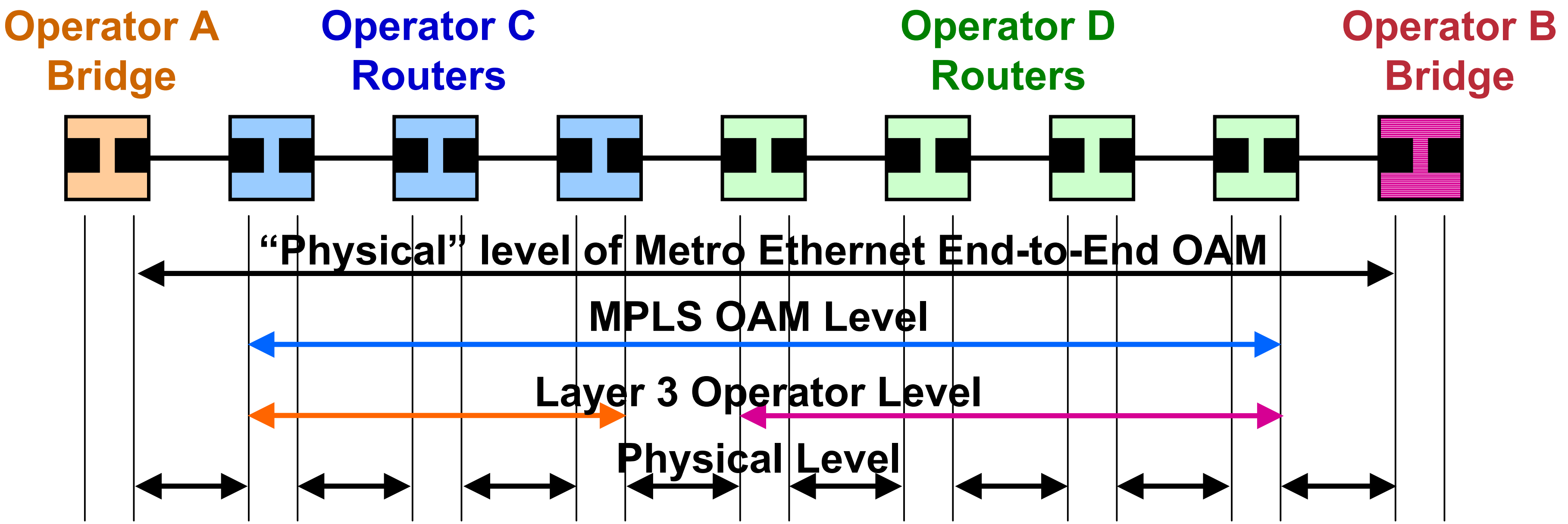
OAM Maintenance Levels



- At (at least) Operator and Physical levels, there may be multiple Domains.

There *could* be multiple Domains at any level.

Lower OAM Levels (Expanding red link in the previous slide)

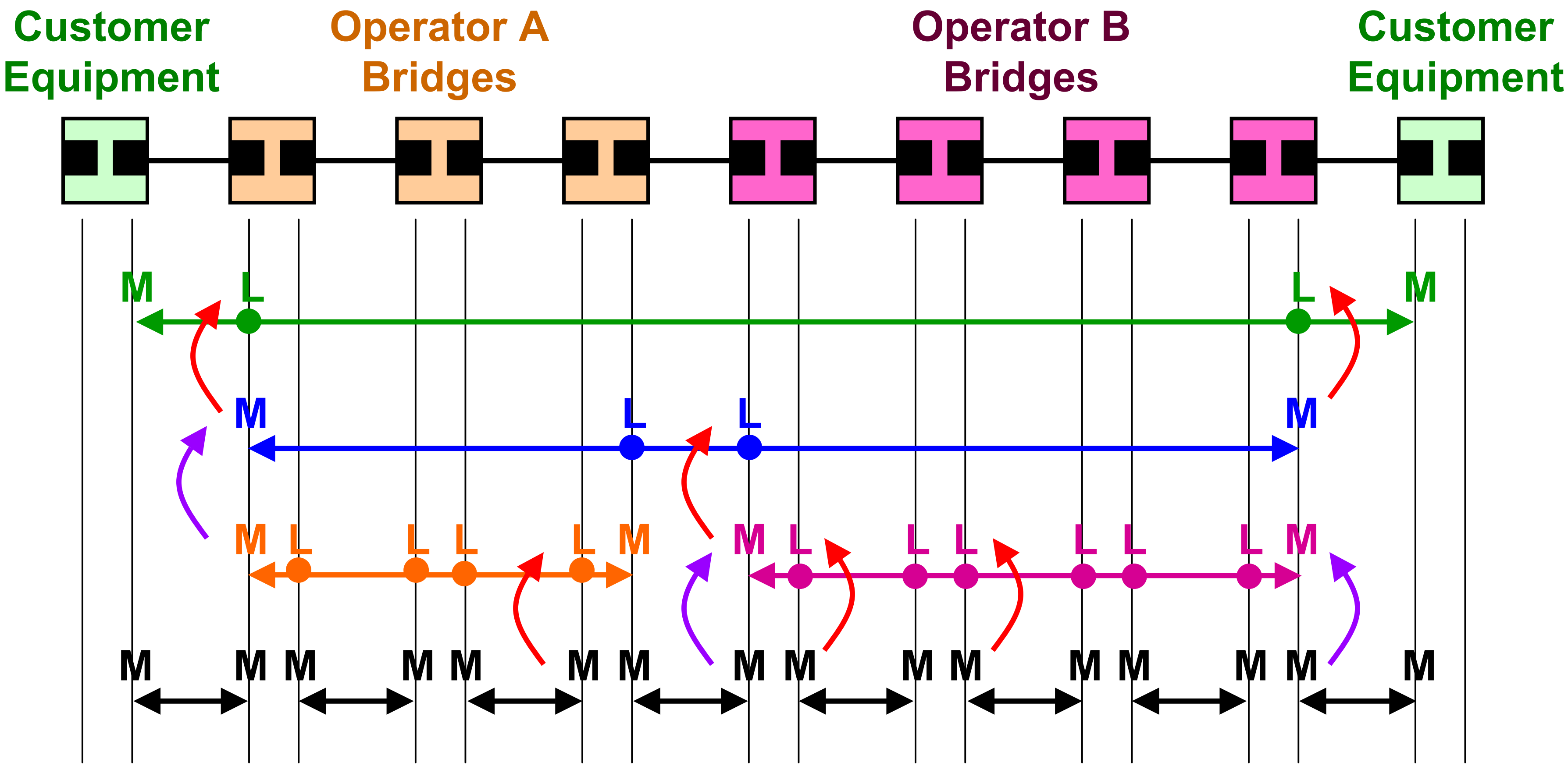


- There are levels below End-to-End OAM.

  This is where MPLS OAM, ATM OAM, etc., live.

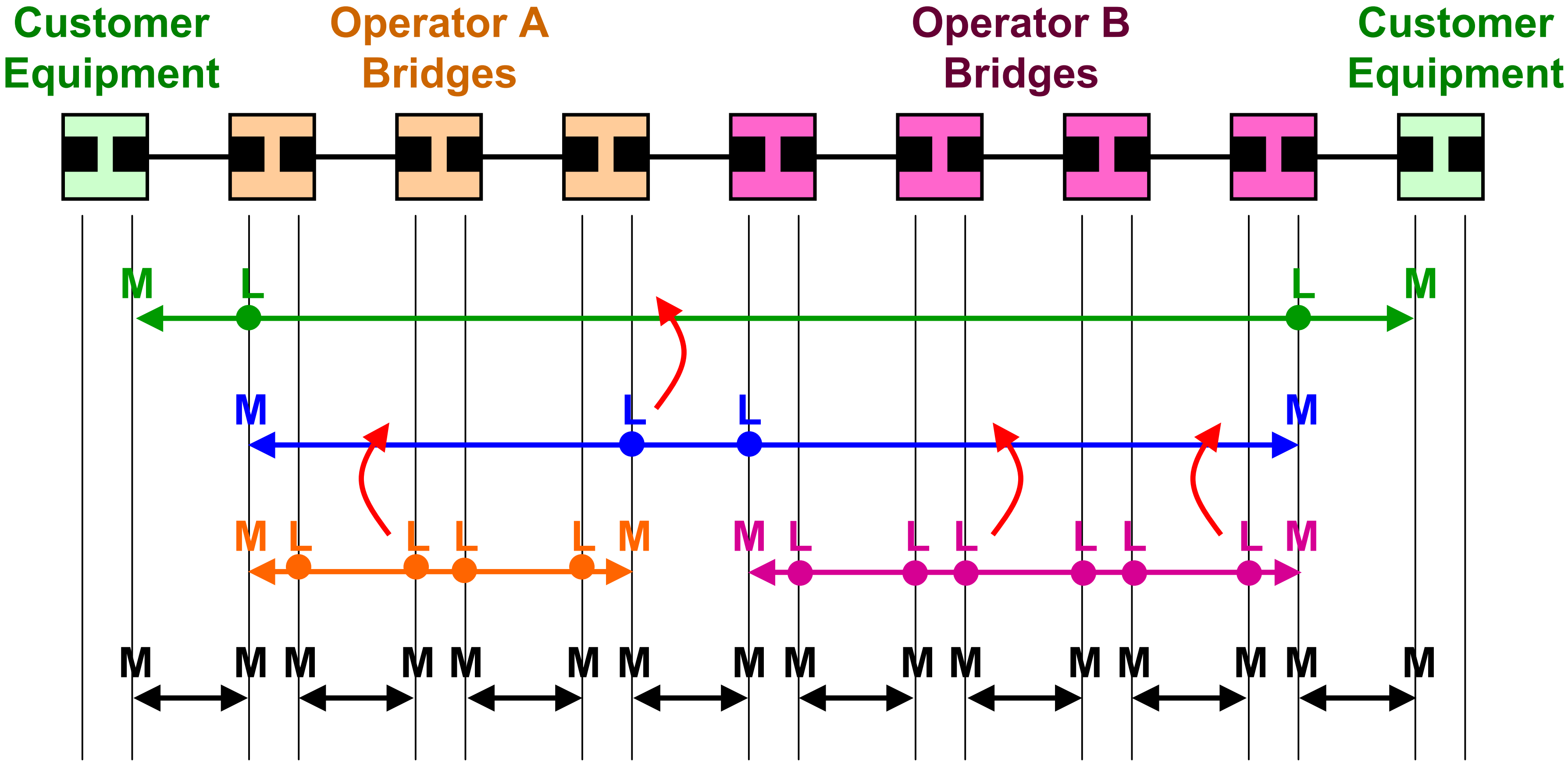
  These OAM functions *ought to*, but do not *necessarily,* follow the MP/LP/FP model.

Loopback Points, Maintenance Points



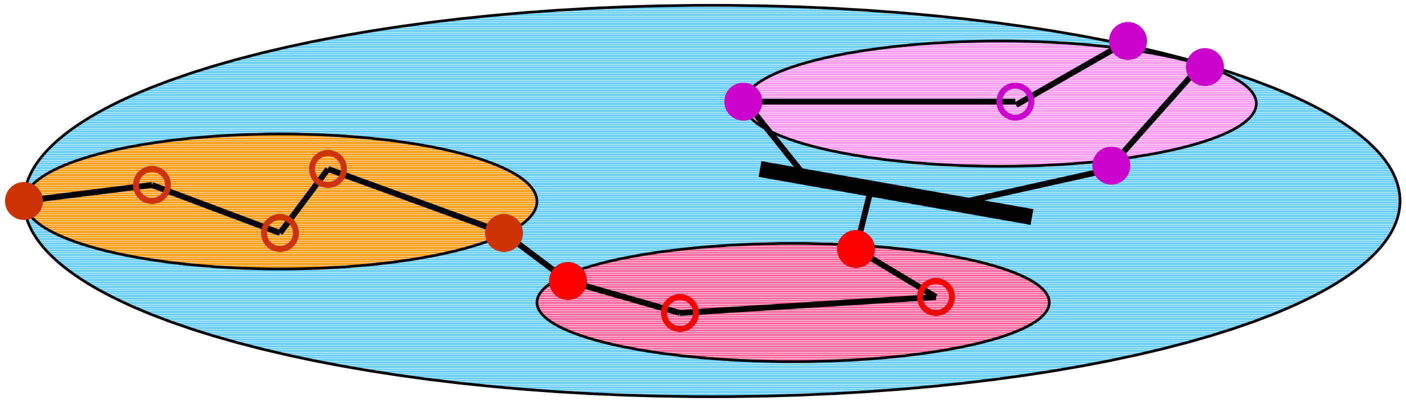
- Each Level’s Maintenance Points are the next-higher-level’s Maintenance or Loopback Points.

Loopback Points, Maintenance Points



- Each Level's Loopback Points are **invisible** to all higher levels.

Loopback Points and Maintenance Points: Operator Levels (x3)

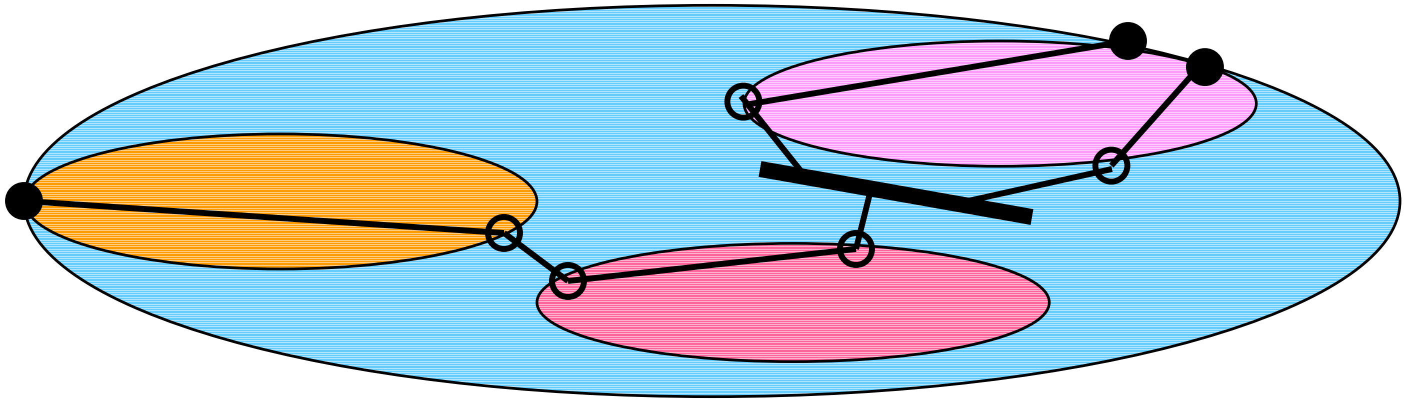


● ● ● Maintenance Points ○ ○ ○ Loopback Points

- Each Level's Maintenance Points are the next-higher-level's Maintenance or Loopback Points.

- Each Level's Loopback Points are invisible to all higher levels.

Loopback Points and Maintenance Points: Provider Level



● Maintenance Points ○ Loopback Points

- Each Level’s Maintenance Points are the next-higher-level’s Maintenance or Loopback Points.
- Each Level’s Loopback Points are invisible to all higher levels.

End-to-End OAM Packet Types

Three Kinds of OAM Packets

- Continuity Check

  Multicast from MP. Received by MPs and LPs. Catalogued by receiving MPs.

- Traceroute

  Next-hop Multicast from MP to next MP or LP along route. Receiver both replies with unicast to original MP, and sends Traceroute to next MP/LP.

- Loopback (Ping)

  Unicast from MP to LP or MP, which replies with unicast to originating MP.

All OAM Packets must include

- Destination MAC addr (uni- or multicast).
- Source MAC addr of sending Flow Point.
- Customer Service Instance (CSI) indicator, typically an S-VLAN ID in an 802.1ad tag.
- EtherType, Version, and Op-code.
- Maintenance Level indicator (an integer).
- Transaction Identifier (an integer).
- Universally unique CSI identifier string (UUCSIID).

All OAM Packets may include

- **Arbitrary amount of arbitrary data.**

  Data MUST NOT be interpreted by replying FP.

  Supports testing frame size limitations and testing of Service Level Agreements.

- **Internal Checksum computed on all of frame except .1Q/.1ad tags, which are subject to alteration.**

  This may detect bad FCS hardware.

Continuity Check: Transmission

- **Transmitted periodically from each Maintenance Point on each CSI (S-VLAN).**
- **Timing depends on administrators' needs; perhaps 1 second intervals, perhaps 60.**
- **Contains the basic info, plus (at least) an MP Identifier (MPID) unique over CSI, a "validity time" of 2.5 or 3.5 times the transmission period, and port state info.**
- **There is one standard reserved Continuity Check multicast destination MAC address defined for *each* Maintenance Level.**

Continuity Check: Transmission

- Separate CCs are used at every level, from each MP (one bridge may have multiple MPs for same CSI), for each CSI.
- Transaction ID is initialized to random value, incremented once per transmission from a given MP, CSI, and Level.
- Failure detection speed must be balanced against CC loss rate and overhead.

  E.g. Validity time of 2.5 transmission interval allows loss of 1 CC.

Continuity Check: Reception

- Received in each Maintenance Point at the Maintenance Level and on the CSI of the sender.
- No reply is made to a Continuity Check.
- Received information is saved in a table indexed by MPID for use by Traceroute or Ethernet Line Management Interface (ELMI).
- If validity time of a table entry expires, saved information is discarded.

Continuity Check: Errors Detected

- Receiver may compare list of saved MPIDs to a configured list, and complain of mismatches.
- A CC received with a UUCSIID / S-VLAN mismatch indicates cross-connected CSIs.
- Duplicate or out-of-order Transaction IDs may indicate a forwarding loop in Provider space.
- Port state is communicated to all MPs for relay to the customer via ELMI.

Continuity Check: Loopback Points

- **Loopback Points also receive Continuity Checks.**
- **Unlike MPs, LPs *need not* log CCs by source MP.**
- **LPs remember only the source MAC address, Maintenance Level, S-VLAN, and ingress bridge port in a (software?) CC database separate from the bridge's main Filtering Database (L2 forwarding table).**
- **This information is used by the Traceroute function.**

Traceroute Message Initiation

- Originator MP initiates Traceroute Message (TM). Included in packet are:

  Target MAC address, always a Maintenance Point

  Originator's MAC address

- TM also includes a Hop Count TLV.

- There is one standard reserved Traceroute multicast destination MAC address defined for each Maintenance Level.

  (For tracing at higher Maintenance Levels.)

Traceroute Message forwarding

- All MPs and LPs at a given level intercept all TMs at that level, using destination MAC address.
- Each MP or LP intercepting a TM:

1. If TTL is not 0, decrements TTL and:
2. Sends unicast Traceroute Reply (TR) to Originator MP.
3. If TTL still not 0, forwards the TM to next hop along path to target MAC address.

Traceroute Message
Next Hop determination

- The bridge looks up the target MAC address in the Filtering Database to find the egress FP.
- If not found, the Continuity Check database is examined.
- If the egress FP matches the ingress FP, then no Traceroute Reply is returned.

  (This prevents multiple answers after hopping over a shared medium.)
- Else, the next hop cannot be found.

Traceroute Reply

- Traceroute Reply always contains the TTL value, after decrementing, of the intercepted Traceroute Message, the Transaction ID, the device ID of the replying bridge that owns the FP, and the replying bridge's Layer 3 management address, all in 802.1ab LLDP format.

Obsolete Slide!

- In theory, TR comes from a Flow Point, not a Bridge's brain. Therefore, a Bridge may issue either two TRs for the ingress and egress FPs, or one TR for both.

Traceroute Reply Contents:
Flow Point at ingress to Bridge

- Ingress Flow Point ID and Type and Egress Flow Point ID, in LLDP format.
- Reason for not forwarding TM, either:

  Ingress FP’s STP/VLAN state is not Forwarding;
  Ingress FP’s ifOperState is not UP;
  Data frame would be filtered on ingress;
  Traceroute Message’s TTL expired;
  Data frame would be filtered by relay function; or
  No egress FP could be identified.

Obsolete Slide!

- If forwarded, method for finding Egress FP, either:

  Egress FP from Filtering Database; or
  Egress FP from Continuity Check database.

Traceroute Reply Contents:
Flow Point at egress from Bridge

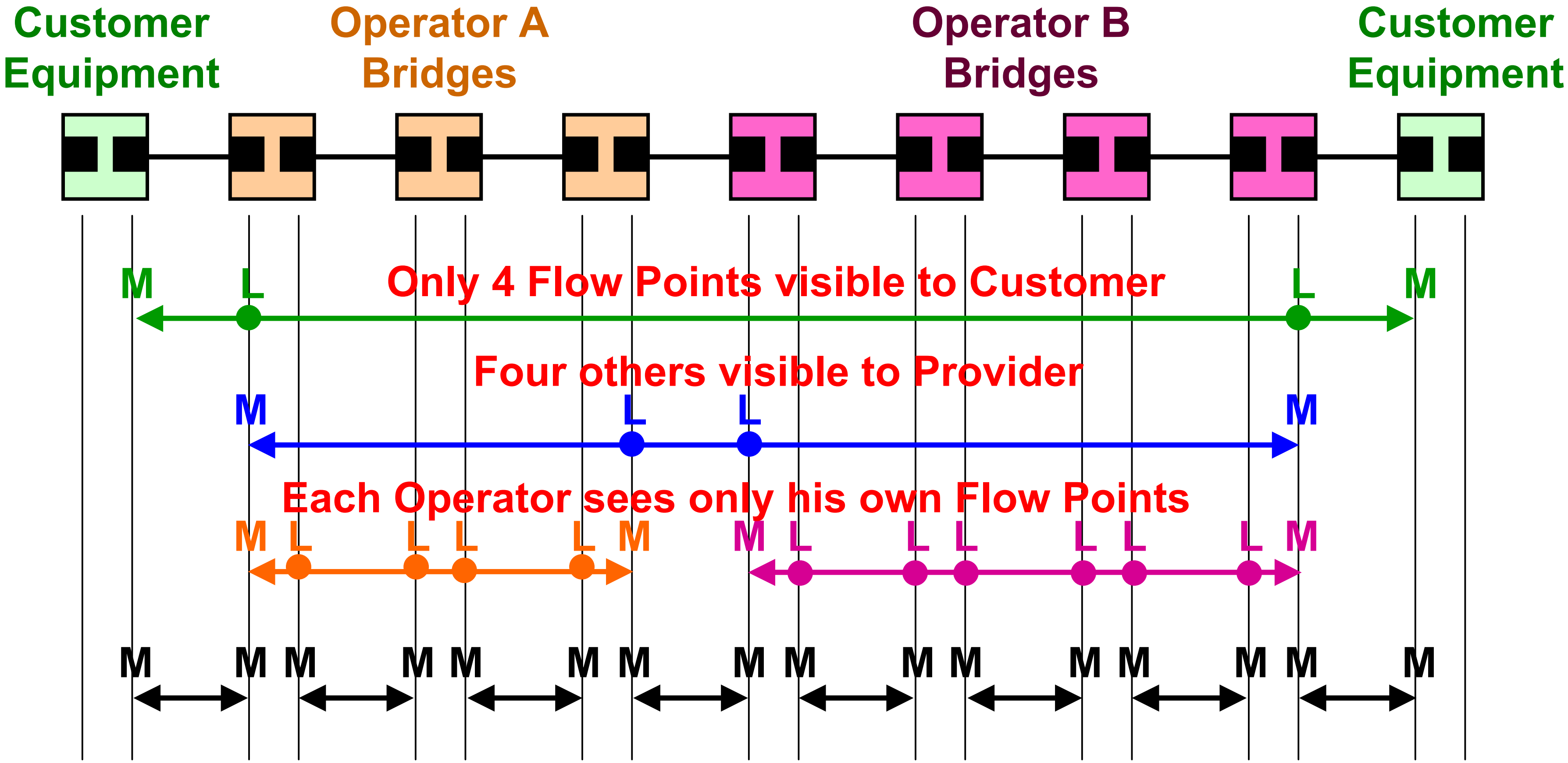
- Egress Flow Point ID and Type in LLDP format.
- Device and Flow Point IDs of next hop bridge (or other device), if unambiguous, in LLDP format.
- Reason for not forwarding TM (if needed), either:

TM was forwarded; OR

Egress FP's ifOperState is not UP;
Egress FP/VLAN STP state is not Forwarding;
Egress FP is blocked by GVRP;
Data frame would be filtered on egress; or
Traceroute Message's TTL expired.

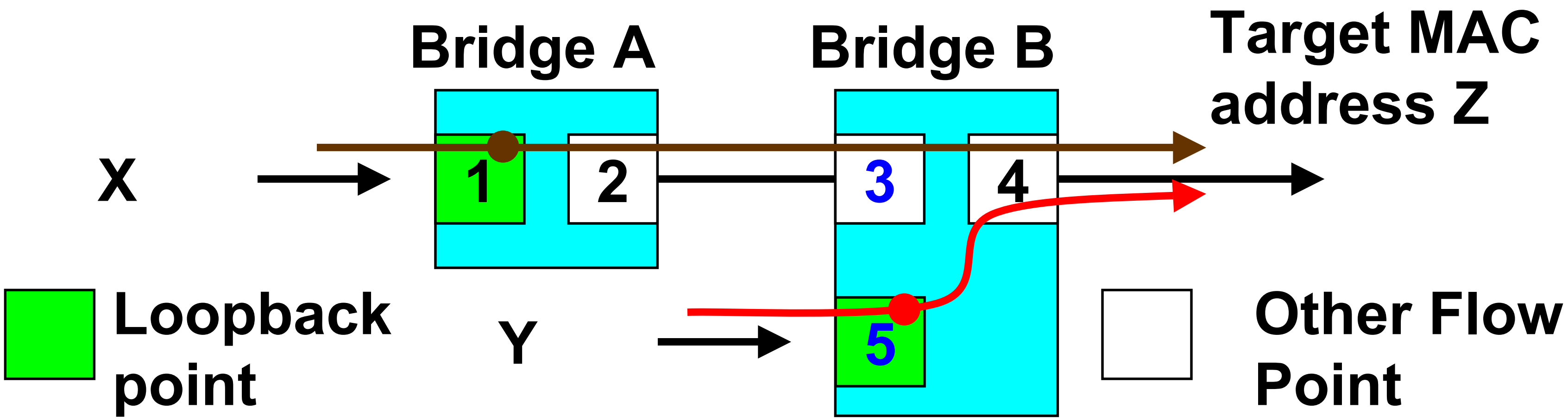
Obsolete Slide!

Traceroutes at different Maintenance Levels see different pictures


Customer Equipment
Operator A Bridges
Operator B Bridges
Customer Equipment
M
L
Only 4 Flow Points visible to Customer
L
M
Four others visible to Provider
M
L
L
M
Each Operator sees only his own Flow Points
M L L L L M
M L L L L L L M
M M M M M M M M M M M M M M M M

- **Each Level’s Maintenance Points see only that level’s Loopback Points.**

Problem: Traceroute at Domain Edge



Bridge A

Bridge B

Target MAC address Z

X

1

2

3

4

Y

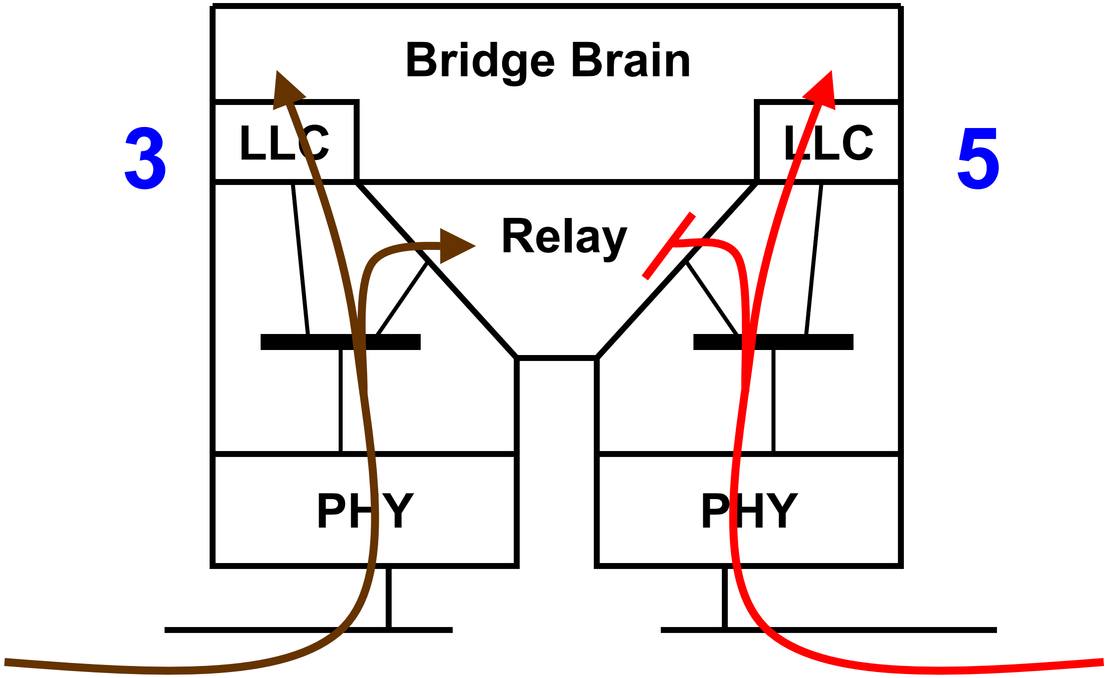
5

🟩 Loopback point

□ Other Flow Point

- Port 3 must pass Traceroute at Level *n* through without stopping.
- Port 5 must stop Traceroute at Level *n* and reply to it.

Traceroute at Domain Edge



- No way, so far, to pass both messages to LLC, but for Relay Function to relay only one.
- So Relay must block both, and software must forward.

Loopback Message and Loopback Reply

- Always sent by an MP *only* to a unicast MAC address of another MP or LP.

  Address is obtained from a previous Continuity Check or Traceroute operation.

- Receiving MP or LP responds with a Loopback Reply, which has a different Op-code.

  All optional fields are returned in the Loopback Reply.

- Transaction ID allows originating MP to match Loopback Messages to Replies.

Controlling OAM

A Layer 3 overlay is required

- At least one “Customer Service Instance” will be used by the Provider(s) and/or Operator(s) to provider Layer 3 services for managing the bridges.
- This may be a private network, or may be attached to the Big-I Internet.
- Routing functions may be embedded in the Provider Bridges, or may be separate devices.

Different organizations may control the same Bridge

- One Operator might provide two UNI ports for two Customers of two different Providers on the same Provider Bridge.
- Although the Operator has the primary responsibility of configuring the Bridge, each Provider requires control of the OAM functions at the Provider Level on those Ports assigned to that Provider as Maintenance Points.
- SNMPv3 is typically the best way to allow a Provider to access OAM controls.

Continuity Check Management

- Controls:
  - Transmission period and validity time
  - Optional TLVs (e.g. interior checksum)
  - Configured list of peer MPs
  - Layer 3 address and options for error reporting
- Reports:
  - Maintenance Level (read-only for CC operator, read/write for Bridge owner/operator)
  - Table of discovered CC information
  - Error conditions

Traceroute Management

- **Controls:**

  Initiate Traceroute

  Select MP (MAC address) to trace

  TTL values

  Layer 3 address and options for error reporting

- **Reports:**

  Traceroute Replies

  Error conditions

Loopback Management

- Controls:

  Initiate Loopback sequence

  Select FP or MP (MAC address) to ping

  Optional TLVs (e.g. arbitrary data)

  Layer 3 address and options for error reporting

- Reports:

  Successful Loopback Replies

  Error conditions

Usage Scenarios

Continuity Check is the backbone of OAM

- Continuity Check enables the Ethernet Line Management Interface (ELMI) to report the status of the other UNIs of the service to the Customer.

- Continuity Check also allows MPs to catalog each other, and to associate a MAC address with each other MP.

- MAC addresses of other devices must be discovered via CC, not configured.

  Line cards may be swapped. spanning tree paths may change, and MP MAC addresses may be addresses of Ports, and not Brains.

Traceroute discovers the Loopback Points

- If CC indicates loss of another MP, Traceroute Message is sent to trace the path from the MP detecting the failure to the saved MAC address of the failed MP.

  This may yield the reason for the loss of connectivity.

  This will also yield a list of Loopback Points' MAC addresses to use in Loopback Messages.

- Traceroute may also be used to obtain the list of Loopback Points' MAC addresses when CC is *not* indicating any error.

Loopback allows detailed checking of paths

- Loopback Messages to successive Loopback Points along the path to another Maintenance Point may detect failures not found by CC or Traceroute.

  TM is hop-by-hop, and CC is a multicast. LM/LR are unicast.

  LM/LR can be used to detect links with too-small MTUs.

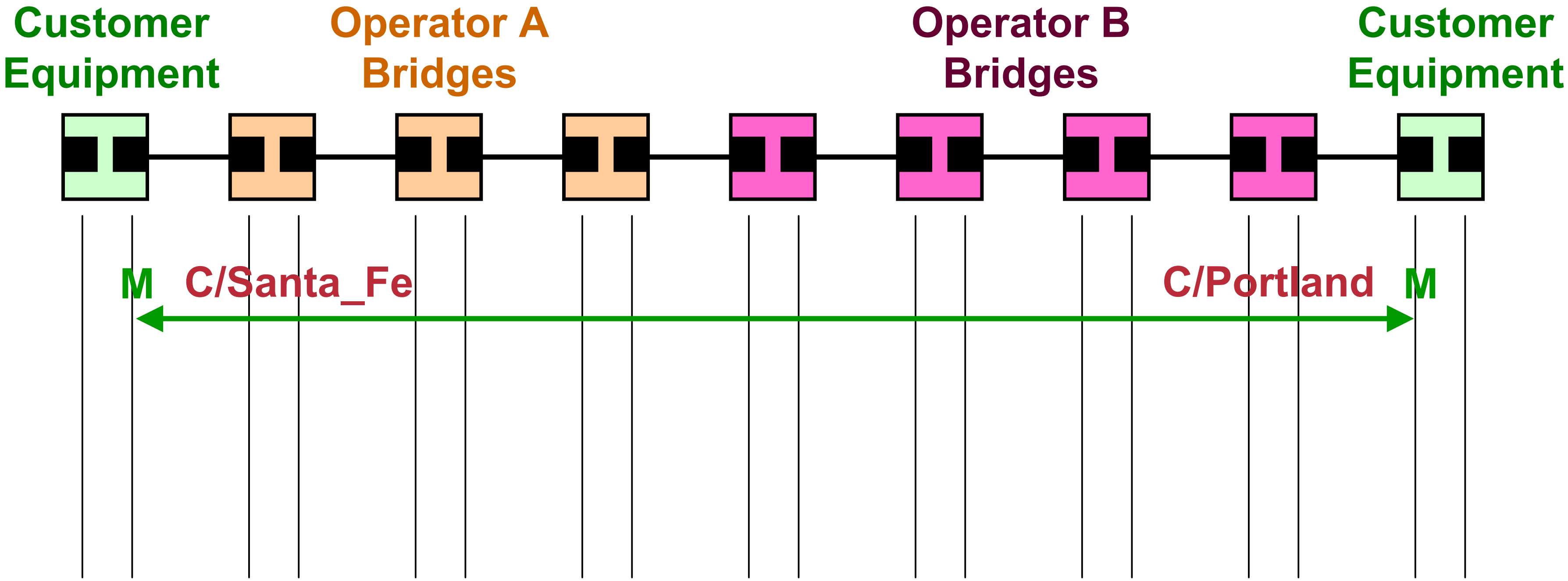
- Loopback may also be used for testing Service Level Agreement compliance: throughput, packet loss rate, etc.

Drilling Down

- If the failure can be isolated to a particular link using the above methods, then that link can be debugged.
- If the failing link is not a simple physical link, e.g. it is an Emulated LAN using Pseudowires over MPLS, then the OAM capabilities of that link must be used to debug that link.

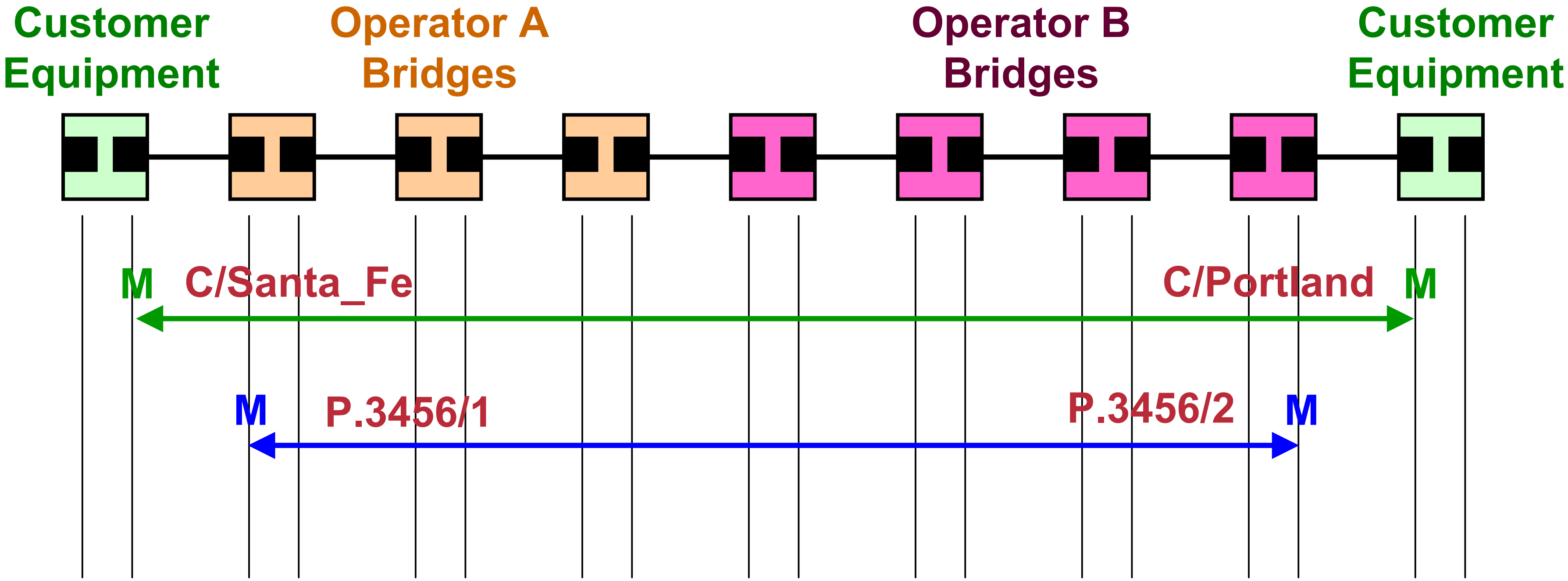
Configuration Information Flow

1: Customer MP Selection



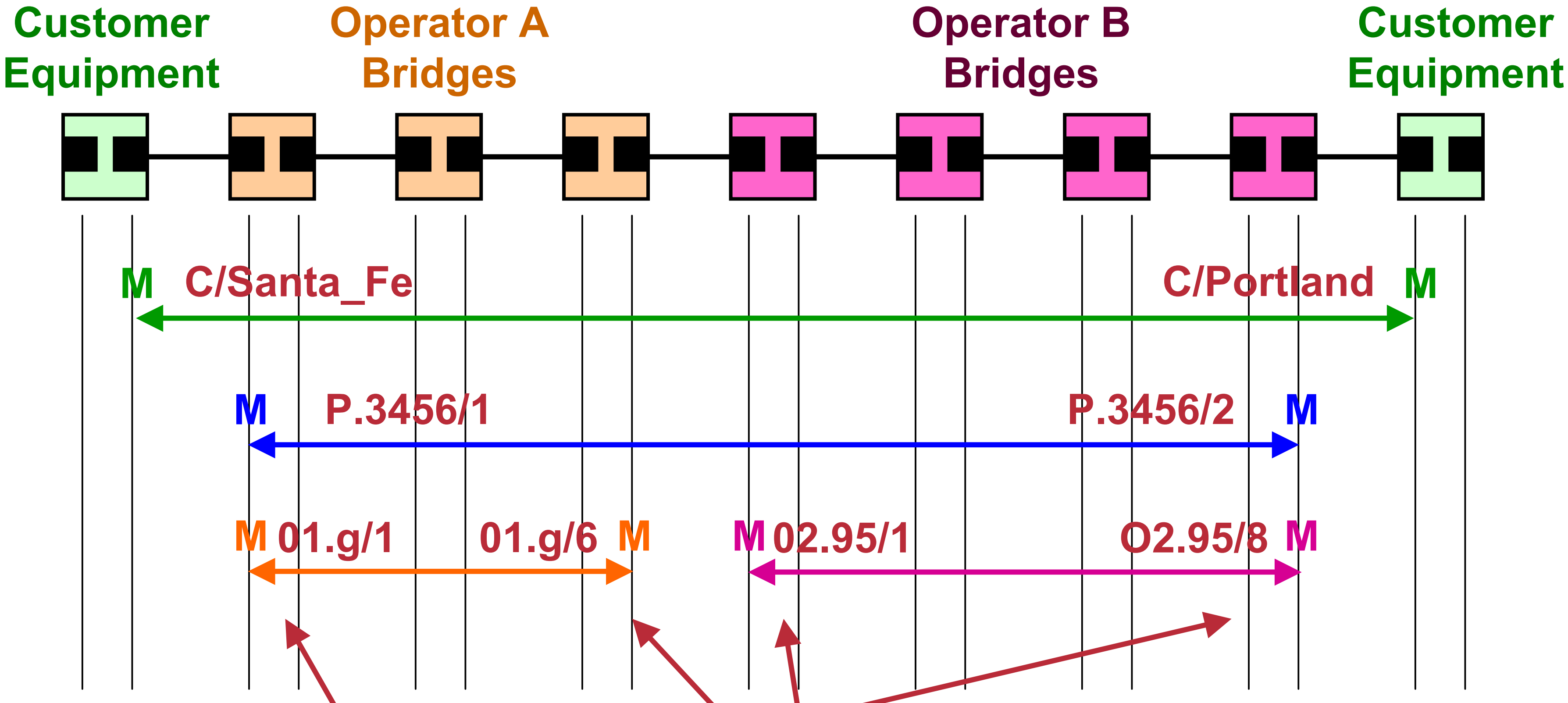
- Customer selects UUCSIIDs / MPIDs for his UNI-C Maintenance Points.
- Customer contracts with Provider

2: Provider MP Selection



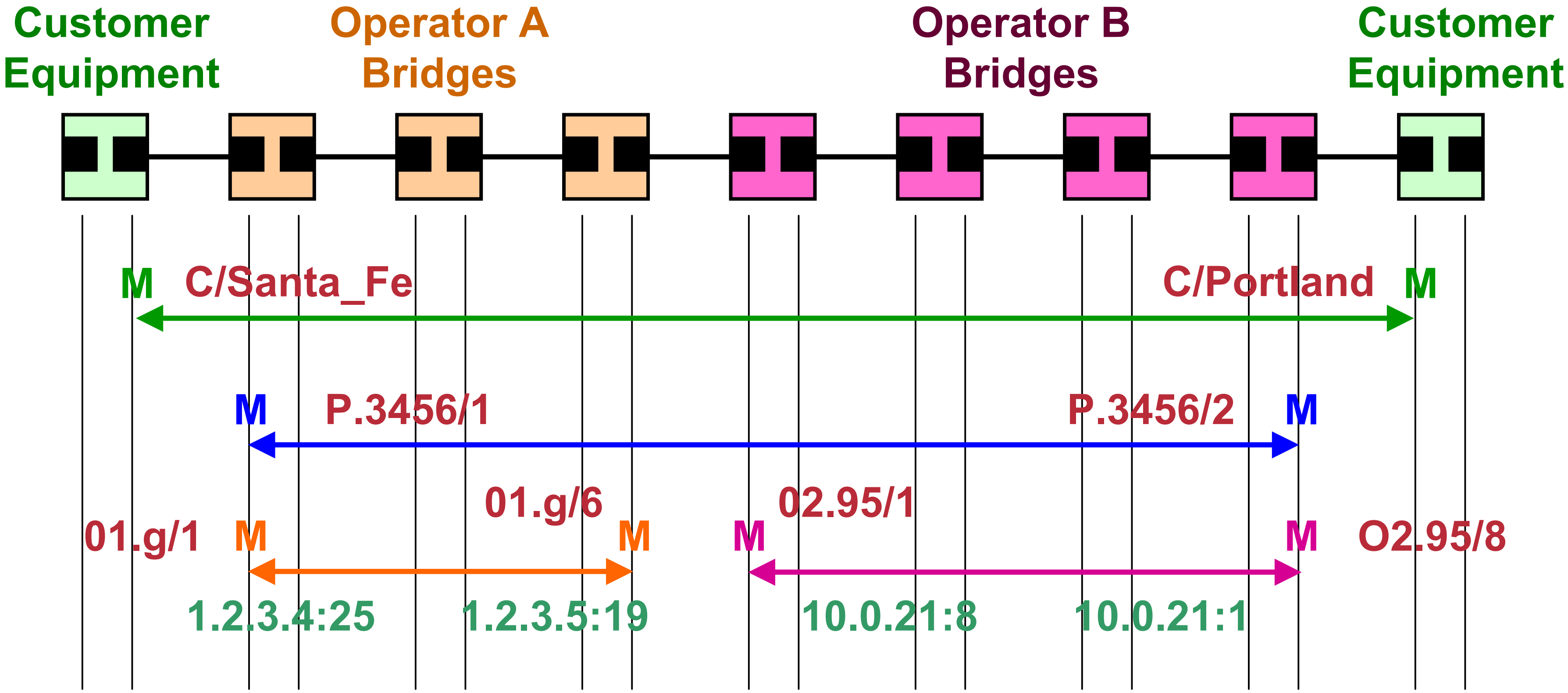
- Provider selects UUCSIIDs / MPIDs for her UNI-N Maintenance Points.
- Provider contracts with Operator(s)

3: Operator MP Selection



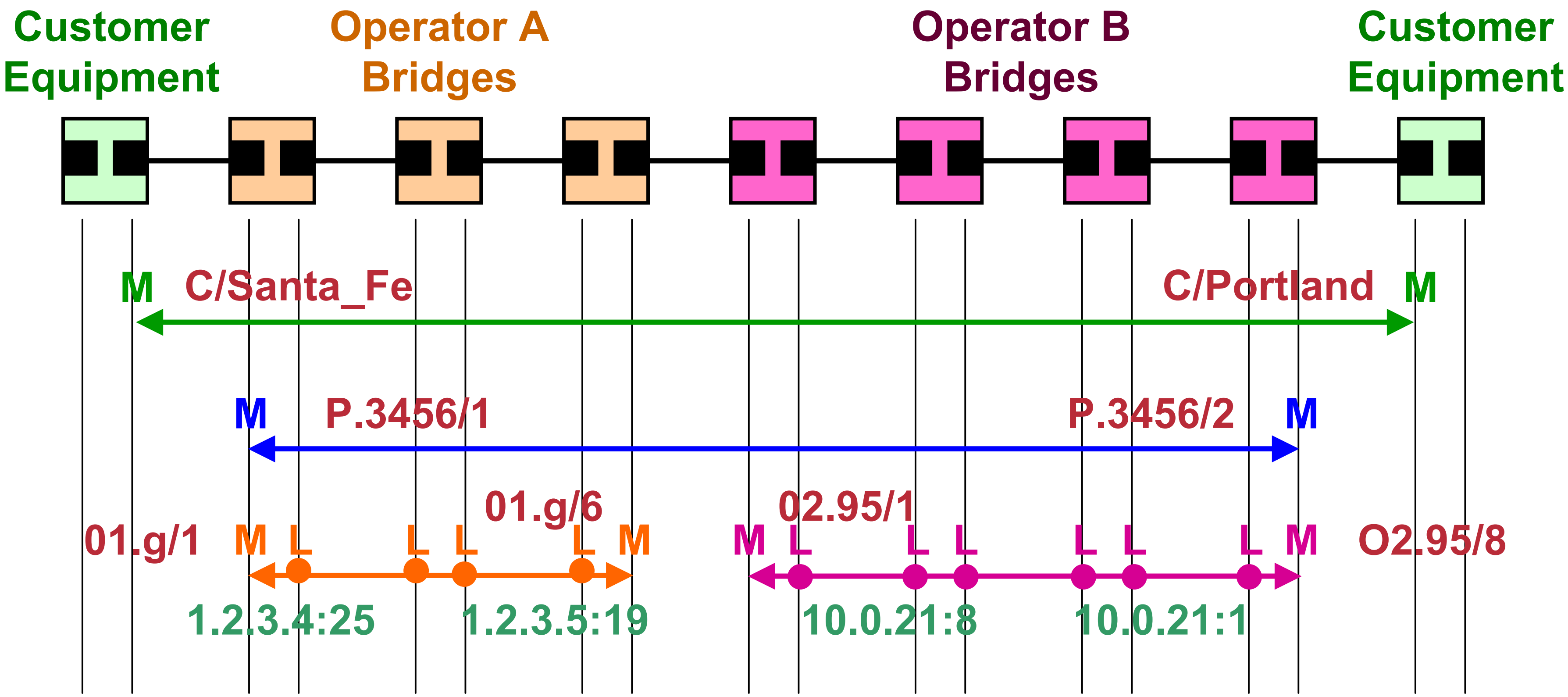
- Operators select UUCSIIDs / MPIDs for their UNI-N and NNI MPs.

4: MP / Interface Binding



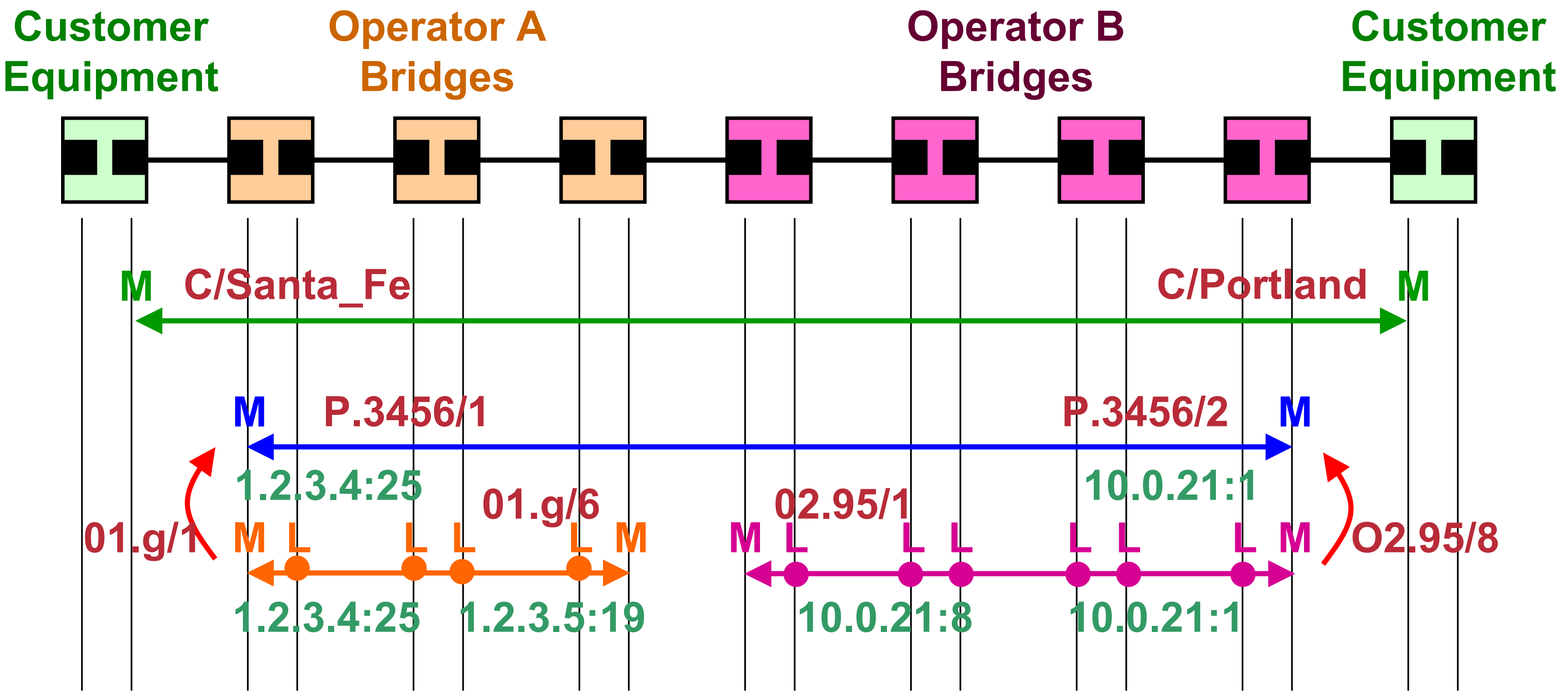
- Operators assign **MPIDs** to specific **hardware Interfaces** (IP addr : ifIndex)

5: Operator LP Discovery



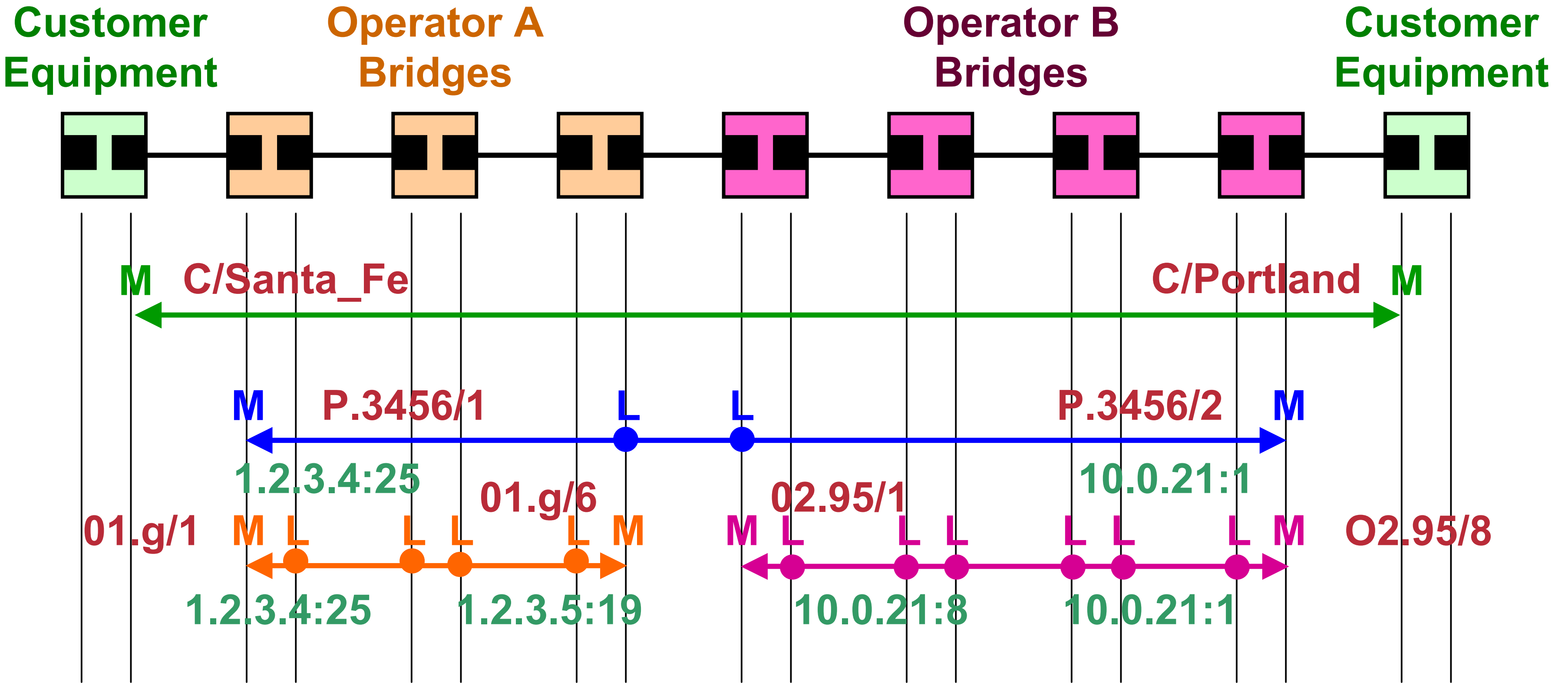
- Operators' MPs *discover* their LPs' hardware Interfaces and MAC addrs through E2E OAM and *start running*.

6: Provider MP Binding



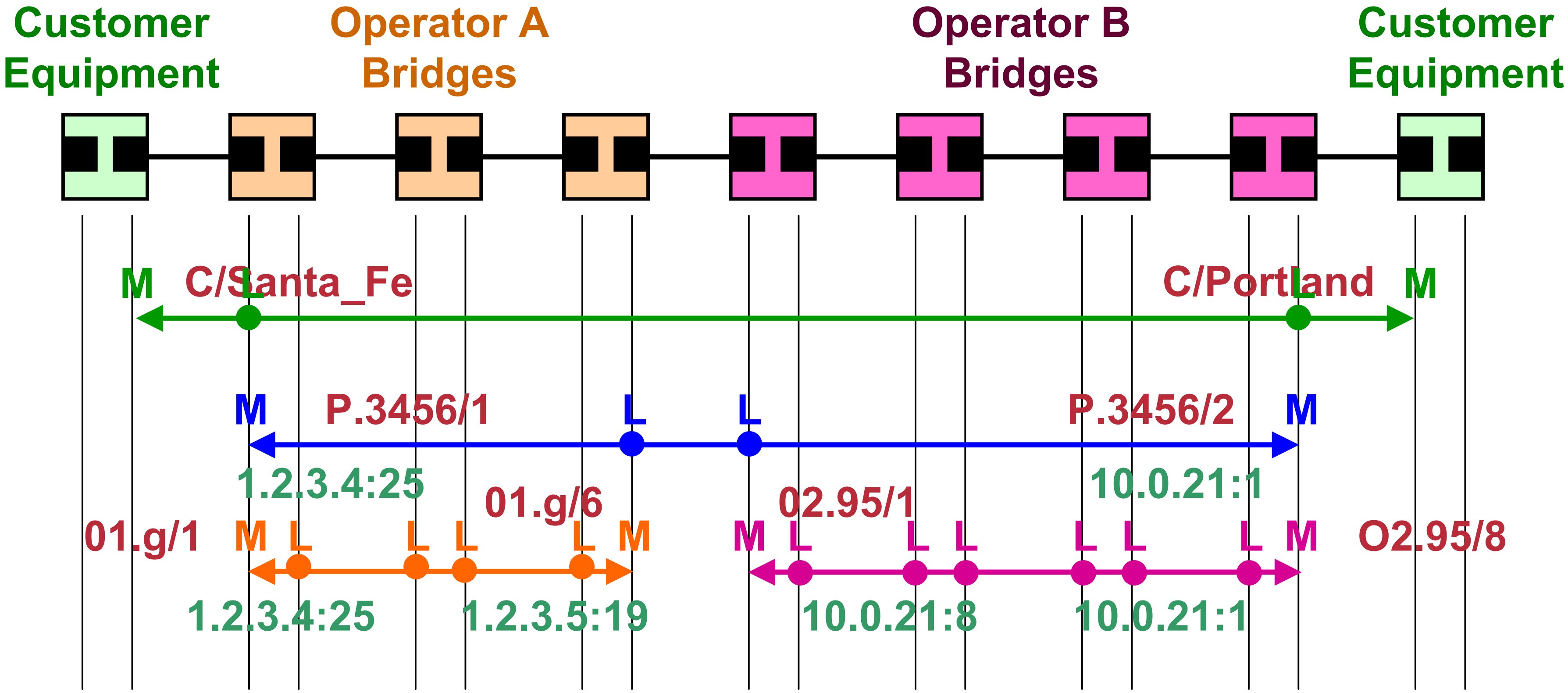
- Operators' report hardware Interface bindings of ***MPs to MPs*** to Provider

7: Provider LP Discovery



- **Provider's MPs discover their LPs' hardware Interfaces and MAC addrs through E2E OAM and *start running*.**

8: Customer LP Discovery



- Customer's MPs discover their LPs' hardware Interfaces (?) and MAC addrs through E2E OAM and ***start running***.

Implementation Choices

Implementation Choices: Three Methods

- A. Bridge "brain"

  All OAM goes through Bridges' internal MAC interfaces.

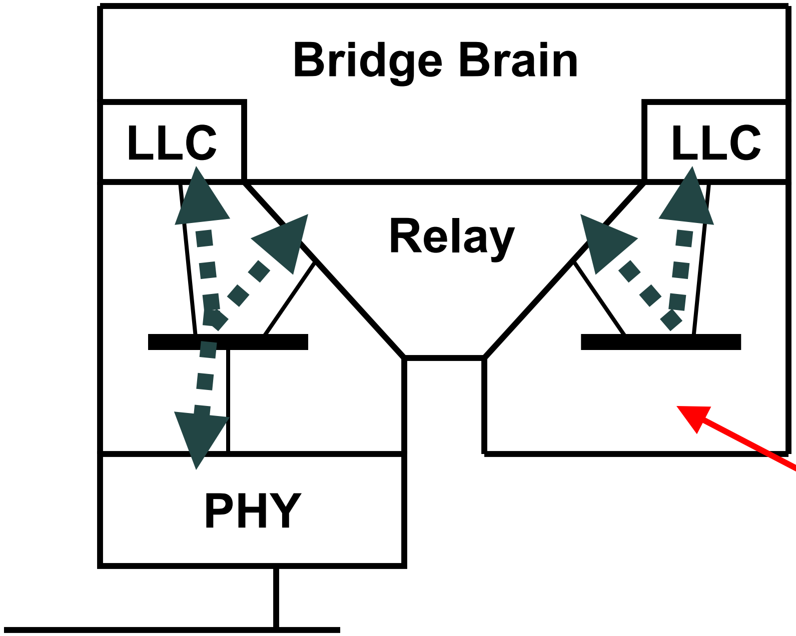
- B. Bridge Baggy Pants MAC Service (LLC)

  Most OAM goes through the LLC module in Bridge Ports' pants legs.

- C. Bridge Baggy Pants OAM Shim

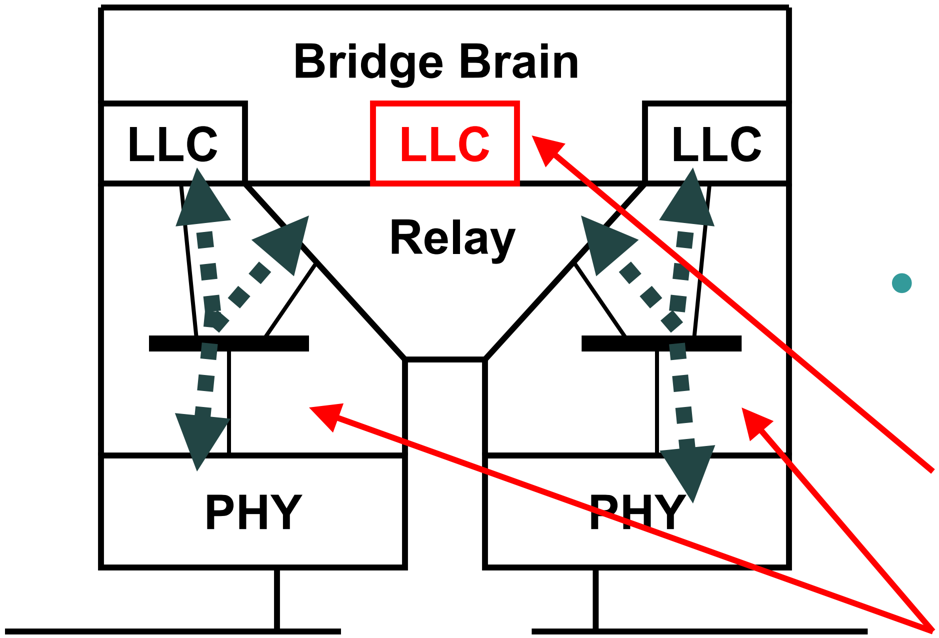
  Most OAM goes through an OAM shim inserted between MACSec and Media-dependent functions in Bridge Ports' pants legs

The 802.1Q “baggy pants” diagram: Bridge Brain model



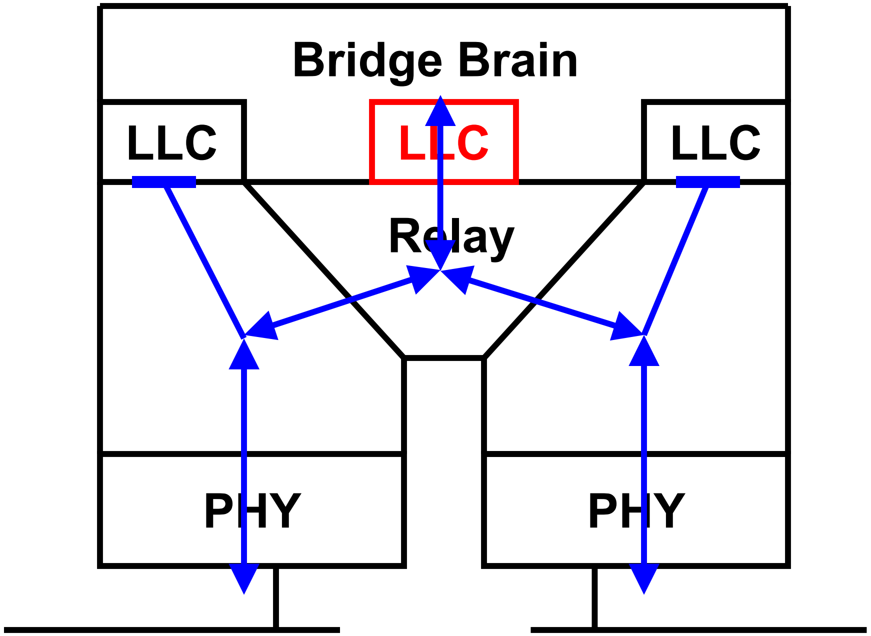
- Bridge transmits and receives many frames, typically including IP bridge management traffic, and specifically for this example, E2E OAM packets, through a separate (virtual?) interface through the Relay Function.

The 802.1Q “baggy pants” diagram: Bridge Brain model



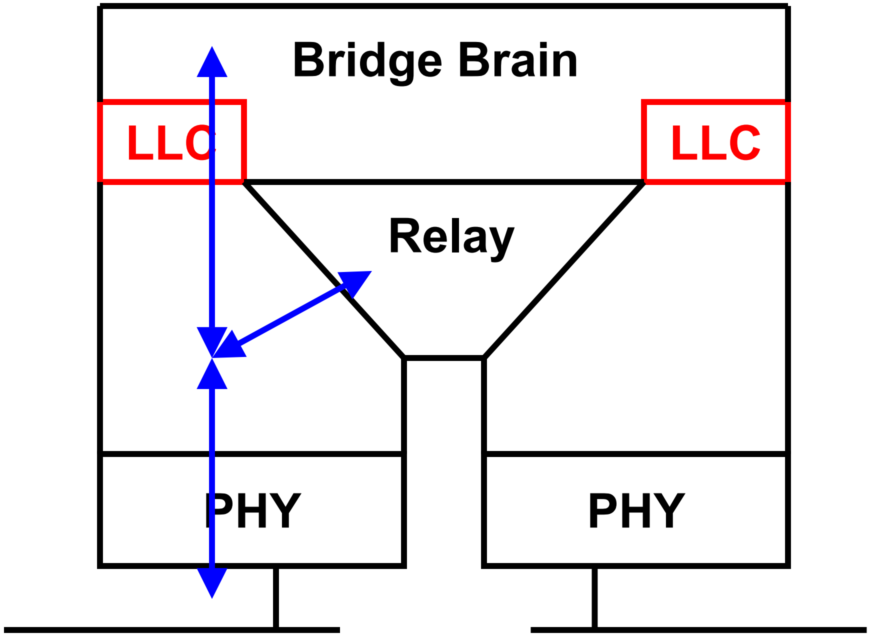
- We’ll draw the Bridge Brain interface like **this** in order to get two normal Ethernet **interfaces** into the diagram.

The 802.1Q “baggy pants” diagram: Bridge Brain model



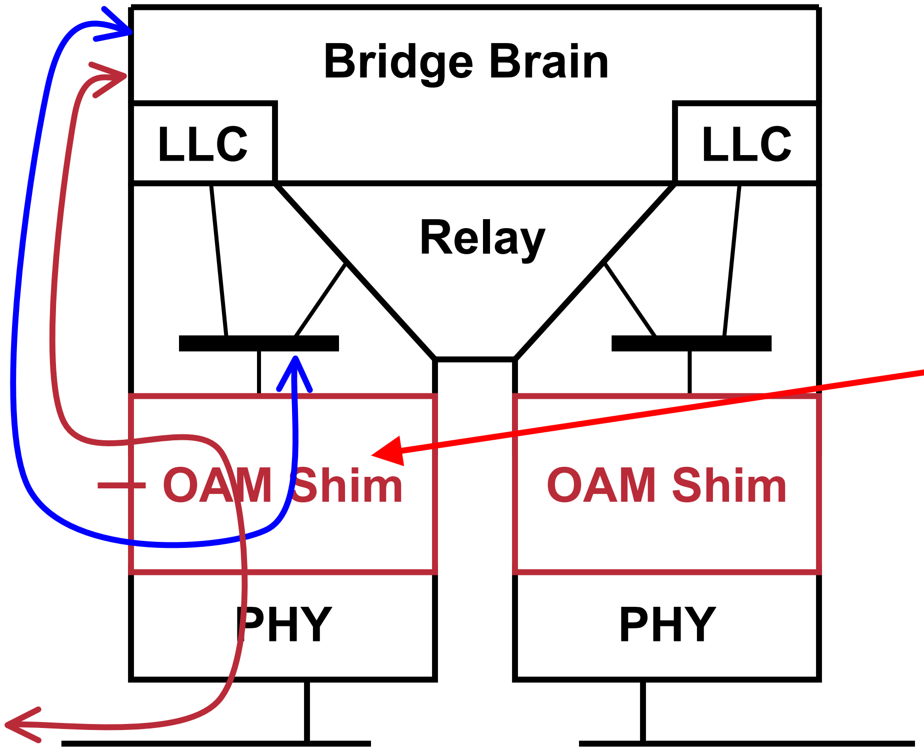
- The **OAM packets** are *all* sent through the Bridge Brain port.

The 802.1Q “baggy pants” diagram: LLC model



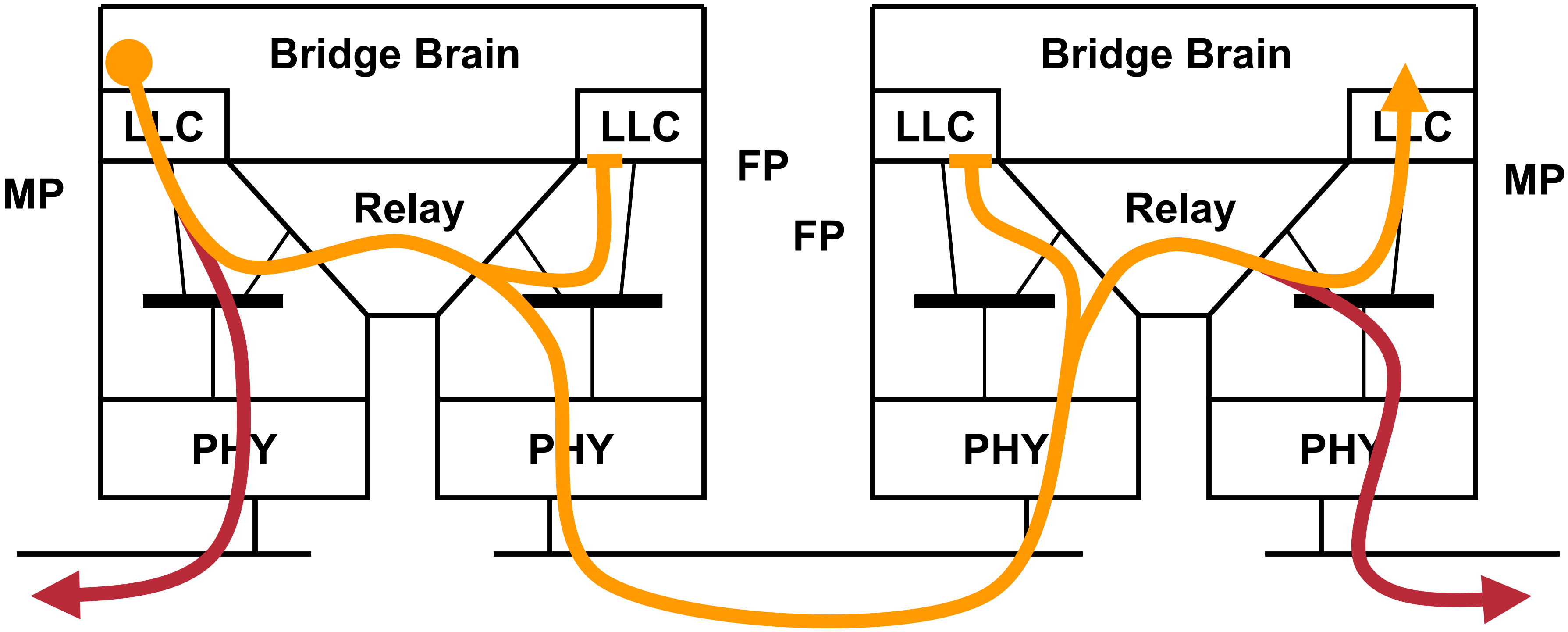
- The **OAM packets** are *mostly* sent through the MAC Service (LLC) of the relevant port.

The 802.1Q “baggy pants” diagram: OAM Shim model



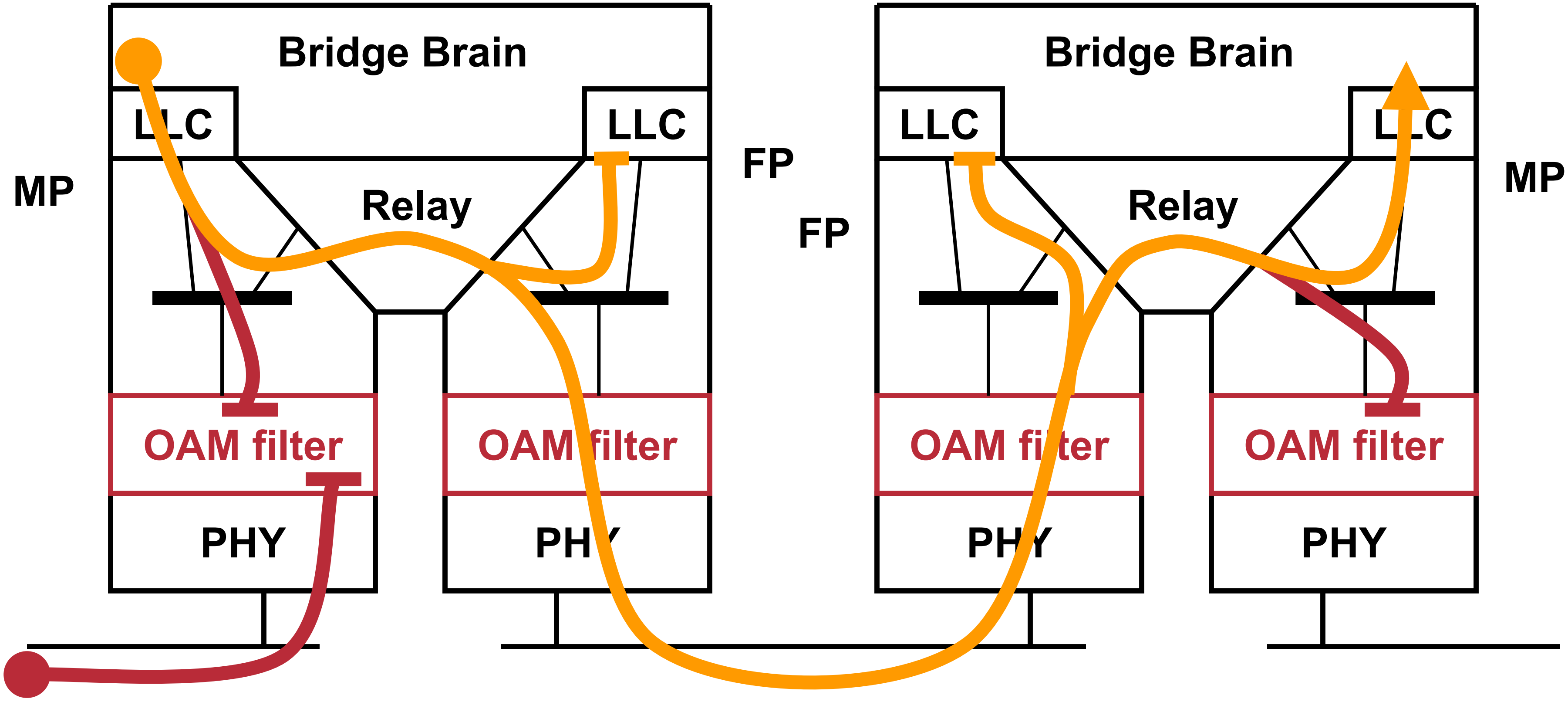
- OAM packets are *mostly* sent through an OAM Shim inserted into the MAC stack.
- OAM packets can be directed downward or upward through Shim, copied or stopped, etc.
- Normal data passes unaffected through OAM Shim.

LLC Model:
Frame flow for *multicast* or *unicast*



- But the OAM packets would escape from both ends of the path and another Domain (e.g. the Customer) would receive them!

LLC Model:
OAM Filter needed!

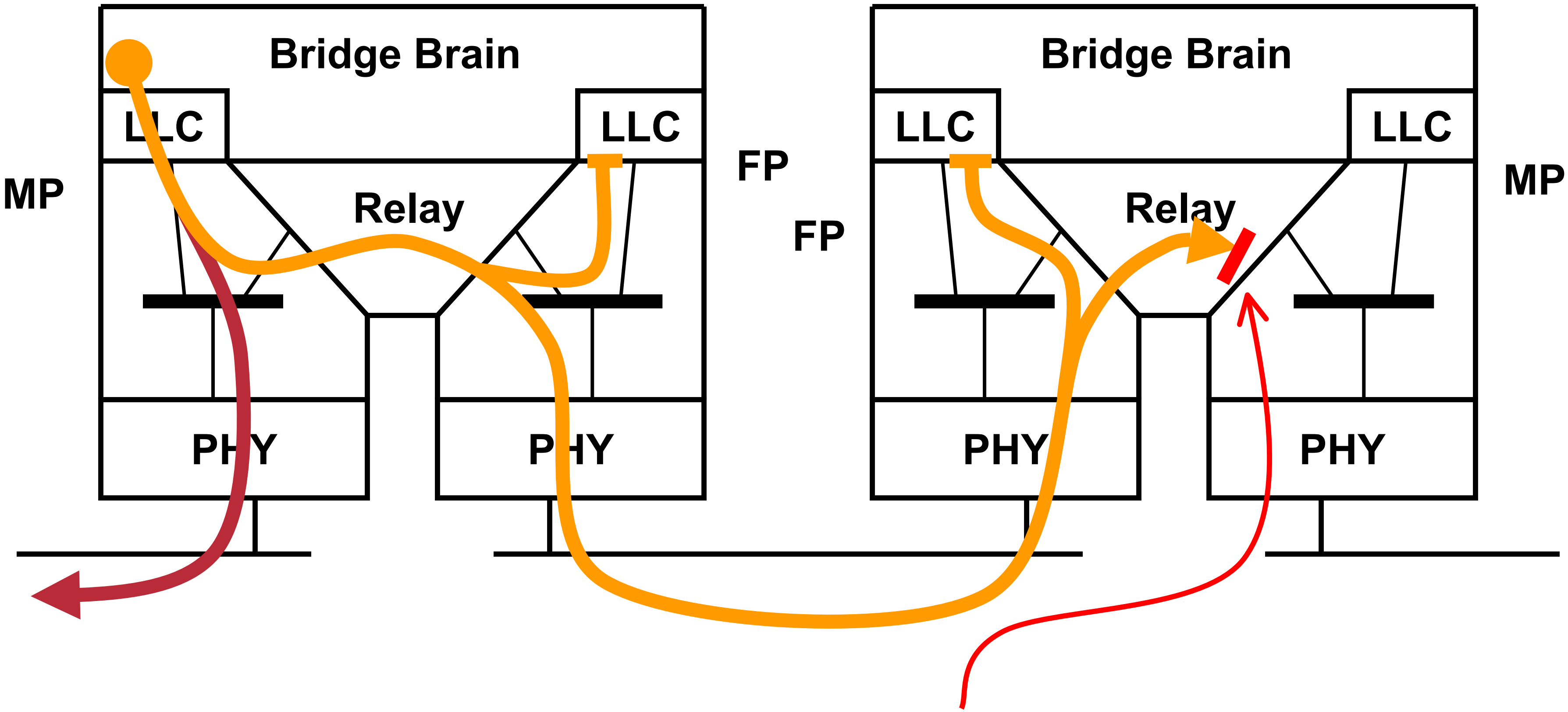


- An **OAM filter layer** blocks OAM frames based on Maintenance Level.

LLC Model:
The "Baggy Pants" OAM Filter

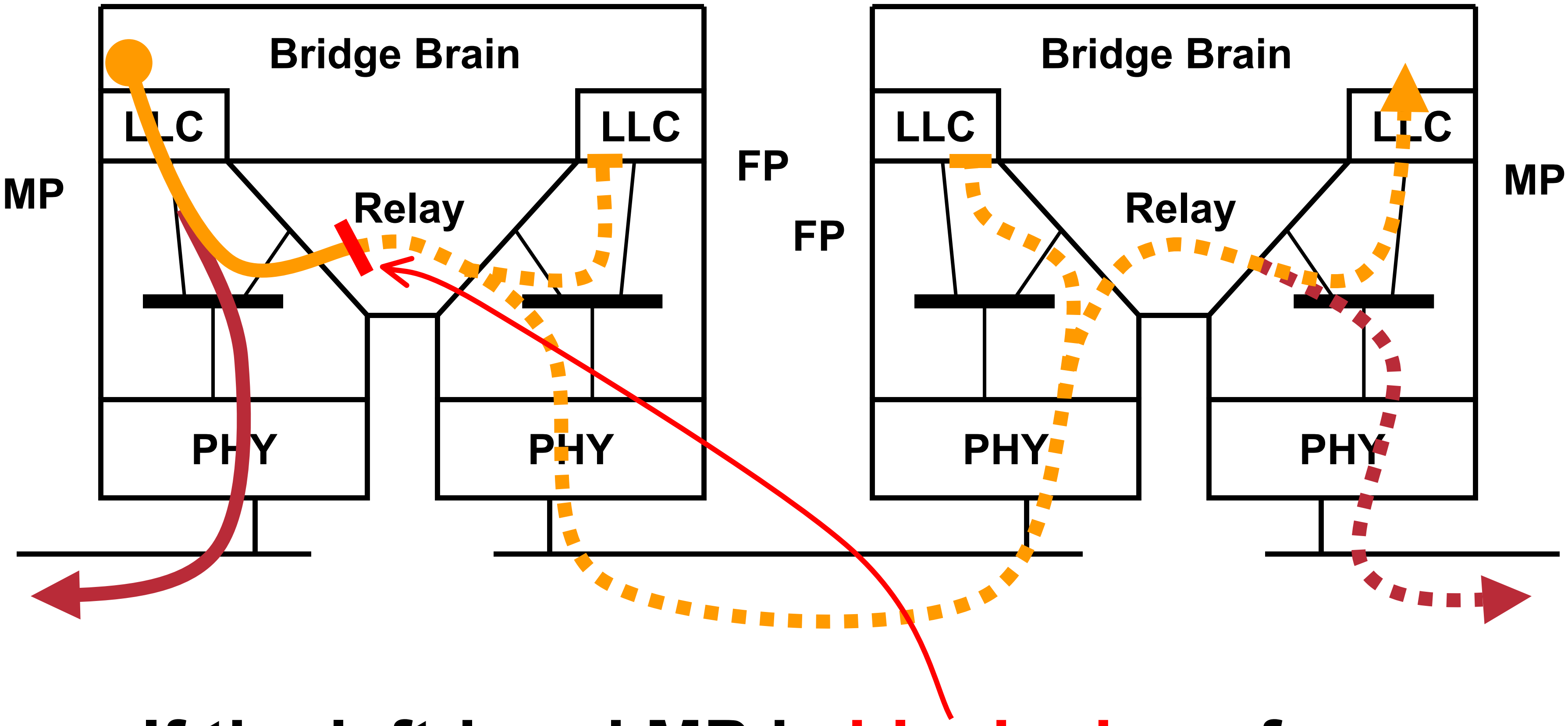
- OAM Filter stops OAM packets that are at or below a certain Maintenance Level.
- Required for Maintenance Points only.
- Multicast OAM packets can be filtered by destination MAC address to achieve *almost* the same result.
- But the OAM Filter is needed to prevent the inadvertent distribution of OAM packets destined for unknown unicast addresses.

OAM Shim or LLC: STP Blocking occurs in Relay Function, not in Bridge Ports



- If the right-hand MP is blocked, no frames can reach that MP through the Domain.

OAM Shim or LLC: STP Blocking occurs in Relay Function, not in Bridge Ports



- If the left-hand MP is **blocked**, no frames can go **anywhere** through the Domain.

And if blocked ports are isolated ...

- Blocked Maintenance Points cannot announce the fact that they are ready to work, but not actually in use.
- Blocked Maintenance Points cannot collect lists of other Maintenance Points in preparation for a fast shift to Forwarding.
- Traceroute towards a blocked Maintenance Points cannot reach that MP, and thus cannot send a Traceroute Reply.
- Blocked MPs are not available to *any* protocol from inside the Domain.

Bridge Brain, LLC, or OAM Shim models: Must we pick one or the other?

- Whether there is really a difference in path coverage between one or the other depends on the architecture of the bridge.
- Since Ethernets may be virtual, bridge can report the same MAC address for multiple MACs, so one cannot tell whether a bridge is using one or the other method.
- OAM to a blocked port requires the Bridge Brain model.

Bridge Brain, LLC, or OAM Shim models: We may observe:

- The only difference between the Bridge Brain and [either the LLC or OAM Shim] model is the MAC address reported by the Maintenance Point or Loopback Point.
- The Configuration Information flow shows that MAC addresses are discovered, not configured.
- MAC addresses (and sometimes, IP:Interface address) are variable, anyway, and must be discovered.

The Answer

- One may implement any of the three models, and use that model’s MAC address for most OAM operations.
- One must use either the Bridge Brain model (or use another Port’s LLC port) to access a blocked port via OAM.
- For the purposes of defining an End-to-End OAM protocol, the choice of implementation is irrelevant.

Summary

Metro Ethernet End-to-End OAM on one slide

- Different organizations' Maintenance Domains are nested in Levels.
- Maintenance Points are active, Loopback Points are passive, other Flow Points are transparent.
- Five kinds of OAM packets: Continuity Check, Traceroute Message / Reply, Loopback Message / Reply.
- Configuration info flow and access to blocked Bridge Ports lead to (at least) 3 interoperable implementation models.

Cisco Systems®