ICMP based OAM Solution for TRILL
draft-tissa-trill-oam-00.txt

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/1id-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 20, 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.

Abstract

This document presents a solution suite for TRILL data plane monitoring and failure detection. Methods presented herein allow in-cooperating IP payloads, exercising multi-paths, verifying multicast trees, locating end stations, virtual segments and diagnosing connectivity problems. ICMP protocol is proposed as framework for error reporting. Document also presents network wide health monitoring, distribution and reporting methods that are intended for efficient troubleshooting.

Table of Contents

1. Introduction...................................................4
   1.1. Motivation................................................5
   1.2. Contributors............................................6
2. Conventions used in this document..............................7
3. Protocol Architecture Overview................................7
   3.1. Overview of Tools........................................8
   3.2. TRILL Data Plane........................................9
   3.3. Monitoring..............................................10
   3.4. Distribution............................................10
   3.5. ISIS..................................................11
   3.6. Reporting...............................................11
4. Frame Format..................................................11
   4.1. Encoding of Request message............................11
   4.2. Encoding of Response Message............................13
   4.3. Encoding of Notification Message........................13
5. 127/8 in-band OAM IP address................................14
   5.1. IPv6 default in-band address............................15
6. Identification of Diagnostic frames............................15
   6.1. Identification of Layer 2 Flow..........................15
   6.2. Identification of IP Flows................................16
   6.3. Identification of Multicast Flows.......................18
      6.3.1. Identification of overall tree verification frames..18
      6.3.2. Identification of Layer 2 Multicast group verification frames...........................................18
6.3.3. Identification of IP Multicast group verification frames

6.4. Default OAM flow Parameters

7. ISIS Extensions

8. ICMP multi part extensions

8.1. C-Type Definitions

9. Details of Diagnostic tools

9.1. Loopback Message

9.1.1. Theory of Operation

9.1.1.1. Originator RBridge

9.1.1.2. Intermediate RBridge

9.1.1.3. Destination RBridge

9.2. Path Trace Message

9.2.1. Theory of Operation

9.2.1.1. Originator RBridge

9.2.1.2. Intermediate RBridge

9.2.1.3. Destination RBridge

9.3. Multicast Tree Verification (MTV) Message

9.3.1. Theory of Operation

9.3.1.1. Originator RBridge

9.3.1.2. Intermediate RBridge

9.3.1.3. In scope RBridges

9.4. MAC address discovery Message

9.4.1. Theory of Operation

9.4.1.1. Originator RBridge

9.4.1.2. Receiving RBridges

9.5. Address-Binding Verification Message

9.5.1. Extension to ARP and invARP

9.5.1.1. Encoding ARP-invARP extensions

9.6. End-Station Attachment Point Discovery

9.7. DRB and AF Discovery

9.7.1. Theory of Operation

9.7.1.1. Originator RBridge

9.7.1.2. Receiving RBridge

9.8. Notification Messages

10. Monitoring and Reporting

10.1. Data categories

10.2. Advertising Policy

10.3. Summary Category

10.4. Detail Category

10.5. Vendor Specific Category

11. Security Considerations

12. IANA Considerations

12.1. IANA considerations

12.1.1. ICMP Extensions

12.1.2. ARP Extensions

12.1.3. Well known Multicast MAC
1. Introduction

TRILL protocol has revolutionized how Layer 2 networks are being built and used. Legacy Ethernet networks provide single path for forwarding traffic and require all of the switches in the network to learn end-station MAC addresses. TRILL, on the other hand utilize multiple active links for forwarding thereby maximizing the overall network bandwidth utilization. TRILL is simple plug-and-play solution and does not require intermediate devices to learn MAC addresses of end-stations. These powerful characteristics of TRILL optimize performance and increase scaling limits. However, with that comes increased difficulty in diagnosing connectivity problems and locating end stations.

Network operators are used to troubleshooting legacy networks with single paths. Legacy devices maintain forwarding database of all end-station addresses in the Layer 2 network. Network administrators can trace the path taken by specific MAC address by examining the forwarding databases of devices. TRILL core switches, by design do not maintain end-station address database. Hence, administrators may not be able to trace a path taken by a specific MAC address by tracing the forwarding databases. Additionally, a given device may utilize multiple active paths to reach to a destination and may use a completely different forwarding topology for multicast traffic than it would use for unicast traffic. These challenges mandate the presence of an effective tool set to monitor and diagnose data plane failures in TRILL networks. These tools and protocols must stay as close as possible to the forwarding paths taken by actual data. OAM frames should not leak to end stations or out of the TRILL network to legacy networks.

TRILL base protocol specification [RFC6325] does not specify algorithm for selecting a path from a set of equal cost paths to forward a given flow. The majority of traffic in the networks is IP centric and most devices deploy some sort of hashing algorithm to identify the forwarding path from set of equal cost paths for a
given flow. Thus, it is desirable to use IP address and TCP/UDP port information as inputs to the ECMP selection hash function. Use of such higher level information provides better distribution of flows across multiple equal cost paths. This document, propose a framework that allow specifying, various combinations of payloads including IP payloads and actual payloads.

As TRILL based networks get deployed, during the transition period, it may be required for TRILL RBridges to co-exist with legacy networks. It is very helpful for the network operator if TRILL data plane failure detection tools allow isolating problem to specific legacy device or at least to the interface(s) that the downstream legacy device is connected. Solutions presented in this document facilitate identifying legacy devices or RBridge interfaces legacy devices are connected to.

ICMP (Internet Control Message Protocol) [RFC 792] has been in use for nearly three decades. ICMP multipart extensions [RFC4884], propose methods to extend ICMP messages to include additional information, without changing or inventing new ICMP message types. In this document we utilize ICMP for reporting of errors. ICMP multipart extensions will be utilized to define additional information that is specific to TRILL. Additionally use of ICMP allows sending error reports either in-band or out-of-band. Use of out-of-band ICMP allows network operators to diagnose uni-directional path failures easily. Also, the same ICMP infrastructure can be utilized to generate unsolicited error notifications for TRILL data plane failures, such as Destination unreachable, Time Exceed (TTL expiry), Parameter Mismatch (MTU mismatch) etc..

Availability of Network health information is a valuable starting point for any failure detection process. In this document we present the concept of network regions, monitoring of network regions and distribution of network health.

Diagnostic tools are also commonly referred to as OAM (Operations, Administration and Maintenance). In this document we use words diagnostics and OAM interchangeably. Unless explicitly specified both the words means the same.

1.1. Motivation

Currently published TRILL OAM solutions, [TRILLCH] and [TRILLOAM], mainly focus on data plane encoding and individual tools. The encoding methods presented in [TRILLCH] and [TRILLOAM], require defining OAM channel that utilize a special EtherType. Implementations that utilize ECMP selection algorithms based on
higher layer address information may require flexible OAM channel 
that allow specifying different payloads including IP based 
payloads.

Availability of network health information is important for 
efficient isolation of network connectivity problems. Currently 
there are neither standard sets of such data to be distributed nor 
framework to distribute network health data. Lack of such leads to 
cumbersome and time consuming troubleshooting of network 
connectivity issues, especially in multi-vendor networks.

Device virtualization is an increasing trend in datacenters and 
large enterprises. Physical servers may host multiple virtual 
servers and these virtual servers may move from physical server to 
physical server based on load balancing policies. As part of network 
connectivity problem isolation, it is important to identify the 
location of the virtual servers and R Bridges they are connected to. 
Currently, administrators are required to utilize multiple tools to 
locate these virtual machines and connecting R Bridges.

ICMP has been in use over three decades as the primary OAM tool of 
IP infrastructure. It is highly desirable to utilize the framework 
of existing infrastructure such as ICMP, thereby leveraging 
knowledge, implementation and time to market.

TRILL networks can co-exist with multi access LAN networks at the 
boundary of the TRILL network. TRILL protocol [RFC6325], introduced 
Designated R Bridge (DRB) and Appointed Forwarder (AF) concepts to 
ensure loop free forwarding and load splitting at the boundary of 
TRILL and multi access LAN networks. Discovery of DRB, AF and 
associated VLANs are important for effective fault isolation at the 
TRILL and multi access LAN boundary. Currently there are no known 
tools available for the purpose.

In this document we propose a framework and solution suite that will 
address the above.

1.2. Contributors

Many people contributed with ideas and comments. Among all, 
following people made notable contributions to all parts of this 
document and spend time reviewing, debating and commenting to ensure 
this specification addressees the problem space.
Ian Cox, Ronak Desai, Satya Dillikar, Rohit Watve, Ashok Ganesan and Leonard Tracy.

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].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

3. Protocol Architecture Overview

Effective OAM solution is not only a set of tools but a wholesome solution that covers all aspects of OAM, such as tools, monitoring, reporting etc. Solution presented in this document contains multiple subcomponents that cover various elements of the total solution. There are six subcomponents in the proposed architecture. These subcomponents collectively are called TRILL OAM Protocol. Here we present an overview of the architecture of the solution and explain the purpose of each of subcomponents and interaction between different subcomponents. Subsequent sections cover details of each of the subcomponents.
3.1. Overview of Tools

The Tools subcomponent consists of series of utilities to implement various data plane monitoring and failure detections methods. Individual tools are invoked directly by the user or by the monitoring subcomponent. Individual tools allow, where applicable, for callers to specify options such as ECMP coverage, destination RBridge nickname, pay-load etc. Tools interface with the TRILL data plane layer to send and receive OAM frames. At the time of writing following tools are included as part of the tool set.

1. Loopback Message (Ping)

2. Path Trace Message (Trace route)

3. Multicast Tree Verification (mtv)
4. MAC discovery

5. Address Binding Verification

6. IP End-station Locator

7. DRB-AF discovery

8. Notification messages

Tools, based on their intended use, can be classified into 3 broader categories as below.

<table>
<thead>
<tr>
<th>Category</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fault Verification</td>
<td>Loop Back Message</td>
</tr>
<tr>
<td>Fault Isolation</td>
<td>Path Trace Message, Multicast Tree Verification</td>
</tr>
<tr>
<td>Auxiliary</td>
<td>MAC discovery, Address Binding Verification, IP End-station Locator, DRB-AF Discovery, Error Notification</td>
</tr>
</tbody>
</table>

3.2. TRILL Data Plane

The TRILL data plane receives and transmits frames on behalf of the tools subcomponent. As far as the encapsulation is concern, TRILL data plane layer treat these frames exactly as it would treat a regular data frame. In fact one of the key design goals is to maintain TRILL data plane diagnostic (OAM) frames as close as possible to actual data frames. Additionally, implementation MUST satisfy the following requirements:

1. OAM frames SHOULD NOT leak in to legacy Ethernet or to end stations outside the TRILL cloud

2. RBridge MUST have ability to identify OAM (diagnostics) frames intended for a destination RBridge.
3. RBridgeS SHOULD have ability to identify TRILL data OAM frames that are not intended for itself and forward such frames without assistance from the CPU.

We explain in Section 6 various methods available to identify TRILL OAM (diagnostic) frames intended for the local RBridge and satisfy above requirements.

3.3. Monitoring

The Monitoring subcomponent utilize the tools subcomponent to monitor the TRILL data plane and proactively detect connectivity faults, configuration errors (cross connect errors) etc. The monitoring subcomponent provides options to specify frequency, retransmission count, ECMP choice and all other applicable options to the specific tool being used to implement the monitoring service. Based on the configuration specified by the user, the monitoring subcomponent periodically invokes the applicable tools. Additionally, based on configuration, monitoring results are propagated to the distribution subcomponent. Monitoring results are always associated with a monitoring region. The monitoring region is an administrative partition of the network such that it: 1. Maximize the fault coverage, 2. Optimize network health data summarization. More details of regions are discussed in Section 10.

3.4. Distribution

The distribution subcomponent has two primary inputs

- Data from the Monitoring Layer
- Data from other RBridges via ISIS GenApp

The distribution subcomponent performs the following functions:

- Advertising locally generated data
- Applying Advertising policies and re-advertising received data
- Maintaining the network health Database

Details of distribution layer and data handling are presented in section 10.
3.5. ISIS

TRILL OAM protocol suite proposed in this document utilize ISIS to distribute

OAM capability of individual RBridge

In-band OAM IP and MAC address

Above, OAM capability and In-band OAM address information are advertised using ISIS MT-Protocol extensions.[section 7. ]

Network monitoring data are distributed using ISIS GenApp extension methods specified in [GenApp]. Details of encoding and proposed TLV definitions are defined in detail in section 7.

3.6. Reporting

The Reporting subcomponent allows users to define and use various reports on network health. The Reporting subcomponent utilize data available in the distribution subcomponent to generate requested reports. Sample reports are listed in Appendix A.

4. Frame Format

TRILL data plane diagnostic (OAM) frames can be broadly classified in to three types: request, response and notification. Request messages are generated to measure TRILL data plane characteristics, such as connectivity. Response messages are generated by a RBridge in response to a request. Notifications are unsolicited messages generated due to certain failures such as unreachable destination. Details of individual messages are covered in later sections. Here we present frame encoding format for Request, Response and Notification messages.

4.1. Encoding of Request message
The above diagram depicts encapsulation of TRILL data plane diagnostic request frames. Encoded in the frame is the diagnostic payload. The diagnostic payload is a flexible structure that allow user to specify different kinds of payloads, including actual payloads. Most hardware implementations use IPDA:IPSA:DestPor:SrcPort based hash method to select ECMP paths for IP frames. For non IP payloads, RBridges normally uses a Layer 2 MAC DA and SA based hash for selecting an ECMP path. Flexible diagnostic payload allow user to drive end to end ECMP selection based on payload without needing additional hardware. Also, in terms of forwarding, this keeps diagnostic frame as close as possible to data frames. The length of the diagnostic payload must be deterministic. We propose a fixed 128 byte size for the diagnostic payload section of the OAM frame. This allows including IPv6 frames with multiple 802.1Q tags in to the diagnostic payload. The remaining bytes are set to zero, if the specified frame is smaller than the 128 byte fixed size.

ICMP header immediately follows the diagnostic payload. The ICMP header is constructed as defined in [RFC792].

ICMP multi part extensions [RFC 4884] are defined to carry additional information and are encoded after the ICMP header.[section 8. ]
4.2. Encoding of Response Message

Figure 3 Encoding of OAM response message

The above diagram depicts encoding of OAM response messages. If in-band delivery is requested, the OAM response message MUST be encoded as payload in a TRILL data frame. The ingress RBridge nickname MUST be set to the RBridge nickname of the node generating the response. Egress RBridge nickname MUST be set to the ingress RBridge nickname of the original TRILL data frame that triggered this response.

Normal IP forwarding rules MUST be followed, if an out-of-band response is requested.

4.3. Encoding of Notification Message

Notification messages are generated in response to an error condition such as delivery failure due to incompatible MTU or destination RBridge not in the forwarding table etc.. Out-of-band responses are generally indicated by explicitly including the indication to receive an out-of-band response in the TRILL OAM request frame. Since notifications are generated in response to regular data frames, the originator RBridge may not have methods to identify IP address required to deliver an out-of-band response. Hence, in this document we propose to deliver Notification messages in-band. Delivery of out-of-band messages are outside the scope of this document.

The RBridge generating the Notification message MUST include up to 128bytes of the original frame that triggered the notification message. If the original frame contains less than 128 bytes, then the remaining bytes MUST be padded with zeros.
The TRILL outer header of the frame that triggered the notification message is not included in the notification message. The Next-Hop header information in the original frame is of local significance to the specific link and may not be of interest to the originator of the data frame.

The following error messages are currently supported:

- Time Expiry
- Destination Unreachable
- Parameter Problem

Additional TRILL OAM error codes may be specified as ICMP multipart extensions in above notifications messages. These error codes indicate the cause of the error. Please see section 8 for error code definitions and section 9.8 for theory of operation.

5. 127/8 in-band OAM IP address

In this document we propose to use the same ICMP framework deployed in IP infrastructure for communicating OAM information. RBridges are not required to have IP interfaces enabled. However, in order to receive and process ICMP messages, RBridges are required to have at least a pseudo IP address. In this document, we propose to use 127/8 addressing scheme similar to the MPLS data plane failure detection methods [RFC 4373]. It is important that each RBridge have a straightforward method of identifying corresponding in-band OAM IP address of any given RBridge, without additional processing or lookups.
The 127/8 Address range is allocated for internal loopback addresses [RFC 1122] and required not to be routed. RFC 4373 updates RFC 1122 to utilize 127/8 addressing to communicate between devices in a peer-to-peer model that does not require routing. In this document, we propose to use 127/8 addressing model to identify in-band IP address required for OAM purposes. Additional methods are provided as ISIS LSP extension to announce, other addresses, user may desire to use for OAM in-band purpose. By default all RBridges MUST support the 127/8 addressing model specified here.

Each RBridge nickname is 16bits wide [RFC6325]. Let’s assume RBridge nickname RB is divided in RB(msb) and RB(lsb), such that, RB(msb) takes the upper 8bits of the RB and RB(lsb) takes the lower 8bits of the RB. Corresponding in-band IP address of RB is 127.RB(msb).RB( ls b).100. Implementation MUST facilitate methods to avoid conflicts between in-band OAM address and implementation specific 127/8 address allocations.

5.1. IPv6 default in-band address

IPv6 based systems have two options to derive the in-band IP address. The systems may choose, IPv6 native loopback address ::RBid:100 or IPv4 mapped IPv6 addressing format of ::FFFF:127.RB(msb).RB( lsb ).100.

RFC 4379, MPLS Data Plane failure detection methods, utilize IPv4 mapped IPv6 addressing. One of the design objectives of the proposal is to re-use as many existing OAM extensions as possible. Hence, implementation that support IPv6 MUST utilize the IPv4 mapped IPv6 addressing format for default IPv6 in-band address. Deployments that desire to utilize native addressing MAY advertise native IPv6 in-band address using OAM extensions in section 7.

6. Identification of Diagnostic frames

In this document we have proposed to use the TRILL header as defined in [RFC6325], without modifications. The standard TRILL header currently, does not provide option to identify diagnostic frames. Hence, it is important to have circumstantial methods to identify diagnostic frames intended for the local RBridge and prevent leaking of diagnostic frames outside of TRILL network. In this section we explain, various methods to attain the above goals.

6.1. Identification of Layer 2 Flow

As stated earlier, most RBridges use Destination and Source MAC address, combination to determine the next hop ECMP interface to
forward non IP frames. It is required to provide flexibility for the user to specify destination MAC address and source MAC address. We propose to use special EthType (TBD) to indentify OAM (diagnostic) frames that contain non IP diagnostic payloads.

Each RBridge, if TRILL data plane OAM enabled, MUST provide following processing:

- Forward frames that have egress RBridge nickname equal to local RBridge nickname and EthType equal to Diagnostic Ethtype, to the Central Processing Unit (CPU). Such frames SHOULD NOT egress out of the RBridge.
- The RBridge SHOULD not egress frames with Diagnostic Ethtype to non TRILL interfaces.

6.2. Identification of IP Flows

As stated earlier, most RBridges use combination of IP address and Layer 4 information such as UDP/TCP Port, to determine the next hop ECMP interface to forward IP frames. Hence, it is important to provide flexibility for users to specify destination IP addressing and payload information.

In this section we propose several approaches to identify OAM (diagnostic) frames with IP payloads that are addressed to the local RBridge for processing

Method 1:

Use of Well know Destination MAC address:

We propose to use a well known diagnostic MAC address (TBD-DMAC-1), as the Destination MAC address of the inner Layer 2 header.

Each RBridge, if TRILL data plane diagnostic is enabled, MUST provide the following processing:

- Forward frames which have egress RBridge nickname equal to the local RBridge nickname and Destination MAC address of the inner Layer 2 header equal to the Well Known Diagnostic MAC address (TBD-DMAC-1) to the Central Processing Unit (CPU). If RBridge nickname is not equal to the local RBridge nickname, frame MUST be forwarded normally.
- RBridge SHOULD NOT egress frames with the Diagnostic MAC address (TBD-DMAC-1) as destination address to non TRILL interfaces.
Method 2:

Use of Well know Source MAC address:

We propose to use a well known source MAC address (TBD-SMAC-1), as the source MAC address of the inner Layer 2 header.

Each RBridge, if TRILL data plane diagnostic is enabled, MUST provide following processing:

- Forward frames that have egress RBridge nickname equal to the local RBridge nickname and source MAC address of the inner Layer 2 header equal to Well Known source MAC address (TBD-SMAC-1), to the Central Processing Unit (CPU). If the egress RBridge nickname is not equal to the local RBridge nickname then the frame MUST be forwarded normally.
- Each RBridge SHOULD NOT egress frames with Well known MAC address as source address to non TRILL interfaces.
- RBridge SHOULD NOT dynamically learn the well known Source MAC address (TBD-SMAC-1) specified above.

Method 3:

Use of RBridge specific OAM MAC address:

Each RBridge may advertise, MAC address for the purpose of receiving OAM frames with IP payloads. Sending RBridges may use the advertised MAC address as the destination MAC address of the inner Layer 2 header of originating diagnostic request frames.

Each RBridge, if TRILL OAM is enabled MUST provide following processing:

- Forward frames that has egress RBridge equal to the local RBridge nickname AND Destination MAC address of the inner Layer 2 header equal to the advertised RBridge specific OAM MAC address, to the Central Processing Unit (CPU).
- RBridge SHOULD NOT egress frames with RBridge specific OAM MAC address as destination address to non TRILL interfaces.
6.3. Identification of Multicast Flows

Multicast frames are forwarded using one of the available multicast trees in the TRILL network [RFC6325]. Selection of a multicast tree is done at the ingress RBridge. Multicast frames are directed to a selected multicast tree at the ingress. Hence exact payload definition is not required for the purpose of ECMP selection. However, based on multicast pruning, certain multicast addresses may not be required to be forwarded to all members of the tree. Intermediate switches perform, (S,G) or (*,G), forwarding based on IP addresses for IP frames and MAC address for non IP frames. Hence, in order to verify the effect of multicast pruning users may require methods to specify Layer 2 and/or IP addressing information, as applicable. There are two types of multicast tree verification messages:

- Overall Tree Verification Messages
- Pruned Tree Verification Messages

6.3.1. Identification of overall tree verification frames

We propose to utilize a well known multicast diagnostic MAC address (TBD-GMAC-1) for this purpose. If TRILL data plane diagnostics are enabled, this specific MAC address MUST be installed on every RBridge for all trees and MUST NOT be subject to pruning.

Each RBridge performs (*,G) forwarding of the frames based on the well known multicast diagnostic MAC address (TBD-GMAC-1) in the inner Layer 2 destination address. Additionally, it send a copy of the frame to the CPU for analysis and generates a response to the requester. Please see section 8.3 for details of multicast tree verification message processing.

A RBridge SHOULD NOT egress multicast frames with above diagnostic MAC address in to non TRILL interfaces. Also, RBridge MUST discard any native frame received on non TRILL interfaces with the above diagnostic MAC address as the destination MAC address.

6.3.2. Identification of Layer 2 Multicast group verification frames

We propose to utilize the diagnostic EthType (TBD) that was defined earlier for identification of Layer 2 group verification frames. User SHOULD have the ability specify destination MAC address, source MAC Address, VLAN and payload data up to 128 octets.

Each RBridge, performs standard multicast forwarding. Additionally, if EthType of the frame is equal to the well known diagnostic
Ethtype (TBD), the RBridge sends a copy of the frame to the CPU for analysis and generating response to the requester. Please see section 9.3 for details of multicast tree verification message processing.

RBridge MUST NOT egress multicast frames with above EthType in to non TRILL interfaces. Also, RBridge MUST discard any native frame received on non TRILL interfaces with the above EthType.

6.3.3. Identification of IP Multicast group verification frames

We propose to use the well known MAC address (TBD-SMAC-1) defined in section 6.2 as the source MAC address. Users have flexibility to define, IP Address, VLAN and other payload data upto 128 octets. The Destination MAC address is derived based on the IP Multicast destination address.

R Bridges perform (S,G) or (*.G) forwarding using the IP address information. Additionally, each RBridge send a copy of the frame to the CPU, if the source MAC address matches the well known MAC address defined here in.

RBridge MUST NOT egress multicast frames with above source MAC address to non TRILL interfaces. Also, each RBridge MUST discard any native frame received on a non TRILL interfaces with the above source MAC address.

RBridge MUST NOT dynamically learn the well known source MAC address specified here.

6.4. Default OAM flow Parameters

Parameters specified herein SHOULD be utilized as default parameters. Parameters specified under the Fixed category MUST not be changed based on user specification and MUST be followed exactly as specified below.
<table>
<thead>
<tr>
<th>Flow type</th>
<th>Default Values</th>
<th>Fixed fields</th>
</tr>
</thead>
</table>
| Layer 2   | DA= Well Known MAC  
SA= RBridge Interface MAC  
VLAN=native VLAN | EthType=OAM(TBD) |
| IPv4 OR IPv6 | IP Address = in-band address  
IP Dest. Port = 3503  
IP Src. Port = TBD  
DA = OAM MAC of egress  
RBridge  
SA =ingress RBr interface  
MAC  
VLAN=native VLAN | EthType=0x8000 OR  
EthType=0x86DD |
| Multicast Tree Verification | SA= RBridge Interface MAC  
VLAN=native VLAN | DA=Well Known Multicast MAC  
EthType=OAM(TBD) |
| Layer 2 Multicast | DA= Well Known MAC  
SA= RBridge Interface MAC  
VLAN=native VLAN | EthType=OAM(TBD) |
| IP Multicast | IP Dest Address = Default OAM MCast address  
in-band-address  
IP Src. Address =  
IP Dest. Port = 3503  
IP Src. Port = TBD  
DA = OAM MAC of egress  
RBridge  
SA =ingress RBr interface  
MAC  
VLAN=native VLAN | EthType=0x8000 OR  
EthType=0x86DD |

Figure 5 Default Parameters of Diagnostic(OAM) Payloads
7. ISIS Extensions

A new ISIS subTLV definition is required to announce the following OAM related information:

- OAM capability
- OAM in-band IP address
- OAM in-band MAC address

We propose to define a single sub TLV structure within ROUETER-CAPABILITY ISIS TLV (242), to announce the above OAM information.

```
+--------+  
| Type   |  
+--------+  
| Length |  
+--------+  

|ver| Res |v|i|m|o|
+-----------------+  
| Sender nickname |  
+-----------------+  
| OAM MAC address |  
+-----------------+  
| OAM in-band IP address |  
+-----------------+  
```

Figure 6 ISIS extension for OAM

Type : (1 octet) TBD (one of the sub-TLV definitions under MT-PORT-CAP ISIS TLV)

Length : ( 1 octet) Length of the subTLV, in octets, excluding Type and Length fields. Minimum 2.

Ver : (4 bits) indicate the OAM version. Currently set to zero.

Res : (1 octet), Reserved for future use. Set to zero on transmission and ignored on receipt.

V   : (1 bit) if set, indicates IP address included in the TLV is IPv6. Only one of I or V bit MUST be set. If both are set, it is a malformed TLV and must be discarded without further processing.
I : (1 bit) if set indicate IP address included in the TLV is IPv4. Only one of I or V bits MUST be set. If both are set, it is malformed TLV and must be discarded without any further processing.

M : (1 bit) If set, indicates MAC address is included in the TLV.

O : (1 bit) If set, indicates announcing RBridge is OAM capable.

MAC Address : (6 octets), IEEE MAC address, associated with the in-band IP address. If included, the MAC address MUST precede the IP address.

IP Address : (4 or 16 octets), OAM in-band IP address. If present MUST follow MAC address.

Above PDU encoding MUST follow exact order as specified and fields are not interchangeable.

NOTE: Both I and V flags MAY be set to zero to indicate that announcing RBridge desire to use the default OAM address. The default OAM address is the 127/8 address derived as specified in section 5.

8. ICMP multi part extensions

We propose to utilize a new Class-Num [RFC4884] to identify TRILL OAM related extensions specified in this document and other related documents. IANA has established a registry for ICMP extensions and we intend to seek a Class-Num assigned for this purpose.

Within the TRILL OAM Class-Nume, C-Types are defined and registered in the IANA to identify various different extensions specified herein and other related future documents.

8.1. C-Type Definitions

Version: C-Type 1

Contain Version number, code and associated flags. Currently Out-of-band Request, Final and Cross Connect Error flags are defined.
Bits

<table>
<thead>
<tr>
<th>31</th>
<th>24</th>
<th>16</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
</table>

---+---+---+---+---+---+
| Version | code | Reserved | F | C | o |
---+---+---+---+---+---+

Figure 7 C-Type 1, Version and Flags

Version (8 bits): Currently set to zero

Code (1 octet): TRILL OAM Message codes. See below for currently available TRILL OAM Message codes.

Reserved (22 bits): Set to zero on transmission and ignored on receipt

F (1 bit): Final flag, when set, indicates this is the last response.

C (1 bit): Cross connect error (VLAN mapping error), if set indicates VLAN cross connect error detected. This field is ignored in request messages and MUST only be interpreted in response messages.

O (1 bit): If set, indicates, OAM out-of-band response requested.

TRILL OAM Message codes:

0 : Loopback Message Request
1 : Loopback Message Response
2 : Path Trace Request
3 : Path Trace Response
4 : Time Expiry Notification (error)
5 : Parameter Problem Notification (error)
6 : Destination Unreachable (error)
7 : Multicast Tree Verification Request
8 : Multicast Tree Verification Response
9 : MAC Address discovery Request
10 : MAC Address discovery Response
11 : DRB discovery request
12 : DRB discovery response
13 : AF discovery request
14 : AF discovery response
15 : AF-VLAN discovery request
16 : AF-VLAN discovery response
17 - 255 : Reserved
Originator IP Address: (C-type 2)

Length of the ICMP extension header indicates whether the address is IPv4 or IPv6. Please refer to RFC 4884 for ICMP extension encoding and ICMP header structure.

```
+---------------------------------+
|                                 |
. IP Address                      .
|                                 |
+---------------------------------+
```

Figure 8 C-Type 2 Originator IP address

Upstream Identification: (C-type 3)

The Upstream Identification C-type structure encodes upstream path information such as upstream neighbor nickname, ingress interface index (ifindex) and name of the ingress port.

```
+------------------+---------------+
|  nickname        | Reserved      |
+------------------+---------------+
|            ifindex               |
+------------+---------------------+
|   Slot           |    Port       |
+------------------+---------------|
|    Speed         |    State      |
+------------------+---------------+
```

Figure 9 C-Type 3 Upstream Identification

Nickname (2 octets): TRILL 16 bit nickname of the upstream RBRdigest. [RFCtrill]

Reserved (2 octets) : Reserved, set to zero on transmission and ignored on receipt.

Ifindex (2 octets) : unsigned integer of local significance
Slot     (2 octets) : Slot number
Port     (2 octets) : Port number
Speed    (2 octets) : Speed in 100Mbps. Zero (0) indicates port speeds less than 100Mbps.
State    (2 octets) : Represent the state of the port.
  0: Down - no errors
  1: Disable
  2: Forwarding-no errors
  3: Down - errors
  4: Forwarding - errors
  5: Forwarding - oversubscribed
  6: Link Monitoring disable
     All other values reserved.

Monitored VLAN(diagnostic VLAN ) : (C-type 4)

Monitored VLAN c-type include in the ICMP extensions allows for testing the integrity of the inner payload VLAN and the expected VLAN. The expected VLAN is encoded in the Monitored VLAN c-type. The destination RBRidge, compare the VLAN of the inner payload with the VLAN value encoded in the Monitored VLAN c-type. If these two VLAN values mismatch, RBRidge SHOULD set the cross connect flag in the response. A RBridge MUST NOT set the cross connect error flag for other than the above specified VLAN mismatch scenario.

Bits
+---------------------------------+
| Reserved   |  VLAN              |
+------------+--------------------+
Figure 10   C-Type 4 Diagnostic VLAN

Downstream Identification: (C-Type 5)

The Downstream Identification C-type carries multiple sets of data, each corresponding to individual downstream neighbor among collection of equal cost paths.
Bits

+------------------+---------------+ ----
| ecmp count       | Reserved      | ^
| Reserved         | nickname      | |
| ifindex           | Next hop neighbor information |
| Slot              | Port          | |
| Speed             | State         | v
| Repeat next hop neighbor identification for each neighbor |

Figure 11  C-Type 5 Downstream Identification

Ecmp count (2 octets): Number of equal cost paths to the given destination from this RBridge.

Reserved (4 octets): Reserved, set to zero on transmission and ignored on receipt.

Next-hop neighbor information:

Nickname (16 bits): TRILL 16 bit nickname [RFCtrill]

Ifindex  (2 octets) : unsigned integer of local significance

Slot     (2 octets) : Slot number

Port     (2 octets) : Port number

Speed    (2 octets) : Speed in 100Mbps. Zero (0) indicates port speeds less than 100Mbps.

State    (2 octets) : Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
6: Link monitoring disable
All other values reserved.

NOTE: Repeat Next-hope neighbor identification entry per each ECMP.
Total number of neighbor entries MUST equal to ecmp count.
Individual neighbor entry MAY have variable length.

Path for this payload: (c-Type 6)
Path for this payload indicates the next hop neighbor that this frame could have been forwarded on based on the payload hashing.

<table>
<thead>
<tr>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>+-----------------------------------+</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>+-----------------------------------+</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>+-----------------------------------+</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>+-----------------------------------+</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Figure 12  C-Type 6 Path for this payload

Nickname (16 bits): TRILL 16 bit nickname [RFCtrill]

Ifindex  (2 octets) : unsigned integer of local significance. 0xFFFF indicate CPU.

Slot  (2 octets) : Slot number

Port  (2 octets) : Port number

Speed  (2 octets) : Speed in 100Mbps. Zero (0) indicates port speeds less than 100Mbps.

State  (2 octets) : Represent the state of the port.
0: Down - no errors
1: Disable
2: Forwarding - no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
6: Link monitoring disable
All other values reserved.

DRB Information (c-Type 7)

| 31 | 16 | 8 | 0 |
+---------------+--------+--+-+
| nickname      | state  | R|P|
+---------------+--------+--+-+

Figure 13 Nickname of the DRB

Nickname (2 octets) : TRILL nickname of the DRB
State (1 octets) : DRB state
R (7 bits ) : set to zero on Transmission and ignored on receipt
P (1 bits ) : Set when pseudo node bypass is indicated by the DRB for the link

AF Information (C-Type 7)
Follow the same encoding as C-Type 6, above.
Nickname and state are of the AF.

Enable VLAN List (c-Type 8)

| 31 | 27 | 16 | 12 | 0 |
+---------------+-----------+---+----------+
| R | St-VLAN | R | End-VLAN |
+---------------+-----------+---+----------+

Figure 14 Enabled VLAN List

R (4 bits) : Reserved, set to zero on transmission and ignored on receipt.
St-VLAN (12 bits) : Start VLAN
End-VLAN (12 bits) : End VLAN

Start VLAN and End VLAN represent the range of enabled VLANS. If the VLAN range is non-contiguous, then multiple Enabled VLAN lists MUST be included, each representing a contiguous VLAN set.

Announcing VLAN set (c-Type 9)

Announcing VLAN list uses the same format as the Enable VLAN List (c-Type 8)

```
31 27 16 12 0
+------------------+----------+
| R | St-VLAN | R | End-VLAN |
+------------------+----------+
```

Figure 15 Announcing VLAN List

R (4 bits) : Reserved, set to zero on transmission and ignored on receipt.

St-VLAN (12 bits) : Start VLAN
End-VLAN (12 bits) : End VLAN

Start VLAN and End VLAN represent the range of announcing VLANS. If the VLAN range is non-contiguous, then multiple of announcing VLAN list MUST be included, each representing a contiguous VLAN set.

AF List (c-Type 10)

This c-Type lists the VLANS for which responding RBridge is a the appointed forwarder.
31 27 16 12 0
+---------------+-------------+
| Reserved      | nickname    |
+---------------+-------------+
| R             | St-VLAN     | R | End-VLAN    |
+---------------+-------------+

Figure 16  AF List

Reserved (2 octets) : set to zero on transmission and ignored on receipt.

Nickname (2 octets) : TRILL 16 bit nickname of the RBridge

R (4 bits) : Reserved, set to zero on transmission and ignored on receipt.

St-VLAN (12 bits) : Start VLAN

End-VLAN (12 bits) : End VLAN

AF List MUST be repeated for each of the contiguous VLAN ranges that the responding RBridge function as Appointed Forwarder.

DRB Life Time (c-Type 11)

DRB Life time indicates the Life time, of the DRB operational role, of the RBridge.

31                                      0
+--------------------------------------+
|                                      |0
+   Life Time                         +
|                                      |1
+--------------------------------------+

Figure 17  DRB Life Time

Life Time ( 8 octets): Indicates the Life time of the operational role in seconds.

AF Lifetime (C-Type 12)

AF Life time indicates the Life time, of the AF operational role, of the RBridge for the specified VLAN.
Encoding follow the same format specified in C-Type 11.

Designated VLAN changes (C-Type 13)

Indicates number of times a given RBridge has observed Designated VLAN changes. Each change may potentially lead to traffic disruptions.

```
15            0
+-------------+
| Change count|
+-------------+
```

Figure 18   Number of times Designated VLAN changes

Change count (2 octets): Indicates number of times a given RBridge has observed Designated VLAN changes

RBridge scope List (c-Type 14)

```
15           0
+-----------+
 |  R   |  Nu  |
+-----------+
 | nickname 1|    
+-----------+
 .           
.           
 | nickname n|
+-----------+
```

Figure 19   Scope List c-Type 15

R (1 octet ) : Reserved, zero on transmission and ignored on receipt.

Nu (1 octet) : number of nicknames listed

Nickname 1 .. n (2 octets) each: List TRILL RBridge nickname of in scope RBridges.

Nicknames MUST be numerically sorted. With nickname1 the lowest to nickname n the highest. This facilitate easy processing the receiving RBridge.
Nu = 0 indicate no embedded nicknames in the message and response required from all RBridges, where applicable.

Multicast Tree downstream List (c-Type 15)

Multicast Tree downstream list provides information on downstream leaf Rbridges on the specified tree.

Bits

<table>
<thead>
<tr>
<th>Bit</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>Leaf count</td>
</tr>
<tr>
<td>30-1</td>
<td>Reserved</td>
</tr>
<tr>
<td>31</td>
<td>Nickname</td>
</tr>
<tr>
<td>30-1</td>
<td>Reserved</td>
</tr>
<tr>
<td>31</td>
<td>Ifindex</td>
</tr>
<tr>
<td>30-1</td>
<td>Downstream leaf information</td>
</tr>
<tr>
<td>31</td>
<td>Slot</td>
</tr>
<tr>
<td>30-1</td>
<td>Port</td>
</tr>
<tr>
<td>31</td>
<td>Speed</td>
</tr>
<tr>
<td>30-1</td>
<td>State</td>
</tr>
<tr>
<td>31</td>
<td>Repeat downstream</td>
</tr>
<tr>
<td>30-1</td>
<td>Leaf information for each downstream RBridge</td>
</tr>
</tbody>
</table>

Figure 20  C-Type 5 Downstream Identification

Leaf count (16 bits): Number of RBridges downstream to this RBridge.

Downstream leaf information:

Nickname (16 bits): TRILL 16 bit nickname [RFCtrill]

Ifindex (32 bits): Unsigned 32 bit integer that has only a local significance to the sending RBridge. Value 0xFFFF indicates CPU interface.

Slot     (2 octets): Slot number
Port     (2 octets): Port number
Speed (2 octets) : Speed in 100Mbps. Zero (0) indicates port speeds less than 100Mbps.

State (2 octets) : Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
All other values reserved.

NOTE: Repeat downstream RBridges reachability information per each leaf node. Total number of neighbor entries MUST equal to leaf count. Individual neighbor entry MAY have variable length.

MAC-discovery Address List (c-Type 16)

```
15           0
+------------+
| count      |
+------------+
| MAC        |
+ Address 1 +
|           |
+          |
+ Address n |
+------------+
```

Figure 21   MAC-discovery Address List

Count (2 octets) : Number of MAC addresses embedded in the response

MAC Address (6 octets) : 6 octet MAC address

MAC-discovery response Address List (c-Type 17)
<table>
<thead>
<tr>
<th>Count (2 octets)</th>
<th>Number of MAC addresses embedded in the response</th>
</tr>
</thead>
<tbody>
<tr>
<td>T (4 bits)</td>
<td>Type of MAC address 0 - Dynamic, 1 Static, 2-15 Reserved</td>
</tr>
<tr>
<td>VLAN (12 bits)</td>
<td>VLAN identifier associated with the MAC address</td>
</tr>
<tr>
<td>MAC Address (6 octets)</td>
<td>6 octet MAC address</td>
</tr>
</tbody>
</table>
Age (8 octets): Age of the MAC address in seconds. For a static MAC address, this field is ignored.

Ifindex (4 octets): Interface index on which MAC address is learnt.

Slot (2 octets): Slot number of the interface on which this MAC address is learnt.

Port (2 octets): Port number of the interface on which this MAC address is learnt.

vNTAG (2 octets): Virtual TAG identifier associated with the MAC address. Value 0 indicates no vNTAG association with the MAC address.

Speed (2 octets): Speed in 100Mbps. Zero (0) indicates port speeds less than 100Mbps.

State (2 octets): Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding - no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
6: Un-monitored
All other values reserved.

Error code (c-Type 18)

Error code c-Type allows an RBridge to specify various error codes within high-level notification messages such as Time Expiry, Parameter Problem and Destination unreachable. The sub-error codes within each of the error code allow specifying further details of the error.

```
+-----------------+--------------+
| Error Code      | sub-code     |
+-----------------+--------------+
```

Figure 23 C-Type 18 Error code
Error Code (2 octets) : Identify the error. Currently following errors are defined

0 - VLAN non existent
1 - VLAN in suspended state
2 - Cross connect error
3 - Unknown RBridge nickname
4 - Not AF
5 - MTU mismatch
6 - Interface not in forwarding state
7 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

Sub-code (2 octets) : identify the sub-error code.
0 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

Warning code (c-Type 19)

Warning code c-Type allow a RBridge to specify various error codes within high-level notification messages such as Time Expiry, Parameter Problem and Destination unreachable. The sub-warning codes within each of the warning codes allow to specify further details of the warning.

<table>
<thead>
<tr>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Figure 24   C-Type 19 Warning code

Warning Code (2 octets) : Identify the Warning. Currently following Warnings are defined

0 - Invalid RBridge nickname (RBridge nickname in the range 0xffco to 0xffff)
1 - Invalid VLAN (Reserved VLAN)
2 - AF VLAN list Mismatch
3 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

Sub-code (2 octets) : identify the sub-error code.
0 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

Information code (c-Type 20)

Information code c-Type allow a RBridge to specify various information codes within the high-level notification messages such as Time Expiry, Parameter Problem and Destination unreachable. The sub-info codes within each of the code allow specifying further details of the information.

```
+------------------|--------------+
| Information Code | sub-code     |
+------------------|--------------+
```

Figure 25    C-Type 19 Information code

Information Code (2 octets) : Identify the Information. Currently following Information are defined

0 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

Sub-code (2 octets) : identify the sub-error code.

0 - 0xFFFF - Reserved for future use and MUST not be used in transmission.

9. Details of Diagnostic tools

In this section we present details of various diagnostic tools that are identified as part of the solution. We assume, readers are familiar with frame encoding methods, diagnostic frame identification methods, and ISIS and ICMP extensions presented earlier in the document. In this section we will only make reference to the extensions and methods, please refer to prior section for details.
9.1. Loopback Message

Loopback message is utilized for fault verification. It verifies connectivity between two R Bridges, for a specified flow. Monitoring subsystem may use Loopback Message for connectivity monitoring and proactive fault detection. Users may specify exact flow, part of it or not at all. Additionally, users may also specify, ECMP choice at the ingress. ECMP choice can be a specific index, set of index, all of the index or non. If no ECMP index specified, payload is used to determine the ECMP choice. Method of deriving the ECMP choice using payload is implementation dependent and is outside the scope of this document. However, CPU generating the Loopback message SHOULD use the same ECMP selection algorithm as the data plane. Additionally some implementation may allow users to specify the ingress interface that actual flow may ingress to the RBridge. Although ability to inject the data plane diagnostic frames from the ingress interface is optional feature, it is highly desirable, as it allows verifying end-end connectivity from an ingress port to an egress RBridge.

Egress RBridge can sent its response either in-band or out-of-band. In-band response, additionally allow to measure round trip delay. In-band responses are tagged with the same VLAN as the request frame. ICMP multi part extensions in the request message allow user to specify whether out-of-band response required. If out-of-band request required, IP address it desire to receive the response MUST be specified.

Additionally, diagnostic VLAN, may be specified as part of the ICMP multi part extensions. Receiver RBridge may compare inner VLAN in the payload and the specified diagnostic VLAN. If the two specified VLAN values do not match, C flag in Version C-type SHOULD be set to indicate cross connect error.

9.1.1. Theory of Operation

9.1.1.1. Originator RBridge

Identify the destination RBridge based on user specification or based on location of the specified address (see below sections for MAC discovery and address locator).

Construct the diagnostic payload based on user specified parameters. Default parameters MUST be utilized for unspecified payload parameters. See Figure 5 for default parameters.

Construct the ICMP Echo request header. Assign applicable identification number and sequence number for the request.
ICMP multi part extension Version MUST be included and set appropriate flags. Specify the code as Loopback Message Request(0).

Construct following ICMP multipart extensions, where applicable

- Out-of-band response request
- Out-of-band IP address
- Diagnostic VLAN

Specify the Hop count of the TRILL data frame per user specification. Or utilize the applicable Hop count value, if TRILL TTL is not being specified.

Dispatch the diagnostic frame to the TRILL data plane for transmission.

RBridge may continue to retransmit, the request at periodic interval, until a response received or re-transmission count expires. At each new re-transmission sequence number may be incremented.

9.1.1.2. Intermediate RBridge

Intermediate RBridges forward the frame as a normal data frame and no special handling is required.

9.1.1.3. Destination RBridge

Destination RBridge performs, frame identification methods specified in above section 5. If the Loopback message is addressed to the local RBridge, then the RBridge forward the Loopback messages to the CPU for processing. CPU performs frame validation and constructs the response as stated below.

Construct the IP header for the ICMP echo response. If no out-of-band response requested, IP address in the IP header MUST be in-band IP address. If out-of-band response requested destination IP address is the IP address specified in the request message. Source IP address is derived based on the outgoing IP interface address.

Construct the ICMP echo reply header using the received ICMP echo request.

Include the received TRILL header and diagnostic payload in to the data field of the ICMP echo request frame [section 4.2].
If in-band response was requested, dispatch the frame to the TRILL data plane with request-originator RBridge nickname as the egress RBridge nickname.

If out-of-band response was requested, dispatch the frame to the standard IP forwarding process.

Error handling:

If VLAN cross connect error detected or inner.VLAN does not exist in the RBridge then generate Destination Unreachable message and specify the cause using error codes.

9.2. Path Trace Message

Primary use of Path Trace Message, commonly known in the IP world as "traceroute", is fault isolation. It may also be used for plotting path taken from a given RBridge to another RBridge. Operation of Path Trace message is identical to Loopback message except, that it is first transmitted with a TRILL Hop count field value of 1. Sending RBridge expect a Time Expiry message from the next hop or a successful response. If a Time Expiry message is received as the response, the originator RBridge record the information received from intermediate node that generated the Time Expiry message and resend the message by incrementing the previous Hope count value by 1. This process is continued until, a successful response is received from the destination RBridge or Path Trace process timeout occur.

9.2.1. Theory of Operation

9.2.1.1. Originator RBridge

Identify the destination RBridge based on user specification or based on location of the specified address (see below sections for MAC discovery and address locator).

Construct the diagnostic payload based on user specified parameters. Default parameters MUST be utilized for unspecified payload parameters. See Figure 4 for default parameters.

Construct the ICMP Echo request header. Assign applicable identification number and sequence number for the request.

ICMP multi part extension Version MUST be included and set appropriate flags. Set the code to Path Trace Request (2)
Construct following ICMP multipart extensions, where applicable

- Out-of-band response request
- Out-of-band IP address
- Diagnostic VLAN

Specify the Hope Count of the TRILL data frame as 1 for the first frame. Or use Hope Count value incremented by 1 if this is a retransmission generated in response to received Time Expiry message.

Dispatch the diagnostic frame to the TRILL data plane for transmission.

RBridge may continue to retransmit, the request at periodic interval, until a response received or re-transmission count expires. At each new re-transmission sequence number may be incremented.

9.2.1.2. Intermediate RBridge

Intermediate RBridge receive the diagnostic frame as Hope count expired frame. Based on flow encoding methods explained in above section 5, RBridge identify TRILL data plane diagnostic frames from actual data frames with Hope count expiry. Hop count time expiry messages may be generated for actual data frames as well. However, Hop count expiry message for actual data frames are always sent in-band, as actual payload does not have methods to specify the response delivery method.

CPU of intermediate RBridge that receives OAM frame with Hope count expiry performs following:

Identify whether in-band or out of band response requested.
Construct the IP header accordingly.

Construct the ICMP Time Expiry message as specified in RFC 792 and RFC 4884. RFC 4884 specifies format of ICMP header when including ICMP multipart messages.

Include original TRILL header and diagnostic payload of the original frame as data for ICMP Time Expiry message. Update the length field to reflect the size of the TRILL header and diagnostic payload.
Include following ICMP multipart extensions

Version

Set the code to Path Trace Response (3)

Nickname of the RBridge

Information of the ingress interface (speed, state, slot, port)

Index of the interface where frame was received

Nickname of the upstream RBridge the frame was received

Downstream ecmp count

List of Downstream RBridges (nickname, interface index and interface information)

Downstream path this specific payload take \{ RBridge nickname, interface index and interface information\}

Optionally include following ICMP multipart extensions

If VLAN cross connect error detected, set C flag (Cross connect error detected) in the version.

If in-band response was requested or the message was generated due to actual data frame, dispatch the frame to the TRILL data plane with request-originator nickname as the egress RBridge nickname.

If out-of-band response was requested, dispatch the frame to the standard IP forwarding process.

9.2.1.3. Destination RBridge

Processing is identical to section 8.1.1.3

9.3. Multicast Tree Verification (MTV) Message

Multicast Tree Verification messages allow verifying multicast tree integrity and Multicast address pruning. IGMP snooping is widely deployed in Layer 2 networks for restricting forwarding of multicast traffic to unwanted destinations. This is accomplished by pruning the multicast tree such that for specified (S,G,VLAN) or (*,G,VLAN), only required destinations are included in the outgoing interface list. It is possible due to timing or implementation defects,
inaccurate pruning of multicast tree, may occur. Such events lead to incorrect multicast connectivity. Multicast tree verification and Multicast group verification messages are design to detect such multicast connectivity defects. Additionally, these tools can be used for plotting a given multicast tree within the TRILL network.

Multicast tree verification OAM frames are copied to the CPU of every intermediate RBridge that are part of the Multicast tree being verified. Originator of the Multicast Tree verification message, specify the scope of RBridges that a response is required. Only, the RBridges listed in the scope field response to the request. Other RBridges silently discard the request. Definition of scope parameter is required to prevent receiving large number of responses. Typical scenario of multicast tree verification or group verification involves verifying multicast connectivity to selected set of end-nodes as opposed to the entire network. Availability of the scope, facilitate narrowing down the focus only to the interested RBridges.

Implementations MAY choose to rate limit CPU bound multicast traffic. As result of rate limiting or due to other congestion conditions, time to time, MTV messages may be discarded by the intermediate RBridges and requester may be required to retransmit the request. Implementations SHOULD narrow the embedded scope of retransmission request only to RBridges that has failed to respond.

9.3.1. Theory of Operation

9.3.1.1. Originator RBridge

User is required at minimum to specify either the multicast trees that needed to be verified or Multicast MAC address and VLAN or VLAN and Multicast destination IP address. Alternatively, for more specific multicast flow verification, user MAY specify more information e.g. source MAC address, VLAN, Destination and Source IP addresses. Implementation, at minimum, must allow user to specify, choice of multicast trees, Destination Multicast MAC address and VLAN that needed to be verified. Although, it is not mandatory, it is highly desired to provide option to specify the scope.

Default parameters MUST be used for unspecified parameters. Please refer to Figure 5 for default payload parameters for MTV message.

Based on user specified parameters, originating RBridge identify the nickname that represent the multicast tree.

Obtain the applicable Hop count value for the selected multicast tree.
Construct the diagnostic payload based on user specified parameters. For overall multicast tree verification message only multicast tree is specified as input. For generic multicast group verification, additional information such as group address is specified. Based on user provided parameters, implementation SHOULD identify whether the request is for overall multicast tree verification or for specific group verification.

For overall multicast tree verification, use well known multicast destination MAC address (TBD_GMAC-1) defined in above section 6.3.1, as the inner destination MAC address of the TRILL frame. Remaining parameters are derived based on default values specified in Figure 5.

Construct ICMP echo request message header and include sequence number and identifier. Identifier and sequence number facilitate the originator to map the response to the correct request.

Version ICMP multipart extension MUST be included.

Code MUST be specified as Multicast Tree Verification Request (7).

Optionally, include following ICMP multipart extensions, where applicable:

- Out-of-band response request
- Out-of-band IP address
- Diagnostic VLAN
- In scope RBridge list.

  - NOTE: Nu field in ICMP extension RBridge scope (section 8.1.) MUST be set to zero, if response required from all RBridges.

Specify the Hop count of the TRILL data frame per user specification. Or utilize the applicable Hop count value, if TRILL Hop count is not being specified by the user.

Dispatch the diagnostic frame to the TRILL data plane for transmission.

RBridge may continue to retransmit, the request at a periodic interval, until a response received or re-transmission count expires. At each new re-transmission sequence number may be
incremented. At each re-transmission, RBRidge may further reduce the scope to the RBRidges it has not received a response.

9.3.1.2. Intermediate RBridge

Intermediate RBridges identify multicast verification frames per the procedure explained in section 6.3. 

CPU of the RBridge validate the frame and analyze the scope RBridge list. If the local RBridge nickname is not specified in the scope list, it will silently discard the frame. If the local RBridge is specified in the scope list, RBridge proceed to 9.3.1.3 for further processing.

9.3.1.3. In scope RBridges

RBridge go through following processing, upon identifying that it’s nickname is specified in the scope RBridge list.

Identify whether in-band or out of band response requested. Construct the IP header accordingly.

Construct the ICMP echo response message as specified in RFC 792.

Include TRILL header and diagnostic payload of the received OAM message as data of the ICMP response message.

Include following ICMP multipart extensions

Version, update the code as Multicast Tree Verification Response (8)

Nickname of the RBridge

Name of the ingress interface frame was received

Interface index where frame was received

Nickname of the upstream RBridge the frame was received

Downstream leaf node count

Leaf RBridge list (RBridge nickname, interface index and interface name)
Optionally, if VLAN cross connect error detected, then set C flag (cross connect error) in the versions extension.

If in-band response was requested, dispatch the frame to the TRILL data plane with request-originator RBRidge nickname as the egress nickname.

If out-of-band response was requested, dispatch the frame to the standard IP forwarding process.

Error Handling:

RBridge MUST generate applicable notification messages if any error such as inner VLAN not available, detected against the OAM message.

9.4. MAC address discovery Message

MAC address discovery message is defined to discover following information

- RBridge nickname where the MAC address is learnt
- Interface Index and Name on which the MAC address is learnt
- Type (i.e. Static, Dynamic, Secure etc.)
- Age of the MAC address
- Virtual Interface Tag (vNTAG)
- Interface Type (Legacy or TRILL Shared)
- DRB on the VLAN (If Applicable)
- AF for the VLAN (If Applicable)
- Time AF operational (If Applicable)

Optionally, an implementation may include the following information

- System MAC address of the device connected to the port with which the MAC address is associated.
- System information, such as name, IP address and location of the device connected to the port with which the MAC address is associated.
The method of obtaining the above optional information is outside the scope of this document. However, implementation may consider link level control protocols such as LLDP for the purpose.

9.4.1. Theory of Operation

There are two possible options to implement MAC address discovery. Either we may define a new MAC-discovery ISIS sub-TLV and use ESADI to propagate the request (similar to the MAC-Reachability TLV [RFC6165]) OR we may use multicast tree verification message and include a ICMP multipart extension to indicate that the message is a MAC discovery message.

Using the ISIS based method has disadvantage of being non real time and subjected to protocol delays. The second method above is independent of any control plane protocol implementation and can be exercised in real-time. Hence, in this document, we propose to utilize second method.

9.4.1.1. Originator RBridge

Use the well known Multicast MAC address described in section 6.3.1, above as the inner destination MAC address of the diagnostic payload. Use the applicable source MAC address and VLAN. Use the diagnostic EthType defined earlier as the EthType. Pad the remainder of the diagnostic payload with zero.

Construct ICMP echo request message and include sequence number and identifier. The sequence number and identifier facilitate the originator to map the response to the correct request.

Construct following ICMP multipart extensions

  o Version
  o Set the OAM code to the MAC address discovery request (9)
  o Indicate that this is a MAC discovery message
  o One or more MAC address to be discovered
  o VLAN ID of MAC addresses (optional)
  o Out-of-band response request (optional)
o Out-of-band IP address (optional)

o In scope RBridge list. If response required from all RBridges, then the Nu count in RBridge scope list MUST be set to zero.

Specify the TTL value of the TRILL data frame to the applicable value.

Set the egress RBridge nickname to the nickname of the multicast tree used for broadcast and unknown unicast.

Dispatch the diagnostic frame to the TRILL data plane for transmission.

An RBridge may continue to retransmit the request at periodic interval until re-transmission count expires. At each new re-transmission sequence number may be incremented. The RBridge scope list of re-transmission messages MUST be pruned to include only the response pending RBridges. It is possible that more than one RBridge has learnt the requested MAC address. Hence the implementation MUST wait until the total wait time expires and SHOULD NOT abort the discovery process on receiving a single response.

9.4.1.2. Receiving RBridges

CPU of Intermediate RBridges receives a copy of the MAC discovery frame through methods explained in section 6.3.2. and 6.3.1.

Receiving in scope RBridges analyze the embedded ICMP multipart extensions to identify whether the request is for MAC discovery.

If the request is for MAC discovery, then the receiving RBridge queries its forwarding database to identify, whether requested MAC address is present with specified VLAN information.

The receiving RBridge generate responses only for identified MAC entries. If there are no matching MAC entries, the receiving RBridge silently discards the MAC discovery request.

If a matching MAC address is found, the receiving RBridge generates a Destination unreachable ICMP message (Type = 3) and code = 12, "Destination host unreachable for type of service". This essentially indicates, it has found the MAC address but has reached the end of the TRILL network where the MAC address is located.
RFC 4884 allow extension of ICMP messages. Only ICMP messages Destination Unreachable, Time Expired and Parameter Problem are currently extensible in RFC 4884 compliant manner. Other messages are only extensible for known payload size and considered non compliant to RFC 4884. For MAC discovery messages there is no requirement to include original data payload. Also response to MAC discovery can contain large amount of MAC address information. Hence, we conclude to utilize Destination unreachable message as opposed to using an ICMP echo response with fixed pay load size.

The receiving RBridge constructs the response as follows:

Construct the IP header based on the requested response type, in-band or out-of-band. For an in-band response, use RBridge in-band IP address. For an out-of-band response, use the provided egress RBridge out-of-band address.

Construct the ICMP Destination Unreachable message per section 4.1 of RFC 4884. Specify, ICMP type=3 and code = 12. Specify the length as zero. (i.e, no data included and ICMP extensions directly follow).

Include the following ICMP multi part extensions;

- nickname of this RBridge. (This is required in the event of out-of-band response to identify the originating RBridge nickname)

- Version

- Code, set to MAC address discovery response (10)

Additionally, include the following ICMP multipart extensions, for each MAC address that was specified in the request and is present in the RBridge forwarding DB:

- Interface Index and Interface Information (Speed,Slot,Port,State) on which MAC address learnt

- Type (i.e. static, Dynamic, Secure etc.)

- Age of the MAC address

- Virtual Interface Identification (vNTAG)

- Interface Type (Legacy or Trill Shared)

- DRB on the VLAN (If Applicable)
o AF for the VLAN (If Applicable)

o Time AF operational (If Applicable)

Optionally an implementation may include the following information:

o The system MAC address of the device connected to the port with which the MAC address is associated.

o System information, such as name, IP address and location of the device connected to the port on which MAC address is associated.

o Information related to this MAC address from the remote device.

If the response size is greater than the maximum MTU size of the outgoing interface, then multiple responses MAY be generated. The final response frame MUST contain ICMP multipart extension Version (C-Type 1) with F (final response) flag set.

The response frame is delivered to the TRILL data plane for in-band response.

If out of band response was requested, the response frame is delivered to the IP protocol stack.

9.5. Address-Binding Verification Message

Virtual machine provisioning is a very common practice in data centers and enterprises. It is normal for virtual machines to move from one physical machine to another physical machine. As a result ARP tables on gateways can be stale and network operators may need to resort to multiple tools to identify the location of a given IP address that is being diagnosed for connectivity. Even if the location of the server that host the given IP address is identified using other tools, additional steps may be required to further identify the RBridge that interface with the physical server.

It is important to have set of tools that allow an operator to quickly and easily identify the physical MAC address associated with a given IP address, or IP addresses associated with a given physical
MAC address. Additionally, it may be required to identify the RBridge that connects to the given IP address. In this section, we present methods to identify MAC address to IP addresses or IP address to MAC address bindings.

Address binding tools presented here need to be exercised from either a router or an RBridge that has IP services enabled on a given VLAN.

There are two different address binding resolutions required

1. MAC address to IP addresses binding

2. IP address to MAC address binding.

We propose to use invARP [RFC 2390] to resolve MAC address to IP address(es) binding and ARP [RFC 826] to resolve IP address to MAC binding information. It is possible a given physical server to host multiple virtual machines (i.e. IP Addresses). Hence, it is expected to receive one or more responses, to an invARP request. However, invARP in its current form is incapable of identifying whether a single multi-homed host or multiple virtual hosts. At the time of RFC 2390 and original ARP standard RFC 892 were written, virtual machine concept did not exist. Hence, these protocols in its current form do not include virtual machine identifiers such as vNTAGs. This lapse of identification of virtual machines, make troubleshooting of large virtual machine networks, with dynamic server allocation, very difficult. Hence, we propose to extend, ARP [RFC 892] and invARP [RFC 2390], protocol to carry, virtual machine identification tags.

Upon discovery of MAC address or identification that a given MAC address is associated with a valid IP addresses, user may employ the locator utilities listed in section 9.6. to identify the corresponding RBridge and associated interface information. Alternatively, implementation may support ARP response snooping with extension explained in 9.5.1 to encode RBridge and location information into ARP or invARP responses.

9.5.1. Extension to ARP and invARP

RFC 2390 presents methods to discover protocol address associated with a given hardware address. In this section we propose methods to extend RFC 2390 and RFC 892 to encode additional virtual interface tag information and device information that may facilitate identifying physical machine locations.
It is important the extensions proposed in the standard are transparent to current implementations.

Figure 27, below, depicts the format of an ARP/invARP frame with the proposed extensions embedded.

ARP frame as defined in RFC 892 and RFC 2390 has a fixed structure and include only the length fields for addresses. Implementations index into these fix address fields and do not check the total length of the response frame as part of validation. Hence, we propose to include the extensions at the end after the target protocol address. Implementations that do not support the new extensions will safely ignore these values.

We expect additional identification information carried in ARP and invARP to be limited. Furthermore, these, identification information have compact and deterministic size. Hence, we propose not to use explicit, length identification field, instead derive the length of the value field implicitly, based on the class and class types defined below. ARP and invARP follow identical encoding structures.
ARP Extension encoding structure and proposed extensions are presented in this section. We propose a compact structure for ARP encoding. In Figure 27 "Class" identifies the Object Class and the "Class Type" (c-Type) within the class identify specific data element within the object class. C-Type implicitly indicates the size of the object. The encoded object size MUST NOT exceed the implied size of the corresponding Class and c-Type.

Figure 26  Encoding of ARP and invARP

9.5.1.1. Encoding ARP-invARP extensions
Class : (1 octet). Define to identify the Object Class.

C-Type : (1 octet). Define Object type within Object class.

Object : (Variable octet, depends on the Class and C-Types)

<table>
<thead>
<tr>
<th>Class</th>
<th>C-Type</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>vNTAG</td>
<td>vNTAG of the interface</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>RBridge</td>
<td>TRILL RBridge nickname</td>
</tr>
<tr>
<td>2</td>
<td>ifindex</td>
<td></td>
<td>ifindex of RBridge interface ARP</td>
</tr>
<tr>
<td>3</td>
<td>Slot</td>
<td></td>
<td>Slot id of RBridge interface ARP</td>
</tr>
<tr>
<td>4</td>
<td>Port</td>
<td></td>
<td>Port id of RBridge interface ARP</td>
</tr>
</tbody>
</table>

Figure 27  Encoding of ARP Extensions

Figure 28, above, presents Class, c-Type and application definitions. vNTAG, rBridge, Slot and Port are each 2 octets in length. The length of ifindex is 4 octets. All of the above extensions are optional. vNTAG is inserted by the end station that is responding to the ARP request. All other fields are inserted by the TRILL RBridge that interface with the end-station and implement ARP response snooping. ARP response snooping is similar to Dynamic ARP inspections, implemented by many major vendors. Dynamic ARP inspection validates the Source IP address of ARP response against known IP addresses to prevent ARP cache poisoning by rogue stations. ARP response snooping, on the other hand, intercepts ARP response frames and inserts required fields as defined
in this standard. Implementations may extend the dynamic ARP inspection framework to implement ARP response snooping.

In the interim, most end stations and servers may not insert the proposed vNTAG information. Hence, optionally, ARP response snooping, process on TRILL RBridge, MAY insert vNTAG information on behalf of the end station or server.

9.6. End-Station Attachment Point Discovery

In traditional deployments, end stations and servers were relatively static in their locations. As a result localizing a fault was relatively easier.

The virtual machine concept is an increasing trend in Datacenter and large enterprises. Dynamic load balancing policies of Virtual infrastructure, based on various load balancing policies, move virtual machines between different physical servers. This dynamic motion of virtual machines causes difficulty in associating a given virtual server to a RBridge. As a result, localizing a fault is a difficult task and requires use of multiple applications. Some virtual machine deployments utilize a single MAC address to represent all the virtual servers in a single physical server. Hence, it is important, to identify both the physical attachment point and the virtual segment information, such as VLAN and Virtual Tags.

ARP/invARP extensions presented above facilitate discovery of the attachment information, however, some implementation may face scaling issues due to the large number of ARP requests. An alternative method is presented below.

The End-Station attachment Point Discovery methods presented here, allow discovering, RBridge, interface information, VLAN, virtual Tags, etc, associated with a given IP Address.

The End-Station attachment Point Discovery is a two step process. However implementations may present a single user interface that combines both the steps.

Step 1: Utilize ARP to discover the MAC address associated with the specified IP address. Identify the ingress RBridge nickname by analyzing the TRILL header and identify the VLAN information based on the inner VLAN.

Step 2: Utilize MAC discovery methods explained above to discover, interface and virtual Tag information associated with the MAC.
address discovered in above Step 1. Implementation SHOULD narrow the scope of the MAC discovery to include only the RBridge and VLAN discovered in step 1.

9.7. DRB and AF Discovery

The TRILL Base Protocol standard [RFC 6325] specifies support for multi-access legacy network and shared segments between TRILL and legacy devices. Legacy networks ensure loop free forwarding via the IEEE 802.1D (Spanning Tree) protocol. RFC 6325 and RFC 6327 specify loop prevention methods in mixed environments where the TRILL network borders with a legacy multi-access network. RFC 6325 also provide methods for load splitting of native traffic in to the TRILL network. These are accomplished by having a single Designated RBridge (DRB) for a given LAN segment which designates an Appointed Forwarder (AF) for each VLAN on the segment to ingress and egress traffic originating and destined to and from the legacy network.

Based on network dynamics, configurations, and failures, DRB and/or AF designation may change from time to time. Hence, discovery of DRB and AF is very important to effectively troubleshoot network connectivity problems that involve TRILL and legacy networks connected via non P2P TRILL interfaces.

DRB-AF discovery message has three variations.

1. All DRB discovery

2. All AF discovery

3. VLAN,AF discovery

Above messages are identified with a unique TRILL OAM message code (section 8.).

DRB-AF discovery messages allow for identifying the following parameters:

- Nickname of the DRB
- STP Root Bridge identifier
- Up time of AF (if responder is the AF)
- Up time of DRB (if Responder is DRB)
- Enabled VLAN List
The above parameters are encoded in the response message via ICMP multipart extensions (section 8).

9.7.1. Theory of Operation

DRB-AF discovery message utilize same addressing and format as the MAC discovery message (Section 9.4.)

9.7.1.1. Originator RBridge

Follow the steps specified in section 9.4.1.1., with the following exceptions:

Specify the message as one of the DRB-AF messages.

If the message is VLAN, AF discovery message, then include the interest VLAN list.

9.7.1.2. Receiving RBridge

Follow the processing steps specified in section 9.4.1.2. with the following exceptions:

If RBridge is in the scope list or All-RBridge scope is specified, then the RBridge processes the message as follows:

If the message is DRB discovery message then the receiving RBridge include the following information:

- Response code set to DRB discovery response (12)
- Nickname of the DRB
- Nickname of AF of the specified VLAN
- STP Root Bridge identifier
- DRB Life time
- Enabled VLAN List
- Announcing VLAN List
- DRB State
- Pseudo Node bypass
- Number of times Designated VLAN change
- AF List (nickname, start VLAN, end VLAN)

If the message is an AF discovery or VLAN, AF discovery message, then the receiving RBridge first validate whether the RBridge is the AF for the specified VLAN list and include following information:

- Response code set AF discover response (14) or AF-VLAN discover response (16)
- Nickname of the DRB
- Nickname of AF of the specified VLAN or AF VLAN-List if VLAN is not specified.
- STP Root Bridge identifier
- AF Life time (i.e. How long has been AF)
- Enabled VLAN List
- Announcing VLAN List
- AF State
- Number of times Designated VLAN change

If RBridge is not the AF for specified VLAN then include ERROR code Not AF (4) (see Figure 23).
If RBridge is AF for only a subset of VLANs specified in the request then include WARNING "AF VLAN list Mismatch" (3) and include the VLAN list that the RBridge is functioning as AF. (Figure 24)

9.8. Notification Messages

Notification messages are generated either due to regular TRILL data frames or TRILL OAM frames. Implementation MUST not generate notification messages on notification messages.

There are 3 types of Notification messages:

- Time Expiry
- Destination Unreachable
- Parameter Problem

Within these Notification messages, error, warning and information ICMP extensions may be included to identify the details of the notification message. Section 4.3. above covers details of encoding Notification messages, section 8.1. covers ICMP extensions.

Time expiry messages are generated when TRILL hope-count field reach to zero. If applicable, It may contain additional error, warning or information extensions.

Destination unreachable notification may be generated for following scenarios; additional scenarios may be added later.

- Egress RBridge nickname unknown
- Inner VLAN does not exist or suspended
- Not the AF for inner VLAN

Parameter Problem notification may be generated for following scenarios; additional scenarios may be added later.

- Invalid RBridge nickname (RBridge nickname is one of the reserved 0xFFC0 - 0xFFFF)
- MTU mismatch
- Invalid VLAN (Reserved VLANs)
- Interface state is not forwarding

10. Monitoring and Reporting

Proactive identification of data plane failures are important part of maintaining Service Level Agreements (SLA). In traditional Layer
2, networks, there is only a single active path to monitor and both multicast and unicast traffic follow identical paths. With TRILL, there are multiple active paths and unicast and multicast traffic take potentially different paths, depending on the flow parameters.

TRILL deployment in a typical data center may have 10’s of 1000 of links and 100’s of R Bridges. In such an environment, there may be large number of active paths between two end points. As an example, assume a topology with 4 R Bridges connected serially with 32 ECMP links at each hop. In the stated example topology, there are 32x32x32=32768 possible paths. Monitoring all of the possible path combinations is not scalable. However, skipping some combination of paths leads to reduce coverage and hence reduced effectiveness of monitoring data. Even if one was brave enough to monitor all of the links, analyzing and diagnosing a problem is quite cumbersome due to the large amount of data. In other words, there must be methods to scale the problem and present information in a more concise manner that is still effective.

In this document we propose to use the "region" concept to partition the network into logical sections. Regions are monitored independently. Detailed sets of monitoring data are distributed throughout the region. A summary set of monitoring data is distributed throughout the network. Network operators can obtain a network health snapshot of the entire network from any R Bridge in the network. Detailed health report of a given region can be obtained from any R Bridge in the region.

An R Bridge associate itself with a region through its interfaces. A given interface can belong to one and only one region. An R Bridge can have multiple interfaces belonging to different regions. Each R Bridge is responsible for collecting monitoring data, organizing the data into regions and advertising the data to its peers. Please see section 10.2, Advertising Policy for details.

In theory a network topology can be any arbitrary graph. In practices, however, it is some set of sub-graphs repeating to construct the overall topology. Each sub-graph or set of sub-graphs can be considered a region for monitoring purpose. The manner in which regions are partitioned is an administrative choice such that;

1. Maximize the fault coverage.
2. Optimize network health data summarization.

As an example consider a typical datacenter topology depicted in Figure 10. Typical datacenter may have multiple Points of
Demarcation (POD)s connected with an aggregation layer. A POD can be considered as a region and may be individually monitored.

```
+o--o+  +o--o+  +o--o+  
|     |     |     |
|     |     |     |
|     |     |     |
+o--o+  +o--o+  +o--o+  
```

Region Rm

```
+~~+ RB ~ RB  +o--o+  +o--o+  +o--o+  
|     |     |     |     |     |     |
|     |     |     |     |     |     |
|     |     |     |     |     |     |
```

Region R1  . . Region Rn

```
+~~+ RB ~ RB  +o--o+  +o--o+  +o--o+  
|     |     |     |     |     |     |
|     |     |     |     |     |     |
|     |     |     |     |     |     |
```

Figure 29  Example of "regions"

10.1. Data categories

There are 3 categories of monitoring data. They are, Summary Category, Detail Category and Vendor Specific Category. The Summary and Detail categories are mandatory. That is, every RBridge that is compliant to this standard and support Monitoring, MUST support all the elements defined under the Summary and Detail categories. The Vendor specific Category is optional. Vendor specific data elements are only available within the region. An RBridge that does not understand the Vendor specific data elements forward them to neighboring RBridges per Advertising Policies define in section 10.2. Individual data elements and structure of encoding Summary, Detail and Vendor specific categories are presented in sections 10.3. - 10.5. .
10.2. Advertising Policy

Each RBridge is responsible for advertising monitoring data to the OAM capable neighbors.

Different interfaces on an RBridge can belong to different regions. However a given interface can belong to one and only one region. As a result a given RBridge may receive data from multiple regions. Each RBridge is responsible for advertising proper data categories over a given interface to the neighbor.

Rule 1: No monitoring data are distributed:

- On legacy interfaces
- To neighbors not OAM capable
- When ISIS state is not 2-way
- When monitoring data advertisement is disabled

Rule 2: Distribution of Summary category data:

- Distribute on all OAM capable interfaces
- Do not distribute summary data element of a region back to the originating region. (i.e. do not distribute on to interfaces that have the same region name as the data element)
- Summary data for local region is derived from Detail data. (local summary data is never advertised into the local region per the above rule. However, it is advertised out to other regions the RBridge has interfaces in to)

Rule 3: Distribution of Detail category

- Distributed on OAM capable interfaces
- Region of the data element and region of the interface must match for propagating a data element over an interface (i.e. Do not advertise to other regions)
- Do not advertise data element back in to the originator RBridge.
Each RBridge distributes data at periodic intervals. Each RBridge collects data it has received, analyzes them and redistributes according to the rules specified above. The distribution interval should be appropriately adjusted to not overload ISIS routing operations.

Then Monitoring application is responsible for maintaining the Application specific LSP. We propose to use Generic Application Encoding methods explained in [GenAPP] for distributing Monitoring data. TRILL operates in ISIS Level-1 layer, hence S,D flags defined in [GenAPP] MUST be set to zero.

We propose to obtain specific Application ID [GenAPP][RFC5226] from IANA for the purpose of registering TRILL Monitoring data distribution.

Within the Application ID, context, a series of sub-TLV are defined to carry specified information.

10.3. Summary Category

Then following individual data elements are defined within the summary category.

- Name of the region
- Total number of RBridges in the regions
- Total number of TRILL enabled ports in the region
- Percentage of TRILL enabled ports down
- Percentage of TRILL enabled ports oversubscribed
- Maximum number of paths in the largest ECMP in the region

Then following structure encodes each of the data elements within the summary category.
Figure 30  Encoding Summary Category Data

subType : (2 octets) is always 1 for summary category

Region-ID : (4 octets) is unsigned 32 bit integer identifier of the region

L     : ( 1 octet), length of the subsequent field

Region Name : ‘\0’ terminated ASCII string of region name of variable size to maximum of 255 octets.

#Rbridge: (2 octets), number of RBridges in the region

#Ports: (2 octets) Total number of TRILL enabled ports available on this RBridge

#Up Ports: (2 octets) Total number of TRILL enabled ports that are operationally up.

#OSPorts : (2 octets) Total number of TRILL enabled ports that are oversubscribed.
#ErrPorts : (2 octets) Total number of TRILL enabled ports that are indicating errors.

#DwnPorts : (2 octets) Total number of TRILL enabled ports that are operationally down.

#ECMP : (2 octets) Maximum number of ECMP as seen by this region ISIS routing table.

10.4. Detail Category

Following data elements MUST be present within the detail category.

- Name of the region
- Name of the RBridge
- RBridge up time
- Total number of neighbors
- Total number of TRILL enabled ports in the RBridge
- Total number of TRILL enabled ports Up
- Total number of TRILL enabled ports oversubscribed
- Total number of TRILL enabled ports observing errors
- Maximum number of links in the largest ECMP of the switch
- Port data: Name of each TRILL enabled Port and Port state (Up, oversubscribed, error) and interface index.

- Adjacency Matrix
  - List of {neighbor RBridge nickname and interface index of ports connecting to the neighbor RBridge}.

  NOTE: Interface index in the Adjacency matrix is used as key in to port data to obtain Port name and state.
Figure 31  Encoding Detail Category Data

subType : (2 octets) always 2 for Detail category

RBridge: (2 octets) TRILL RBridge nickname [RFCtrill]
Region-ID : (4 octets) unsigned 32 bit integer identifier of the region

L     : ( 1 octet), length of the subsequent field

Region Name : ‘\0’ terminated ASCII string of region name of variable size to maximum of 255 octets.

Up Time: (8 octets), number of seconds RBridge has been operational. If an RBridge reaches maximum count, it MUST NOT rollover.

#Ports: (2 octets) Total number of TRILL enabled ports available on this RBridge

#Up Ports: (2 octets) Total number of TRILL enabled ports that are operationally up.

#OSPorts : (2 octets) Total number of TRILL enabled ports that are oversubscribed.

#ErrPorts : (2 octets) Total number of TRILL enabled ports that are indicating errors.

#DwnPorts : (2 octets) Total number of TRILL enabled ports that are operationally down.

#ECMP : (2 octets) Maximum number of ECMP as seen by this RBridge ISIS routing table.

subtype-2: (2 octets): Set to 3. Following this sub type is the variable length Port Data. See below for details

subtype-3: (2 octets): Set to 4. Following this sub type is the variable length Adjacency Matrix. See below for details
+----------+
| subType  |            2 octets
+----------+
| RBridge  |
+----------+
| F |                   1 octets
+----------+
| subtype-p|            2 octets
+----------+
| ifindex             | 4 octets
+----------+
| Slot     | Port     |
+----------+
| Speed    | State    |
+----------+

Figure 32  Encoding Port data

subType: (2 octets) Set to 3 for Port Data

RBridge: (2 octets) TRILL RBridge nickname [RFCtrill]

Region-ID: (4 octets) unsigned 32 bit integer identifier of the region

L: (1 octet), length of the subsequent field in octets.

Region Name: '\0' terminated ASCII string of region name of variable size to maximum of 255 octets.

F: (1 octet) Flag. When set, indicates this is the last Port data set from this node. It is possible Port data encoding to exceed MTU size due to large number of interfaces. The F flag allows to for advertising the information in multiple LSP packets.

subtype-p: (2 octets) set to 5 to indicate that this is a single Port entry within subtype 3. SubType 5 MUST always be embedded with subtype 3. Within subtype 3 there can be multiple subtype 5, one for each port entry.

Ifindex: (4 octets) 32 bit unsigned integer, used as key to port data advertised.

Slot  (2 octets): Slot number

Port   (2 octets): Port number
Speed (2 octets) : Speed in 100Mbps. Zero (0) indicates port speeds less than 100Mbps.

State (2 octets) : Represent the state of the port.

0: Down - no errors
1: Disable
2: Forwarding-no errors
3: Down - errors
4: Forwarding - errors
5: Forwarding - oversubscribed
6: Link Monitoring disable
All other values reserved.

Figure 33   Encoding Adjacency Matrix

subType : (2 octets) set to 4 for Adjacency Matrix

RBridge: (2 octets) TRILL RBridge nickname [RFCtrill]

Region-id : (4 octets) unsigned 32 bit integer identifier of the region

L : ( 1 octet), length of the region name in octets
Region Name : ’\0’ terminated ASCII string of region name of variable size to a maximum of 255 octets.

F : (1 octet) Flag. When set, indicates this is the last Port data set from this node. It is possible Port data encoding to exceed MTU size due to large number of interfaces. The F flag allows to for advertising the information in multiple LSP packets.

subtype-a: (2 octets) set to 6 to indicates a single adjacency entry within subtype 4. SubType 6 MUST always be embedded with subtype 4. Within subtype 4, there can be multiple subtype 6, one for each adjacency.

nrBRIDGE : (2 octets), nickname of the next hop RBridge

#ports : (2 octets), total number of parallel links from RBridge to nrBRIDGE

Ifindex : (4 octets) 32 bit unsigned integer, used as key to port data advertised.

10.5. Vendor Specific Category

Vendors may specify additional data elements to be distributed as part of the monitoring data suite. All vendor specific data elements MUST contain the regions name and follow the structure defined below.
Figure 34  Encoding Vendor specific category Data

subType : (2 octets) set to 250 for Vendor specific category

RBridge: (2 octets) TRILL RBridge nickname [RFCtrill]

Region-ID : (4 octets) unsigned 32 bit integer identifier of the region

L     : (1 octet), length of the region name in octets

Region Name : ‘\0’ terminated ASCII string of region name of variable size to maximum of 255 octets.

Vendor OUI : 3 octets of IEEE vendor OUI. Right justified. Most significant octet in network byte order is set to zero and ignored on receipt.

Vendor specific information : variable size and vendor dependent.

11. Security Considerations

Security considerations are under investigation.
12. IANA Considerations

12.1. IANA considerations

Following IANA considerations are required

12.1.1. ICMP Extensions

Request IANA to assign new Class-Num for TRILL OAM ICMP extensions.

Request to form a sub-registry under ICMP extensions to include c-types defined in this document and allocate future requests. Currently c-types 1-20 are defined in section 8.1.

12.1.2. ARP Extensions

Request IANA to form a new registry to allocate ARP extensions defined in section 9.5.1. Class-Num allocated within ARP extensions are allocated by IANA on first come first serve basis. C-type within a given Class-Num are defined by owners of the Class-Num and sub-registry MUST be established within ARP extensions.

12.1.3. Well known Multicast MAC

Request IETF authority to allocate one of the TRILL allocated Multicast MAC address (01-80-C2-00-00-43 to 01-80-C2-00-00-4F) for the purpose.

12.2. IEEE Registration Authority Consideration

Well known unicast MAC address for the purpose of identifying OAM frames.

Well known unciast MAC address for the purpose of identifying certain OAM frames.

EthType <TBD> for the purpose of identifying OAM frames.

13. References

13.1. Normative References

13.2. Informative References

[RFC792] Postel, J. "Internet Control Message Proctocol (ICMP)", September, 1981.


14. Acknowledgments

Authors wish to thank people who volunteered to review this document and provided comments.

This document was prepared using 2-Word-v2.0.template.dot.
### Appendix A. Reports

#### A.1. Sample Reports

In this section we present sample reports of summary data and sample output of detail data.

#### A.2. Summary Report

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of switches</th>
<th>Max ECMP</th>
<th>Total# Of Ports</th>
<th>% of Up</th>
<th>% of Ports Oversubscribed</th>
<th>Err Ports</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxx</td>
<td>40</td>
<td>16</td>
<td>400</td>
<td>100</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>yyy</td>
<td>8</td>
<td>2</td>
<td>25</td>
<td>75</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>
A.3. Detail Report

Region Name: <xx>

Total Number of Switches in the region: 10
Total Number of Core Ports in the region: 16
Number of Operationally up Core Ports: 14
Number of Oversubscribed Core Ports: 2
Number of Error Core Ports: 0

Maximum Switch Up Time: 15 days: 8 hr: 10 min: 0 sec
Minimum Switch Up Time: 0 days: 0 hr: 1 min: 0 sec

Switch Adjacency Matrix:

(*) oversubscribed Links
(x) down Links
(?) error Links

<table>
<thead>
<tr>
<th>Switch</th>
<th>Next Hop switch</th>
<th>Interfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>S2</td>
<td>eth81, eth8/2(*)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eth10/2(x)</td>
</tr>
<tr>
<td></td>
<td>S3</td>
<td>eth5/1 (?)</td>
</tr>
<tr>
<td>S2</td>
<td>S1</td>
<td>eth4/1, eth4/2, eth3/1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eth3/2(x)</td>
</tr>
<tr>
<td></td>
<td>S4</td>
<td>eth5/2, eth7/1</td>
</tr>
</tbody>
</table>

Authors’ Addresses

Tissa Senevirathne
CISCO Systems
375 East Tasman Drive,
San Jose, CA 95134

Phone: 408-853-2291
Email: tsenevir@cisco.com
Dinesh G Dutt  
CISCO Systems  
3800 Zankar Road  
San Jose, CA 95134  

Email: ddutt@cisco.com

Vishwas Manral  
Hewlett-Packard Co.  
19111 Pruneridge Ave.  
Cupertino, CA  95014

Phone: 408-447-0000  
EMail: vishwas.manral@hp.com