Internet Engineering Task Force (IETF)
Request for Comments: 7275
Category: Standards Track
ISSN: 2070-1721

L. Martini
S. Salam
A. Sajassi
Cisco
M. Bocci
Alcatel-Lucent
S. Matsushima
Softbank Telecom
T. Nadeau
Brocade

June 2014

Inter-Chassis Communication Protocol for Layer 2 Virtual Private Network (L2VPN) Provider Edge (PE) Redundancy

Abstract

This document specifies an Inter-Chassis Communication Protocol (ICCP) that enables Provider Edge (PE) device redundancy for Virtual Private Wire Service (VPWS) and Virtual Private LAN Service (VPLS) applications. The protocol runs within a set of two or more PEs, forming a Redundancy Group, for the purpose of synchronizing data among the systems. It accommodates multi-chassis attachment circuit redundancy mechanisms as well as pseudowire redundancy mechanisms.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7275.

Table of Contents

1. Introduction

Network availability is a critical metric for service providers, as it has a direct bearing on their profitability. Outages translate not only to lost revenue but also to potential penalties mandated by contractual agreements with customers running mission-critical applications that require tight Service Level Agreements (SLAs). This is true for any carrier network, and networks employing Layer 2 Virtual Private Network (L2VPN) technology are no exception. A high degree of network availability can be achieved by employing intra- and inter-chassis redundancy mechanisms. The focus of this document is on the latter. This document defines an Inter-Chassis Communication Protocol (ICCP) that allows synchronization of state and configuration data between a set of two or more Provider Edge nodes (PEs) forming a Redundancy Group (RG). The protocol supports multi-chassis redundancy mechanisms that can be employed on either the attachment circuits or pseudowires (PWs). A formal definition of the term "chassis" can be found in [RFC2922]. For the purpose of this document, a chassis is an L2VPN PE node.

This document assumes that it is normal to run the Label Distribution Protocol (LDP) between the PEs in the RG, and that LDP components will in any case be present on the PEs to establish and maintain pseudowires. Therefore, ICCP is built as a secondary protocol running within LDP and taking advantage of the LDP session mechanisms as well as the underlying TCP transport mechanisms and TCP-based security mechanisms already necessary for LDP operation.

2. Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. ICCP Overview

3.1. Redundancy Model and Topology

The focus of this document is on PE node redundancy. It is assumed that a set of two or more PE nodes are designated by the operator to form an RG. Members of an RG fall under a single administration (e.g., service provider) and employ a common redundancy mechanism towards the access (attachment circuits or access pseudowires) and/or towards the core (pseudowires) for any given service instance. It is possible, however, for members of an RG to make use of disparate redundancy mechanisms for disjoint services. The PE devices may be offering any type of L2VPN service, i.e., Virtual Private Wire Service (VPWS) or Virtual Private LAN Service (VPLS). As a matter of

fact, the use of ICCP may even be applicable for Layer 3 service redundancy, but this is considered to be outside the scope of this document.

The PEs in an RG offer multi-homed connectivity to either individual devices (e.g., Customer Edge (CE), Digital Subscriber Line Access Multiplexer (DSLAM)) or entire networks (e.g., access network). Figure 1 below depicts the model.

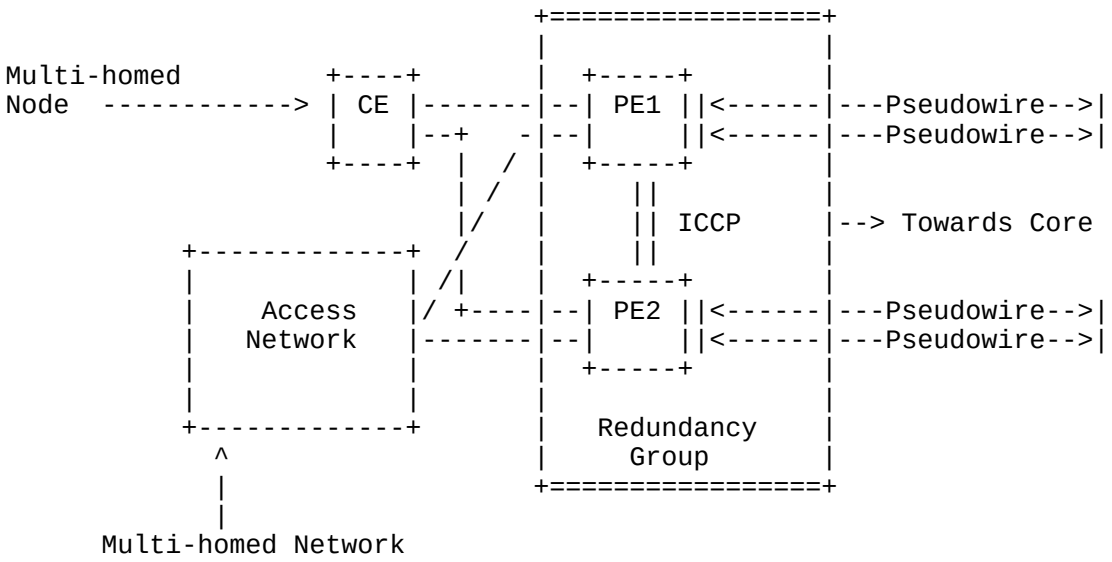


Figure 1: Generic Multi-Chassis Redundancy Model

In the topology shown in Figure 1, the redundancy mechanism employed towards the access node/network can be one of a multitude of technologies, e.g., it could be IEEE 802.1AX Link Aggregation Groups with the Link Aggregation Control Protocol (LACP) or Synchronous Optical Network Automatic Protection Switching (SONET APS). The specifics of the mechanism are outside the scope of this document. However, it is assumed that the PEs in the RG are required to communicate with each other in order for the access redundancy mechanism to operate correctly. As such, it is required that an inter-chassis communication protocol among the PEs in the RG be run in order to synchronize configuration and/or running state data.

Furthermore, the presence of the inter-chassis communication channel allows simplification of the pseudowire redundancy mechanism. This is primarily because it allows the PEs within an RG to run some arbitration algorithm to elect which pseudowire(s) should be in active or standby mode for a given service instance. The PEs can

then advertise the outcome of the arbitration to the remote-end PE(s), as opposed to having to embed a handshake procedure into the pseudowire redundancy status communication mechanism as well as every other possible Layer 2 status communication mechanism.

3.2. ICCP Interconnect Scenarios

When referring to "interconnect" in this section, we are concerned with the links or networks over which Inter-Chassis Communication Protocol messages are transported, and not normal data traffic between PEs. The PEs that are members of an RG may be either physically co-located or geo-redundant. Furthermore, the physical interconnect between the PEs over which ICCP is to run may comprise either dedicated back-to-back links or a shared connection through the packet switched network (PSN), e.g., MPLS core network. This gives rise to a matrix of four interconnect scenarios, as described in the following subsections.

3.2.1. Co-located Dedicated Interconnect

In this scenario, the PEs within an RG are co-located in the same physical location, e.g., point of presence (POP) or central office (CO). Furthermore, dedicated links provide the interconnect for ICCP among the PEs.

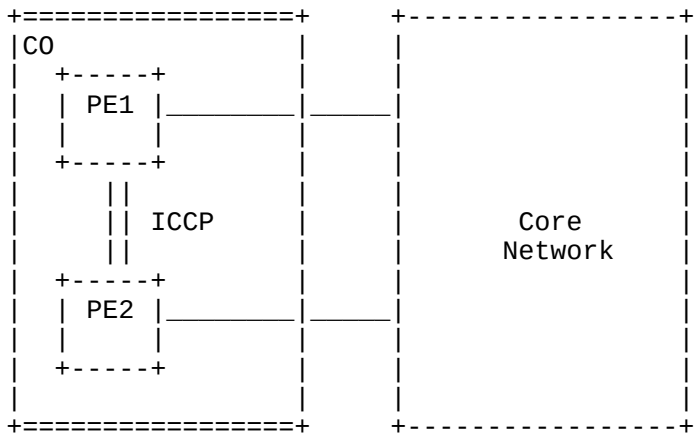


Figure 2: ICCP Co-located PEs Dedicated Interconnect Scenario

Given that the PEs are connected back-to-back in this case, it is possible to rely on Layer 2 redundancy mechanisms to guarantee the robustness of the ICCP interconnect. For example, if the

interconnect comprises IEEE 802.3 Ethernet links, it is possible to provide link redundancy by means of IEEE 802.1AX Link Aggregation Groups.

3.2.2. Co-located Shared Interconnect

In this scenario, the PEs within an RG are co-located in the same physical location (POP, CO). However, unlike the previous scenario, there are no dedicated links between the PEs. The interconnect for ICCP is provided through the core network to which the PEs are connected. Figure 3 depicts this model.

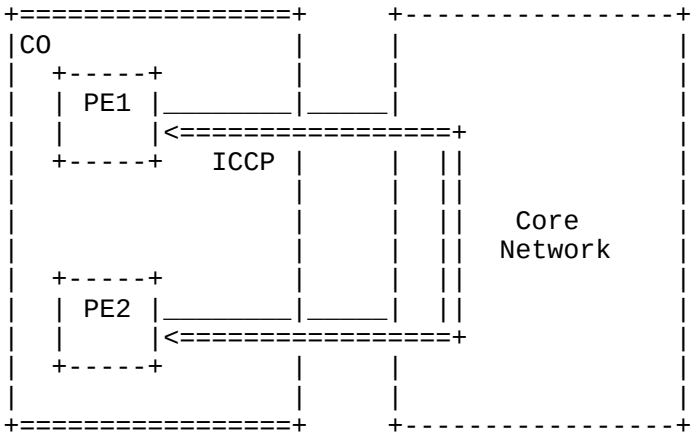


Figure 3: ICCP Co-located PEs Shared Interconnect Scenario

Given that the PEs in the RG are connected over the PSN, PSN Layer mechanisms can be leveraged to ensure the resiliency of the interconnect against connectivity failures. For example, it is possible to employ RSVP Label Switched Paths (LSPs) with Fast Reroute (FRR) and/or end-to-end backup LSPs.

3.2.3. Geo-redundant Dedicated Interconnect

In this variation, the PEs within an RG are located in different physical locations to provide geographic redundancy. This may be desirable, for example, to protect against natural disasters or the like. A dedicated interconnect is provided to link the PEs. This is a costly option, especially when considering the possibility of providing multiple such links for interconnect robustness. The resiliency mechanisms for the interconnect are similar to those highlighted in the co-located interconnect counterpart.

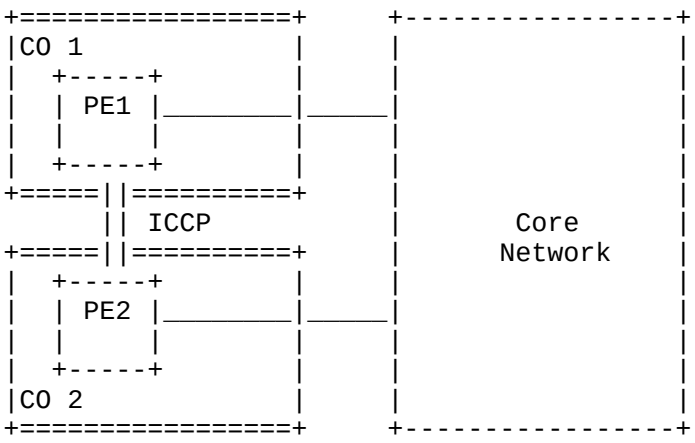


Figure 4: ICCP Geo-redundant PEs Dedicated Interconnect Scenario

3.2.4. Geo-redundant Shared Interconnect

In this scenario, the PEs of an RG are located in different physical locations and the interconnect for ICCP is provided over the PSN network to which the PEs are connected. This interconnect option is more likely to be the one used for geo-redundancy, as it is more economically appealing compared to the geo-redundant dedicated interconnect option. The resiliency mechanisms that can be employed to guarantee the robustness of the ICCP transport are PSN Layer mechanisms, as described in Section 3.2.2 above.

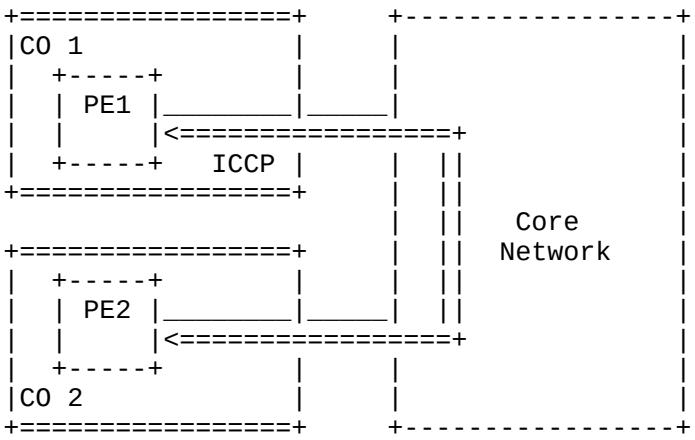


Figure 5: ICCP Geo-redundant PEs Shared Interconnect Scenario

3.3. ICCP Requirements

The requirements for the Inter-Chassis Communication Protocol are as follows:

i. ICCP MUST provide a control channel for communication between PEs in a Redundancy Group (RG). PE nodes may be co-located or remote (refer to Section 3.2 above). Client applications that make use of ICCP services MUST only use this channel to communicate control information and not data traffic. As such, the protocol SHOULD provide relatively low bandwidth, low delay, and highly reliable message transfer.

ii. ICCP MUST accommodate multiple client applications (e.g., multi-chassis LACP, PW redundancy, SONET APS). This implies that the messages SHOULD be extensible (e.g., TLV-based), and the protocol SHOULD provide a robust application registration and versioning scheme.

iii. ICCP MUST provide reliable message transport and in-order delivery between nodes in an RG with secure authentication mechanisms built into the protocol. The redundancy applications that are clients of ICCP expect reliable message transfer and as such will assume that the protocol takes care of flow control and retransmissions. Furthermore, given that the applications will rely on ICCP to communicate data used to synchronize state machines on disparate nodes, it is critical that ICCP guarantees in-order message delivery. Loss of messages or out-of-sequence messages would have adverse effects on the operation of the client applications.

iv. ICCP MUST provide a common mechanism to actively monitor the health of PEs in an RG. This mechanism will be used to detect PE node failure (or isolation from the MPLS network in the case of shared interconnect) and inform the client applications. The applications require that the mechanism trigger failover according to the procedures of the redundancy protocol employed on the attachment circuit (AC) and PW. The solution SHOULD achieve sub-second detection of loss of remote node (˜50-150 msec) in order to give the client applications (redundancy mechanisms) enough reaction time to achieve sub-second service restoration times.

v. ICCP SHOULD provide asynchronous event-driven state update, independent of periodic messages, for immediate notification of client applications' state changes. In other words, the transmission of messages carrying application data SHOULD be on-demand rather than timer-based to minimize inter-chassis state synchronization delay.

vi. ICCP MUST accommodate multi-link and multi-hop interconnects between nodes. When the devices within an RG are located in different physical locations, the physical interconnect between them will comprise a network rather than a link. As such, ICCP MUST accommodate the case where the interconnect involves multiple hops. Furthermore, it is possible to have multiple (redundant) paths or interconnects between a given pair of devices. This is true for both the co-located and geo-redundant scenarios. ICCP MUST handle this as well.

vii. ICCP MUST ensure transport security between devices in an RG. This is especially important in the scenario where the members of an RG are located in different physical locations and connected over a shared network (e.g., PSN). In particular, ICCP MUST NOT accept connections arbitrarily from any device; otherwise, the state of client applications might be compromised. Furthermore, even if an ICCP connection request appears to come from an eligible device, its source address may have been spoofed. Therefore, some means of preventing source address spoofing MUST be in place.

viii. ICCP MUST allow the operator to statically configure members of an RG. Auto-discovery may be considered in the future.

ix. ICCP SHOULD allow for flexible RG membership. It is expected that only two nodes in an RG will cover most of the redundancy applications for common deployments. ICCP SHOULD NOT preclude supporting more than two nodes in an RG by virtue of design. Furthermore, ICCP MUST allow a single node to be a member of multiple RGs simultaneously.

4. ICC LDP Protocol Extension Specification

To address the requirements identified in the previous section, ICCP is modeled to comprise three layers:

i. Application Layer: This provides the interface to the various redundancy applications that make use of the services of ICCP. ICCP is concerned with defining common connection management procedures and the formats of the messages exchanged at this layer; however, beyond that, it does not impose any restrictions

on the procedures or state machines of the clients, as these are deemed application specific and lie outside the scope of ICCP. This guarantees implementation interoperability without placing any unnecessary constraints on internal design specifics.

ii. Inter-Chassis Communication (ICC) Layer: This layer implements the common set of services that ICCP offers to the client applications. It handles protocol versioning, RG membership, Redundant Object identification, PE node identification, and ICCP connection management.

iii. Transport Layer: This layer provides the actual ICCP message transport. It is responsible for addressing, route resolution, flow control, reliable and in-order message delivery, connectivity resiliency/redundancy, and, finally, PE node failure detection. The Transport layer may differ, depending on the Physical Layer of the interconnect.

4.1. LDP ICCP Capability Advertisement

When an RG is enabled on a particular PE, an LDP session to every remote PE in that RG MUST be created, if one does not already exist. The capability of supporting ICCP MUST then be advertised to all of those LDP peers in that RG. This is achieved by using the methods described in [RFC5561] and advertising the "ICCP capability TLV". If an LDP peer supports the dynamic capability advertisement, this can be done by sending a new capability message with the S-bit set for the "ICCP capability TLV" when the first RG is enabled on the PE. If the peer does not support dynamic capability advertisements, then the "ICCP TLV" MUST be included in the LDP initialization procedures in the capability parameter [RFC5561].

4.2. RG Membership Management

ICCP defines a mechanism that enables PE nodes to manage their RG membership. When a PE is configured to be a member of an RG, it will first advertise the ICCP capability to its peers. Subsequently, the PE sends an "RG Connect" message to the peers that have also advertised ICCP capability. The PE then waits for the peers to send their own "RG Connect" messages, if they haven’t done so already. For a given RG, the ICCP connection between two devices is considered to be operational only when both devices have sent and received ICCP "RG Connect" messages for that RG.

If a PE that has sent a particular "RG Connect" message doesn’t receive a corresponding RG Connect (or a Notification message rejecting the connection) from a destination, it will remain in a state of expecting the corresponding "RG Connect" message (or

Notification message). The RG will not become operational until the corresponding "RG Connect" message has been received. If a PE that has sent an "RG Connect" message receives a Notification message rejecting the connection, with a NAK TLV (Negative Acknowledgement TLV) (Section 6.4.1), it will stop attempting to bring up the ICCP connection immediately.

A device MUST reject an incoming "RG Connect" message if at least one of the following conditions is satisfied:

i. the PE is not a member of the RG;

ii. the maximum number of simultaneous ICCP connections that the PE can handle is exceeded.

Otherwise, the PE MUST bring up the connection by responding to the incoming "RG Connect" message with an appropriate RG Connect.

A PE sends an "RG Disconnect" message to tear down the ICCP connection for a given RG. This is a unilateral operation and doesn’t require any acknowledgement from the other PEs. Note that the ICCP connection for an RG MUST be operational before any client application can make use of ICCP services in that RG.

4.2.1. ICCP Connection State Machine

A PE maintains an ICCP Connection state machine instance for every ICCP connection with a remote peer in the RG. This state machine is separate from any Application Connection state machine (Section 4.4.2). The ICCP Connection state machine reacts only to "RG Connect", "RG Disconnect", and "RG Notification" messages that do not contain any "Application TLVs". Actions and state transitions in the Application Connection state machines have no effect on the ICCP Connection state machine.

The ICCP Connection state machine is defined to have six states, as follows:

- NONEXISTENT: This state is the starting point for the state machine. It indicates that no ICCP connection exists and that there’s no LDP session established between the PEs.

- INITIALIZED: This state indicates that an LDP session exists between the PEs but LDP ICCP capability information has not yet been exchanged between them.

- CAPSENT: This state indicates that an LDP session exists between the PEs and that the local PE has advertised LDP ICCP capability to its peer.

- CAPREC: This state indicates that an LDP session exists between the PEs and that the local PE has both received and advertised LDP ICCP capability from/to its peer.

- CONNECTING: This state indicates that the local PE has initiated an ICCP connection to its peer and is awaiting its response.

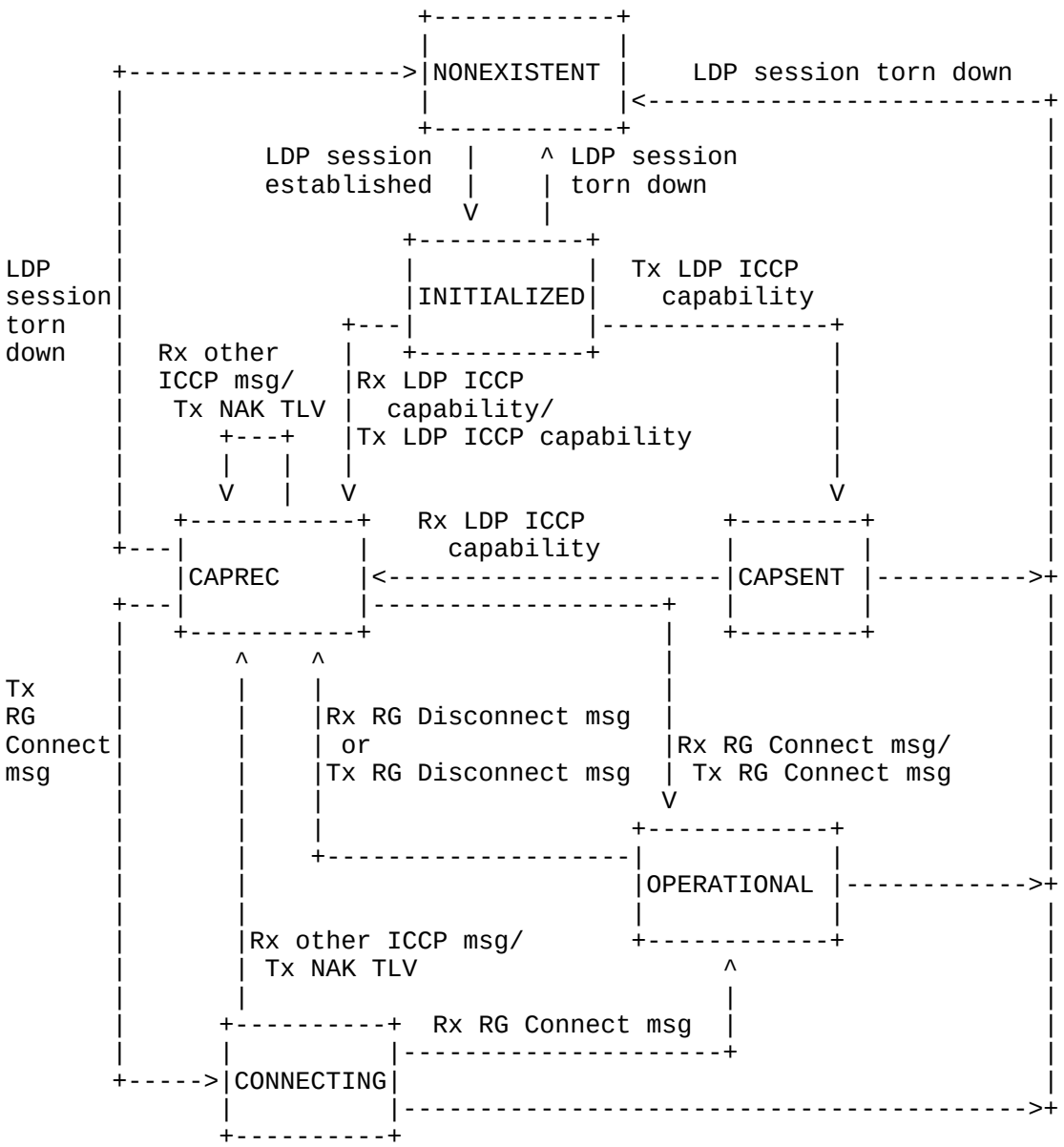
- OPERATIONAL: This state indicates that the ICCP connection is operational.

The state transition table and state transition diagram follow.

ICCP Connection State Transition Table

STATE	EVENT	NEW STATE
NONEXISTENT	LDP session established	INITIALIZED
INITIALIZED	Transmit LDP ICCP capability	CAPSENT
	Receive LDP ICCP capability Action: Transmit LDP ICCP capability	CAPREC
	LDP session torn down	NONEXISTENT
CAPSENT	Receive LDP ICCP capability	CAPREC
	LDP session torn down	NONEXISTENT
CAPREC	Transmit RG Connect message	CONNECTING
	Receive acceptable RG Connect message Action: Transmit RG Connect message	OPERATIONAL
	Receive any other ICCP message Action: Transmit NAK TLV in RG Notification message	CAPREC
	LDP session torn down	NONEXISTENT
CONNECTING	Receive acceptable RG Connect message	OPERATIONAL
	Receive any other ICCP message Action: Transmit NAK TLV in RG Notification message	CAPREC
	LDP session torn down	NONEXISTENT
OPERATIONAL	Receive acceptable RG Disconnect message	CAPREC
	Transmit RG Disconnect message	CAPREC
	LDP session torn down	NONEXISTENT

ICCP Connection State Transition Diagram



4.3. Redundant Object Identification

ICCP offers its client applications a uniform mechanism for identifying links, ports, forwarding constructs, and, more generally, objects (e.g., interfaces, pseudowires, VLANs) that are being protected in a redundant setup. These are referred to as Redundant Objects (ROs). An example of an RO is a multi-chassis link-aggregation group that spans two PEs. ICCP introduces a 64-bit opaque identifier to uniquely identify ROs in an RG. This identifier, referred to as the Redundant Object ID (ROID), MUST match between RG members for the protected object in question; this allows separate systems in an RG to use a common handle to reference the protected entity, irrespective of its nature (e.g., physical or virtual) and in a manner that is agnostic to implementation specifics. Client applications that need to synchronize state pertaining to a particular RO SHOULD embed the corresponding ROID in their TLVs.

4.4. Application Connection Management

ICCP provides a common set of procedures by which applications on one PE can connect to their counterparts on another PE, for the purpose of inter-chassis communication in the context of a given RG. The prerequisite for establishing an Application Connection is to have an operational ICCP RG connection between the two endpoints. It is assumed that the association of applications with RGs is known a priori, e.g., by means of device configuration. ICCP then sends an "Application Connect TLV" (carried in an "RG Connect" message), on behalf of each client application, to each remote PE within the RG. The client may piggyback application-specific information in that "Connect TLV", which, for example, can be used to negotiate parameters or attributes prior to bringing up the actual Application Connection. The procedures for bringing up the Application Connection are similar to those of the ICCP connection: an Application Connection between two nodes is up only when both nodes have sent and received "RG Connect" messages with the proper "Application Connect TLVs". A PE MUST send a Notification message to reject an Application Connection request if one of the following conditions is encountered:

i. the application doesn’t exist or is not configured for that RG;

ii. the Application Connection count exceeds the PE’s capabilities.

When a PE receives such a rejection notification, it MUST stop attempting to bring up the Application Connection until it receives a new Application Connection request from the remote PE. This is done by responding to the incoming "RG Connect" message (carrying an "Application Connect TLV") with an appropriate "RG Connect" message (carrying a corresponding "Application Connect TLV").

When an application is stopped on a device or it is no longer associated with an RG, it MUST signal ICCP to trigger sending an "Application Disconnect TLV" (in the "RG Disconnect" message). This is a unilateral notification to the other PEs within an RG and as such doesn’t trigger any response.

4.4.1. Application Versioning

During Application Connection setup, a given application on one PE can negotiate with its counterpart on a peer PE the proper application version to use for communication. If no common version is agreed upon, then the Application Connection is not brought up. This is achieved through the following set of rules:

- If an application receives an "Application Connect TLV" with a version number that is higher than its own, it MUST send a Notification message with a "NAK TLV" indicating status code "Incompatible Protocol Version" and supplying the version that is locally supported by the PE.

- If an application receives an "Application Connect TLV" with a version number that is lower than its own, it MAY respond with an RG Connect that has an "Application Connect TLV" using the same version that was received. Alternatively, the application MAY respond with a Notification message to reject the request using the "Incompatible Protocol Version" code and supply the version that is supported. This allows an application to operate in either backwards-compatible or incompatible mode.

- If an application receives an "Application Connect TLV" with a version that is equal to its own, then the application MUST honor or reject the request based on whether the application is configured for the RG in question, and whether or not the Application Connection count has been exceeded.

4.4.2. Application Connection State Machine

A PE maintains one Application Connection state machine instance per ICCP application for every ICCP connection with a remote PE in the RG. Each application's state machine reacts only to the "RG Connect", "RG Disconnect", and "RG Notification" messages that contain an "Application TLV" specifying that particular application.

The Application Connection state machine has six states, as follows:

- NONEXISTENT: This state indicates that the Application Connection does not exist, since there is no ICCP connection between the PEs.

- RESET: This state indicates that an ICCP connection is operational between the PEs but that the Application Connection has not been initialized yet or has been resent.

- CONNSENT: This state indicates that the local PE has requested initiation of an Application Connection with its peer but has not received a response yet.

- CONNREC: This state indicates that the local PE has received a request to initiate an Application Connection from its peer but has not responded yet.

- CONNECTING: This state indicates that the local PE has transmitted to its peer an "Application Connection" message with the A-bit set to 1 and is awaiting the peer's response.

- OPERATIONAL: This state indicates that the Application Connection is operational.

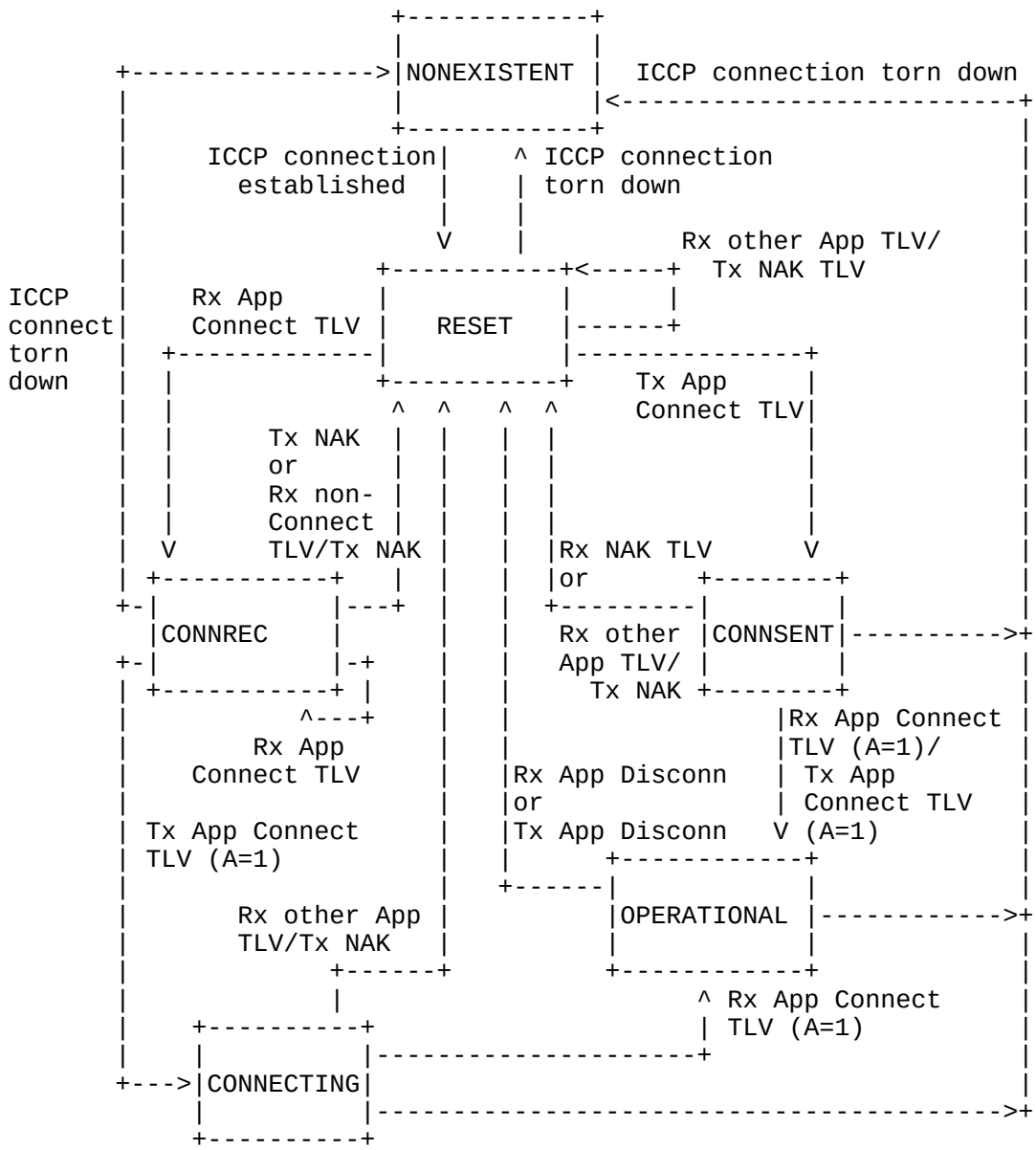
The state transition table and state transition diagram follow.

ICCP Application Connection State Transition Table

STATE	EVENT	NEW STATE
NONEXISTENT	ICCP connection established	RESET
RESET	ICCP connection torn down	NONEXISTENT
	Transmit Application Connect TLV	CONNSENT
	Receive Application Connect TLV	CONNREC
	Receive any other Application TLV Action: Transmit NAK TLV	RESET

CONNSENT	Receive NAK TLV	RESET
	Receive Application Connect TLV with A-bit=1 Action: Transmit Application Connect TLV with A-bit=1	OPERATIONAL
	Receive any other Application TLV Action: Transmit NAK TLV	RESET
	ICCP connection torn down	NONEXISTENT
CONNREC	Transmit NAK TLV	RESET
	Transmit Application Connect TLV with A-bit=1	CONNECTING
	Receive Application Connect TLV	CONNREC
	Receive any Application TLV except Connect Action: Transmit NAK TLV	RESET
	ICCP connection torn down	NONEXISTENT
CONNECTING	Receive Application Connect TLV with A-bit=1	OPERATIONAL
	Receive any other Application TLV Action: Transmit NAK TLV	RESET
	ICCP connection torn down	NONEXISTENT
OPERATIONAL	Receive Application Disconnect TLV	RESET
	Transmit Application Disconnect TLV	RESET
	ICCP connection torn down	NONEXISTENT

ICCP Application Connection State Transition Diagram



4.5. Application Data Transfer

When an application has information to transfer over ICCP, it triggers the transmission of an "Application Data" message. ICCP guarantees in-order and lossless delivery of data. An application may reject a message or a set of one or more TLVs within a message by using the Notification message with a "NAK TLV". Furthermore, an application may implement its own ACK mechanism, if deemed required, by defining an application-specific TLV to be transported in an "Application Data" message. Note that this document does not define a common ACK mechanism for applications.

It is left up to the application to define the procedures to handle the situation where a PE receives a "NAK TLV" in response to a transmitted "Application Data" message. Depending on the specifics of the application, it may be favorable to have the PE that sent the NAK explicitly request retransmission of data. On the other hand, for certain applications it may be more suitable to have the original sender of the "Application Data" message handle retransmissions in response to a NAK. ICCP supports both models.

4.6. Dedicated Redundancy Group LDP Session

For certain ICCP applications, it is required that a fairly large amount of RG information be exchanged in a very short period of time. In order to better distribute the load in a multiple-processor system, and to avoid head-of-line blocking to other LDP applications, initiating a separate TCP/IP session between the two LDP speakers may be required.

This procedure is OPTIONAL and does not change the operation of LDP or ICCP.

A PE that requires a separate LDP session will advertise a separate LDP adjacency with a non-zero label space identifier. This will cause the remote peer to open a separate LDP session for this label space. No labels need to be advertised in this label space, as it is only used for one or a set of ICCP RGs. All relevant LDP and ICCP procedures still apply as described in [RFC5036] and this document.

5. ICCP PE Node Failure / Isolation Detection Mechanism

ICCP provides its client applications a notification when a remote PE that is a member of the RG is no longer reachable. In the case of a dedicated interconnect, this indicates that the remote PE node has failed, whereas in the case of a shared interconnect this indicates that the remote PE node has either failed or become isolated from the MPLS network. This information is used by the client applications to

trigger failover according to the procedures of the redundancy protocol employed on the AC and PW. To that end, ICCP does not define its own Keep-Alive mechanism for the purpose of monitoring the health of remote PE nodes but rather reuses existing fault detection mechanisms. The following mechanisms may be used by ICCP to detect PE node failure:

- Bidirectional Forwarding Detection (BFD)

  Run a BFD session [RFC5880] between the PEs that are members of a given RG, and use that to detect PE node failure. This assumes that resiliency mechanisms are in place to protect connectivity to the remote PE nodes, and hence loss of BFD periodic messages from a given PE node can only mean that the node itself has failed.

- IP Reachability Monitoring

  It is possible for a PE to monitor IP-layer connectivity to other members of an RG that are participating in IGP/BGP. When connectivity to a given PE is lost, the local PE interprets that to mean loss of the remote PE node. This technique assumes that resiliency mechanisms are in place to protect the route to the remote PE nodes, and hence loss of IP reachability to a given node can only mean that the node itself has failed.

It is worth noting here that loss of the LDP session with a PE in an RG is not a reliable indicator that the remote PE itself is down. It is possible, for example, that the remote PE could encounter a local event that would lead to resetting the LDP session, while the PE node would remain operational for traffic forwarding purposes.

6. ICCP Message Formats

This section defines the messages exchanged at the Application and ICC layers.

6.1. Encoding ICC into LDP Messages

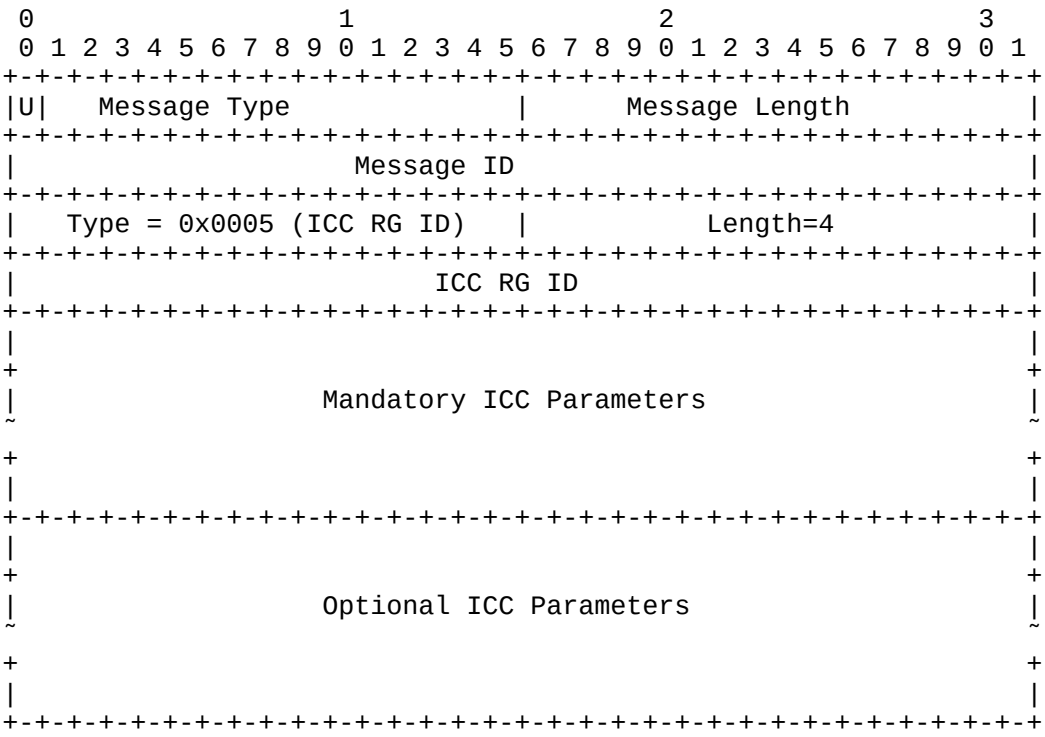
ICCP requires reliable, in-order, stateful message delivery, as well as capability negotiation between PEs. LDP offers all of these features and is already in wide use in the applications that would also require the ICCP protocol extensions. For these reasons, ICCP takes advantage of the already-defined LDP protocol infrastructure.

[RFC5036], Section 3.5 defines a generic LDP message structure. A new set of LDP message types is defined to communicate the ICCP information. LDP message types in the range 0x0700 to 0x070F will be used for ICCP.

Message types have been allocated by IANA; see Section 12 below for details.

6.1.1. ICC Header

Every ICCP message comprises an ICC-specific LDP Header followed by message data. The format of the ICC Header is as follows:



- U-bit

  Unknown message bit. Upon receipt of an unknown message, if U is clear (=0), a notification is returned to the message originator; if U is set (=1), the unknown message is silently ignored. Subsequent sections that define messages specify a value for the U-bit.

- Message Type

  Identifies the type of the ICCP message. Must be in the range 0x0700 to 0x070F.

- Message Length

  2-octet integer specifying the total length of this message in octets, excluding the "U-bit", "Message Type", and "Length" fields.

- Message ID

  4-octet value used to identify this message. Used by the sending PE to facilitate identifying "RG Notification" messages that may apply to this message. A PE sending an "RG Notification" message in response to this message SHOULD include this Message ID in the "NAK TLV" of the "RG Notification" message; see Section 6.4.

- ICC RG ID TLV

  A TLV of type 0x0005, length 4, containing a 4-octet unsigned integer designating the Redundancy Group of which the sending device is a member. RG ID value 0x00000000 is reserved by the protocol.

- Mandatory ICC Parameters

  Variable-length set of required message parameters. Some messages have no required parameters.

  For messages that have required parameters, the required parameters MUST appear in the order specified by the individual message specifications in the sections that follow.

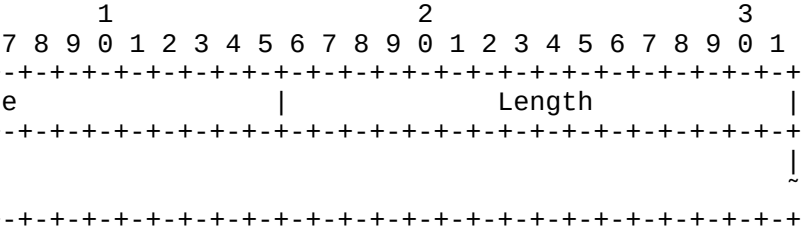
- Optional ICC Parameters

  Variable-length set of optional message parameters. Many messages have no optional parameters.

  For messages that have optional parameters, the optional parameters may appear in any order.

6.1.2. ICC Parameter Encoding

The generic format of an ICC parameter is as follows:



- U-bit

  Unknown TLV bit. Upon receipt of an unknown TLV, if U is clear (=0), a notification MUST be returned to the message originator and the entire message MUST be ignored; if U is set (=1), the unknown TLV MUST be silently ignored and the rest of the message processed as if the unknown TLV did not exist. Subsequent sections that define TLVs specify a value for the U-bit.

- F-bit

  Forward unknown TLV bit. This bit applies only when the U-bit is set and the LDP message containing the unknown TLV is to be forwarded. If F is clear (=0), the unknown TLV is not forwarded with the LDP message; if F is set (=1), the unknown TLV is forwarded with the LDP message. Subsequent sections that define TLVs specify a value for the F-bit. By setting both the U- and F-bits, a TLV can be propagated as opaque data through nodes that do not recognize the TLV.

- Type

  14 bits indicating the ICC Parameter type.

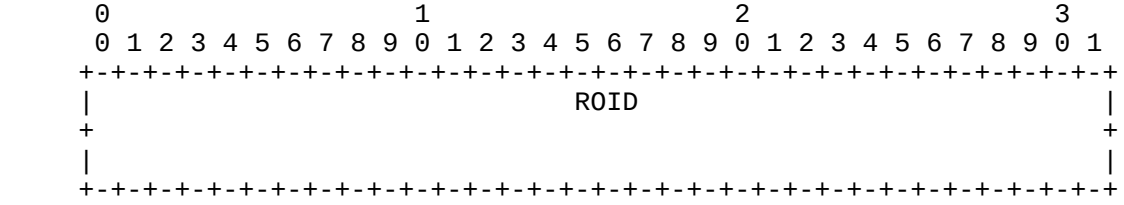
- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- TLV(s): A set of 0 or more TLVs. Contents will vary according to the message type.

6.1.3. Redundant Object Identifier Encoding

The Redundant Object Identifier (ROID) is a generic opaque handle that uniquely identifies a Redundant Object (e.g., link, bundle, VLAN) that is being protected in an RG. It is encoded as follows:



where the ROID is an 8-octet field encoded as an unsigned integer. The ROID value of 0 is reserved.

The ROID is carried within application-specific TLVs.

6.2. RG Connect Message

The "RG Connect" message is used to establish the ICCP RG connection in addition to individual Application Connections between PEs in an RG. An "RG Connect" message with no "Application Connect TLV" signals establishment of the ICCP RG connection, whereas an "RG Connect" message with a valid "Application Connect TLV" signals the establishment of an Application Connection in addition to the ICCP RG connection if the latter is not already established.

An implementation MAY send a dedicated "RG Connect" message to set up the ICCP RG connection and a separate "RG Connect" message for each client application. However, all implementations MUST support the receipt of an "RG Connect" message that triggers the setup of the ICCP RG connection as well as a single Application Connection simultaneously.

A PE sends an "RG Connect" message to declare its membership in a Redundancy Group. One such message should be sent to each PE that is a member of the same RG. The set of PEs to which "RG Connect" messages should be transmitted is known via configuration or an auto-discovery mechanism that is outside the scope of this specification. If a device is a member of multiple RGs, it MUST send separate "RG Connect" messages for each RG even if the receiving device(s) happens to be the same.

The format of the "RG Connect" message is as follows:

i. ICC Header with Message type = "RG Connect Message" (0x0700)

ii. ICC Sender Name TLV

iii. Zero or one "Application Connect TLV"

The currently defined "Application Connect TLVs" are as follows:

- PW-RED Connect TLV (Section 7.1.1)

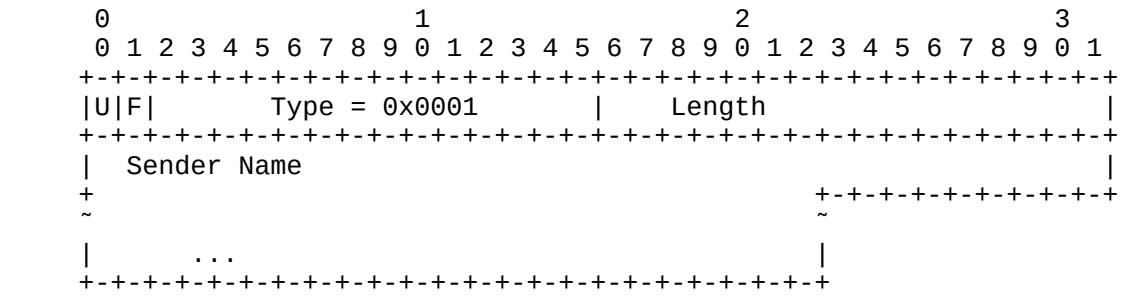
- mLACP Connect TLV (Section 7.2.1)

The details of these TLVs are discussed in Section 7.

The "RG Connect" message can contain zero or one "Application Connect TLV".

6.2.1. ICC Sender Name TLV

The "ICC Sender Name TLV" carries the hostname of the sender, encoded in UTF-8 [RFC3629] format. This is used primarily for the purpose of management of the RG and easing network operations. The specific format is shown below:



- U=F=0

- Type

  Set to 0x0001 (from the ICC parameter name space).

- Length

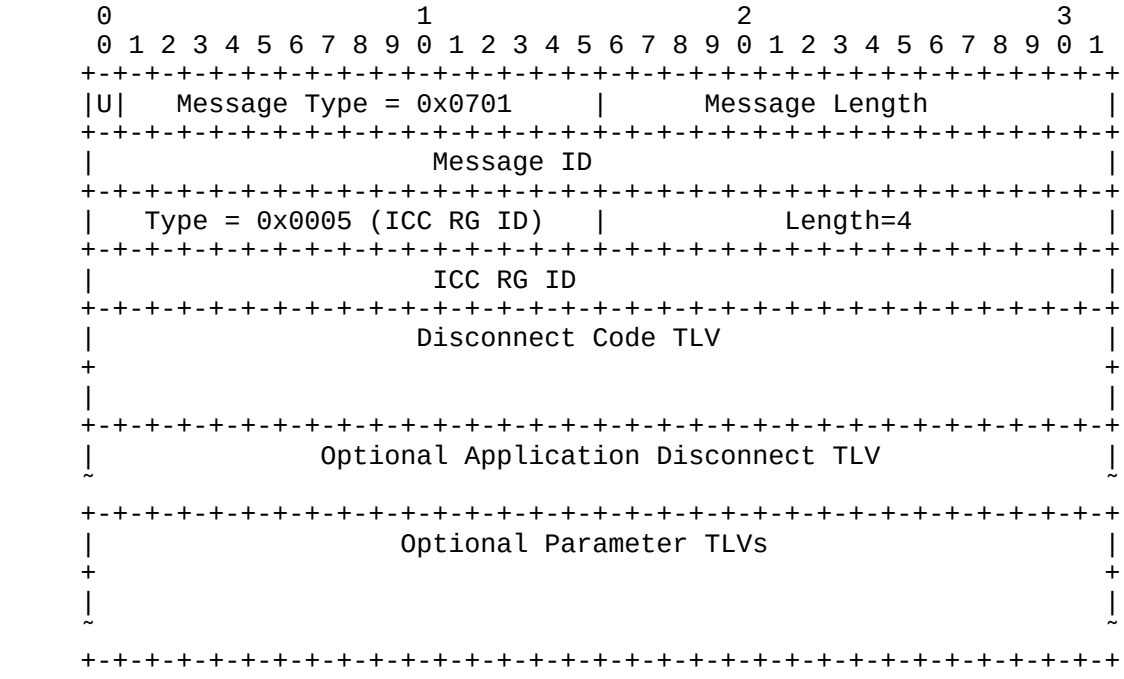
  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Sender Name

  An administratively assigned name of the sending device, encoded in UTF-8 format and limited to a maximum of 80 octets. This field does not include a terminating null character.

6.3. RG Disconnect Message

The "RG Disconnect" message serves a dual purpose: to signal that a particular Application Connection is being closed within an RG or that the ICCP RG connection itself is being disconnected because the PE wishes to leave the RG. The format of this message is as follows:



- U-bit

  U=0

- Message Type

  The message type for the "RG Disconnect" message is set to 0x0701.

- Length

  Length of the TLV in octets, excluding the "U-bit", "Message Type", and "Message Length" fields.

- Message ID

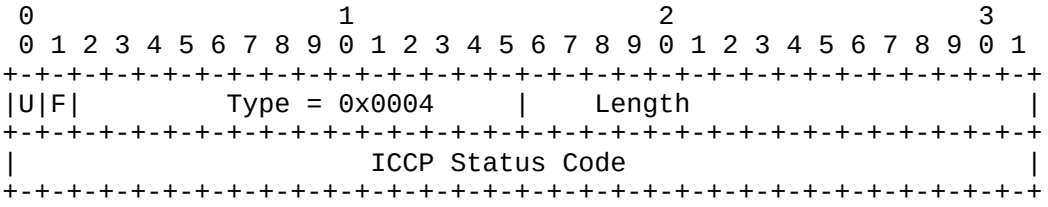
  Defined in Section 6.1.1 above.

- ICC RG ID

  Defined in Section 6.1.1 above.

- Disconnect Code TLV

  The format of this TLV is as follows:



  - U-bit and F-bit

    Both are set to 0.

  - Type

    Set to "Disconnect Code TLV" (0x0004).

  - Length

    Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

  - ICCP Status Code

    A status code that reflects the reason for the disconnect message. Allowed values are "ICCP RG Removed" and "ICCP Application Removed from RG".

- Optional Application Disconnect TLV

  Zero or one "Application Disconnect TLV" (defined in Sections 7.1.2 and 7.2.2). If the "RG Disconnect" message has a status code of "RG Removed", then it MUST NOT contain any "Application Disconnect TLVs", as the sending PE is signaling that it has left the RG and thus is disconnecting the ICCP RG connection with all associated client Application Connections. If the message has a status code of "Application Removed from RG", then it MUST contain exactly one "Application Disconnect TLV", as the sending PE is only tearing down the connection for the specified application. Other applications, and the ICCP RG connection, are not to be affected.

- Optional Parameter TLVs

  None are defined for this message in this document. This is specified to allow for future extensions.

6.4. RG Notification Message

A PE sends an "RG Notification" message to indicate one of the following: to reject an ICCP connection, to reject an Application Connection, to reject an entire message, or to reject one or more TLVs within a message. The Notification message MUST only be sent to a PE that is already part of an RG.

The "RG Notification" message MUST only be used to reject messages or TLVs corresponding to a single ICCP application. In other words, there is a limit of at most a single ICCP application per "RG Notification" message.

The format of the "RG Notification" message is as follows:

i. ICC Header with Message type = "RG Notification Message" (0x0702)

ii. Notification Message TLVs

The currently defined Notification message TLVs are as follows:

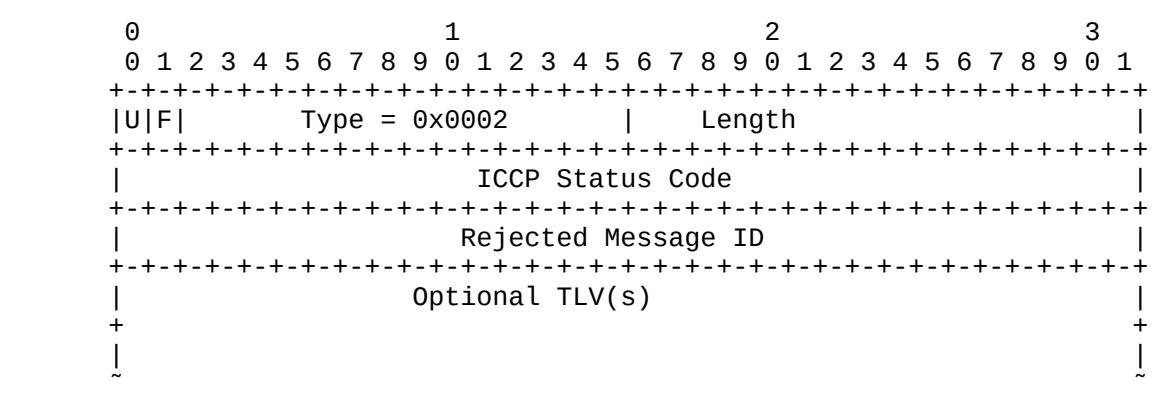
i. ICC Sender Name TLV

ii. Negative Acknowledgement (NAK) TLV

6.4.1. Notification Message TLVs

The "ICC Sender Name TLV" uses the same format as the format used in the "RG Connect" message and was described above.

The "NAK TLV" is defined as follows:



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to "NAK TLV" (0x0002).

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- ICCP Status Code

  A status code that reflects the reason for the "NAK TLV". Allowed values are as follows:

  i. Unknown ICCP RG (0x00010001)

     This code is used to reject a new incoming ICCP connection for an RG that is not configured on the local PE. When this code is used, the "Rejected Message ID" field MUST contain the message ID of the rejected "RG Connect" message.

ii. ICCP Connection Count Exceeded (0x00010002)

This is used to reject a new incoming ICCP connection that would cause the local PE's ICCP connection count to exceed its capabilities. When this code is used, the "Rejected Message ID" field MUST contain the message ID of the rejected "RG Connect" message.

iii. ICCP Application Connection Count Exceeded (0x00010003)

This is used to reject a new incoming Application Connection that would cause the local PE's ICCP connection count to exceed its capabilities. When this code is used, the "Rejected Message ID" field MUST contain the message ID of the rejected "RG Connect" message and the corresponding "Application Connect TLV" MUST be included in the "Optional TLV".

iv. ICCP Application not in RG (0x00010004)

This is used to reject a new incoming Application Connection when the local PE doesn't support the application or the application is not configured in the RG. When this code is used, the "Rejected Message ID" field MUST contain the message ID of the rejected "RG Connect" message and the corresponding "Application Connect TLV" MUST be included in the "Optional TLV".

v. Incompatible ICCP Protocol Version (0x00010005)

This is used to reject a new incoming Application Connection when the local PE has an incompatible version of the application. When this code is used, the "Rejected Message ID" field MUST contain the message ID of the rejected "RG Connect" message and the corresponding "Application Connect TLV" MUST be included in the "Optional TLV".

vi. ICCP Rejected Message (0x00010006)

This is used to reject an "RG Application Data" message, or one or more TLVs within the message. When this code is used, the "Rejected Message ID" field MUST contain the message ID of the rejected "RG Application Data" message.

vii. ICCP Administratively Disabled (0x00010007)

This is used to reject any ICCP messages from a peer from which the PE is not allowed to exchange ICCP messages due to local administrative policy.

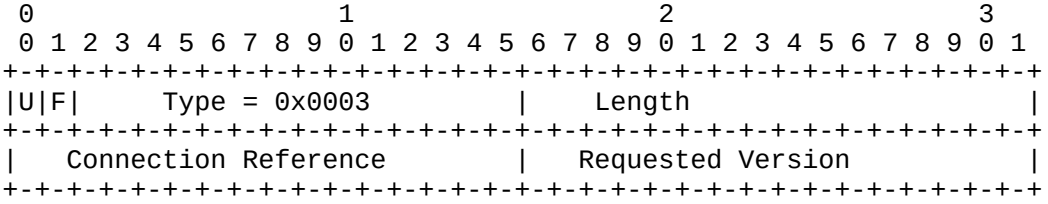
- Rejected Message ID

If non-zero, a 4-octet value that identifies the peer message to which the "NAK TLV" refers. If zero, no specific peer message is being identified.

- Optional TLV(s)

A set of one or more optional TLVs. If the status code is "Rejected Message", then this field contains the TLV or TLVs that were rejected. If the entire message is rejected, all of its TLVs MUST be present in this field; otherwise, the subset of TLVs that were rejected MUST be echoed in this field.

If the status code is "Incompatible Protocol Version", then this field contains the original "Application Connect TLV" sent by the peer, in addition to the "Requested Protocol Version TLV" defined below:



- U-bit and F-bit

Both are set to 0.

- Type

Set to 0x0003 for "Requested Protocol Version TLV".

- Length

Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Connection Reference

  Set to the "Type" field of the "Application Connect TLV" that was rejected because of incompatible version.

- Requested Version

  The version of the application supported by the transmitting device. For this version of the protocol, it is set to 0x0001.

6.5. RG Application Data Message

The "RG Application Data" message is used to transport application data between PEs within an RG. A single message can be used to carry data from only one application. Multiple Application TLVs are allowed in a single message, as long as all of these TLVs belong to the same application. The format of the "Application Data" message is as follows:

i. ICC Header with Message type = "RG Application Data Message" (0x0703)

ii. Application-specific TLVs

The details of these TLVs are discussed in Section 7. All application-specific TLVs in one "RG Application Data" message MUST belong to a single application but MAY reference different ROs.

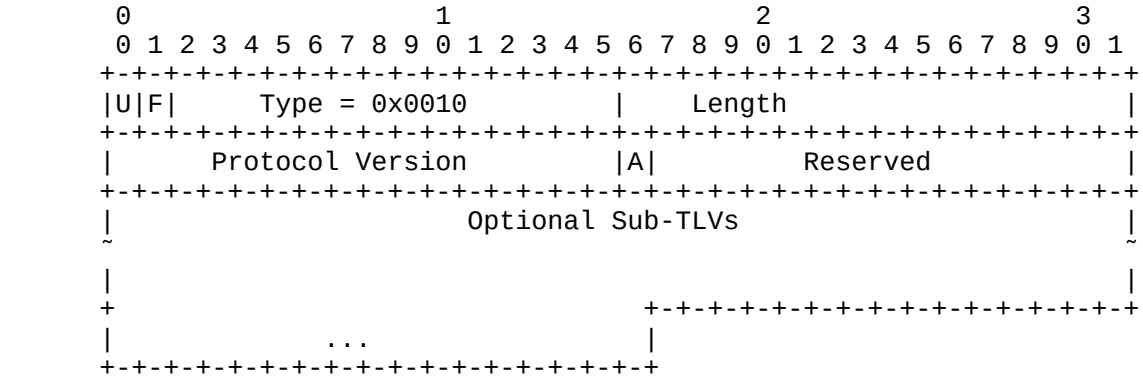
7. Application TLVs

7.1. Pseudowire Redundancy (PW-RED) Application TLVs

This section discusses the "ICCP TLVs" for the Pseudowire Redundancy application.

7.1.1. PW-RED Connect TLV

This TLV is included in the "RG Connect" message to signal the establishment of a PW-RED Application Connection.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0010 for "PW-RED Connect TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Protocol Version

  The version of this particular protocol for the purposes of ICCP. This is set to 0x0001.

- A-bit

  Acknowledgement bit. Set to 1 if the sender has received a "PW-RED Connect TLV" from the recipient. Otherwise, set to 0.

- Reserved

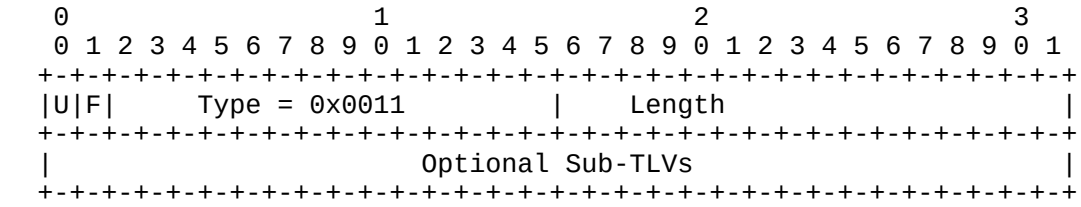
  Reserved for future use.

- Optional Sub-TLVs

  There are no optional sub-TLVs defined for this version of the protocol. This document does not impose any restrictions on the length of the sub-TLVs.

7.1.2. PW-RED Disconnect TLV

This TLV is used in an "RG Disconnect" message to indicate that the connection for the PW-RED application is to be terminated.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0011 for "PW-RED Disconnect TLV".

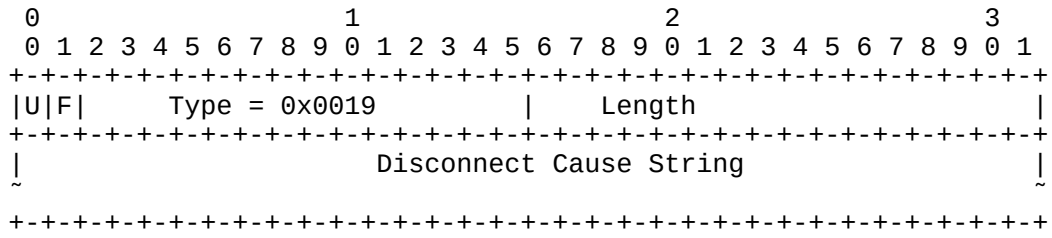
- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Optional Sub-TLVs

  The only optional sub-TLV defined for this version of the protocol is the "PW-RED Disconnect Cause TLV" defined in Section 7.1.2.1.

7.1.2.1. PW-RED Disconnect Cause TLV



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0019 for "PW-RED Disconnect Cause TLV".

- Length

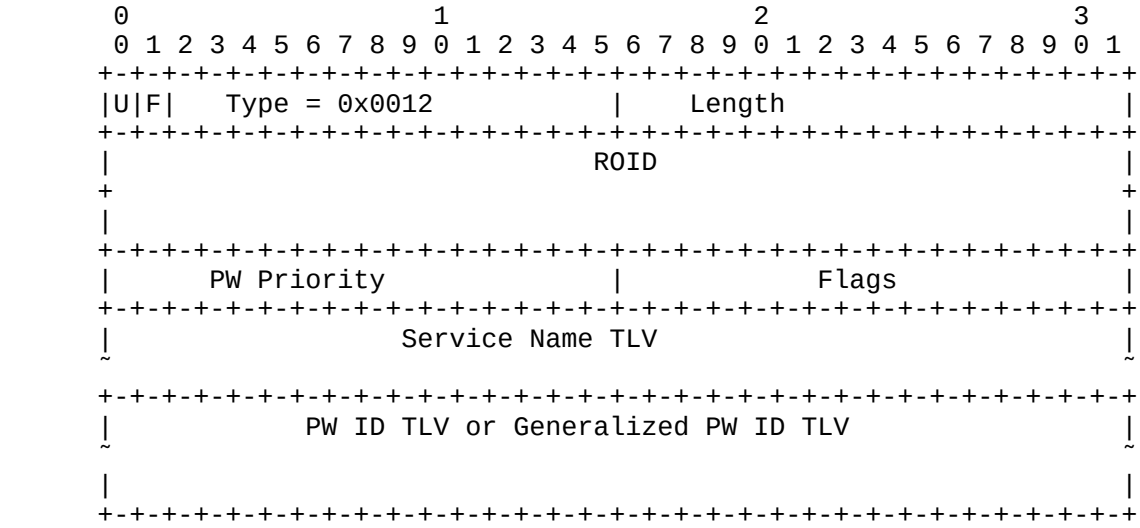
  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Disconnect Cause String

  Variable-length string specifying the reason for the disconnect, encoded in UTF-8 format. The string does not include a terminating null character. Used for network management.

7.1.3. PW-RED Config TLV

The "PW-RED Config TLV" is used in the "RG Application Data" message and has the following format:



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0012 for "PW-RED Config TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- ROID

  As defined in Section 6.1.3.

- PW Priority

  2 octets. Pseudowire Priority. Used to indicate which PW has better priority to go into active state. Numerically lower numbers are better priority. In case of a tie, the PE with the numerically lower identifier (i.e., IP Address) has better priority.

- Flags

  Valid values are as follows:

  i. Synchronized (0x01)

     Indicates that the sender has concluded transmitting all pseudowire configuration for a given service.

  ii. Purge Configuration (0x02)

     Indicates that the pseudowire is no longer configured for PW-RED operation.

  iii. Independent Mode (0x04)

     Indicates that the pseudowire is configured for redundancy using the Independent Mode of operation, per Section 5.1 of [RFC6870].

  iv. Independent Mode with Request Switchover (0x08)

     Indicates that the pseudowire is configured for redundancy using the Independent Mode of operation with the use of the "Request Switchover" bit, per Section 6.3 of [RFC6870].

  v. Master Mode (0x10)

     Indicates that the pseudowire is configured for redundancy using the Master/Slave Mode of operation, with the advertising PE acting as Master, per Section 5.2 of [RFC6870].

  vi. Slave Mode (0x20)

     Indicates that the pseudowire is configured for redundancy using the Master/Slave Mode of operation, with the advertising PE acting as Slave, per Section 5.2 of [RFC6870].

- Sub-TLVs

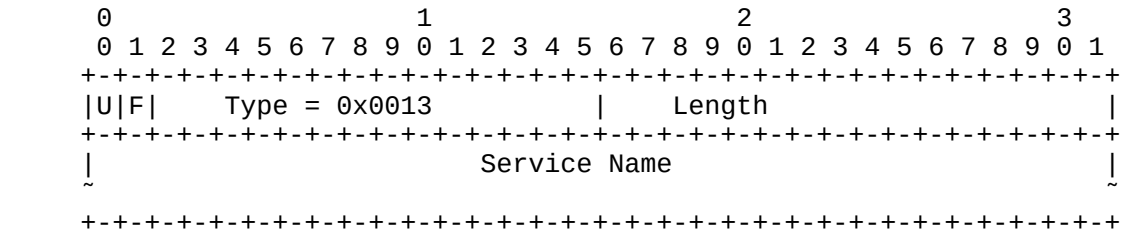
  The "PW-RED Config TLV" includes the following two sub-TLVs:

  i. Service Name TLV

  ii. One of the following: PW ID TLV or Generalized PW ID TLV

  The format of the sub-TLVs is defined in Sections 7.1.3.1 through 7.1.3.3.

7.1.3.1. Service Name TLV



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0013 for "Service Name TLV".

- Length

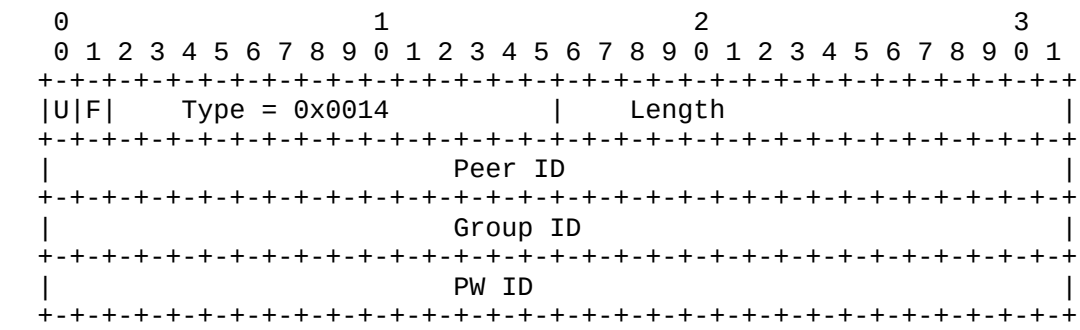
  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Service Name

  The name of the L2VPN service instance, encoded in UTF-8 format and up to 80 octets in length. The string does not include a terminating null character.

7.1.3.2. PW ID TLV

This TLV is used to communicate the configuration of PWs for VPWS.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0014 for "PW ID TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Peer ID

  4-octet LDP Router ID of the peer at the far end of the PW.

- Group ID

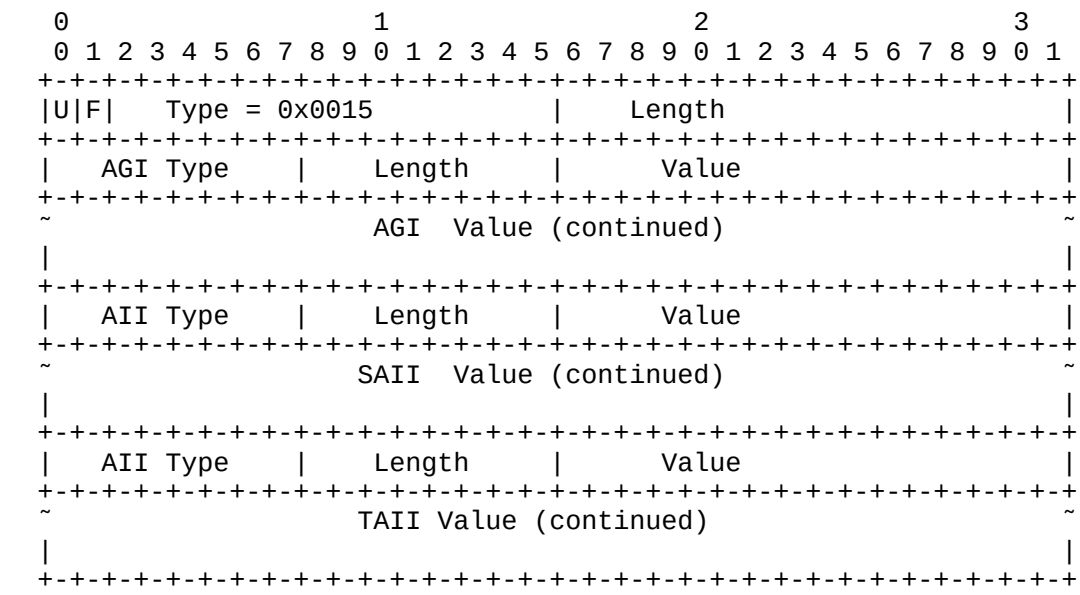
  Same as Group ID in [RFC4447], Section 5.2.

- PW ID

  Same as PW ID in [RFC4447], Section 5.2.

7.1.3.3. Generalized PW ID TLV

This TLV is used to communicate the configuration of PWs for VPLS.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0015 for "Generalized PW ID TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

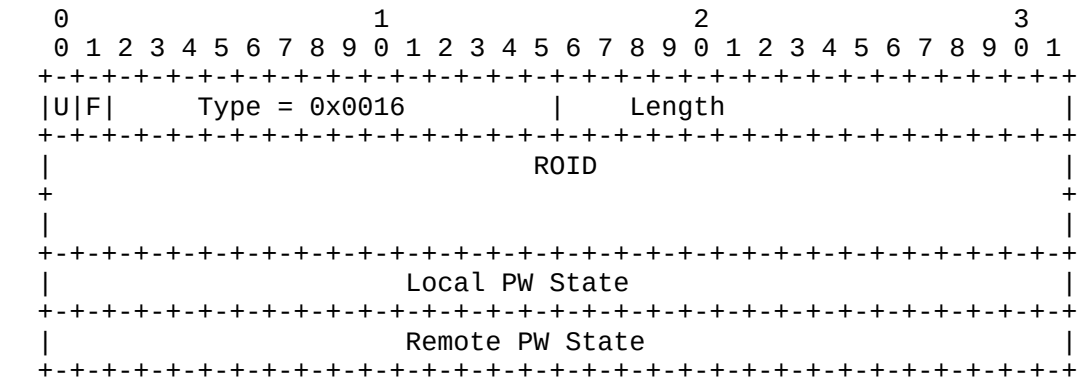
- AGI, AII, SAII, and TAII

  Defined in [RFC4447], Section 5.3.2.

7.1.4. PW-RED State TLV

The "PW-RED State TLV" is used in the "RG Application Data" message. This TLV is used by a device to report its PW status to other members in the RG.

The format of this TLV is as follows:



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0016 for "PW-RED State TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- ROID

  As defined in Section 6.1.3.

- Local PW State

  The status of the PW as determined by the sending PE, encoded in the same format as the "Status Code" field of the "PW Status TLV" defined in [RFC4447] and extended in [RFC6870].

- Remote PW State

  The status of the PW as determined by the remote peer of the sending PE. Encoded in the same format as the "Status Code" field of the "PW Status TLV" defined in [RFC4447] and extended in [RFC6870].

7.1.5. PW-RED Synchronization Request TLV

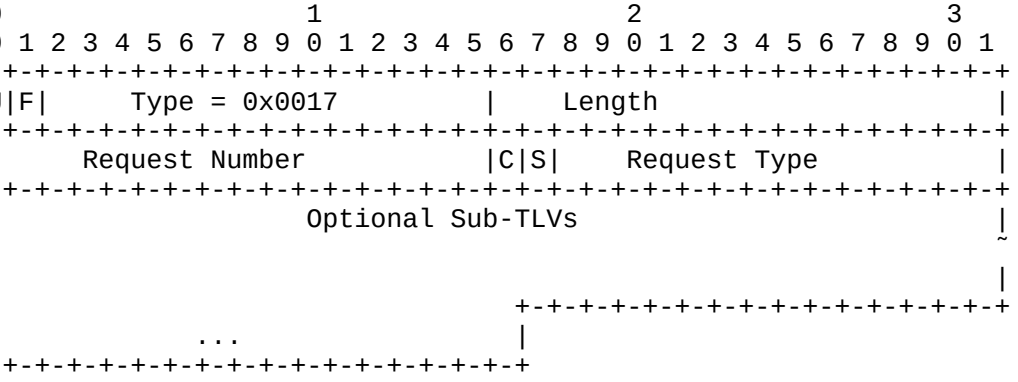
The "PW-RED Synchronization Request TLV" is used in the "RG Application Data" message. This TLV is used by a device to request that its peer retransmit configuration or operational state. The following information can be requested:

- configuration and/or state for one or more pseudowires

- configuration and/or state for all pseudowires

- configuration and/or state for all pseudowires in a given service

The format of the TLV is as follows:



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0017 for "PW-RED Synchronization Request TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Request Number

  2 octets. Unsigned integer uniquely identifying the request. Used to match the request with a response. The value of 0 is reserved for unsolicited synchronization and MUST NOT be used in the "PW-RED Synchronization Request TLV". Given the use of TCP, there are no issues associated with the wrap-around of the Request Number.

- C-bit

  Set to 1 if the request is for configuration data. Otherwise, set to 0.

- S-bit

  Set to 1 if the request is for running state data. Otherwise, set to 0.

- Request Type

  14 bits specifying the request type, encoded as follows:

0x00	Request Data for specified pseudowire(s)
0x01	Request Data for all pseudowires in specified service(s)
0x3FFF	Request All Data

- Optional Sub-TLVs

  A set of zero or more TLVs, as follows:

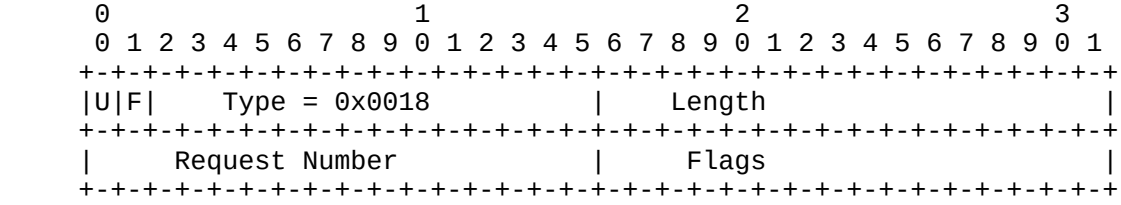
  If the "Request Type" field is set to 0x00, then this field contains one or more "PW ID TLVs" or "Generalized PW ID TLVs". If the "Request Type" field is set to 0x01, then this field contains one or more "Service Name TLVs". If the "Request Type" field is set to 0x3FFF, then this field MUST be empty. This document does not impose any restrictions on the length of the sub-TLVs.

7.1.6. PW-RED Synchronization Data TLV

The "PW-RED Synchronization Data TLV" is used in the "RG Application Data" message. A pair of these TLVs is used by a device to delimit a set of TLVs that are sent in response to a "PW-RED Synchronization Request TLV". The delimiting TLVs signal the start and end of the synchronization data and associate the response with its corresponding request via the "Request Number" field.

The "PW-RED Synchronization Data TLVs" are also used for unsolicited advertisements of complete PW-RED configuration and operational state data. In this case, the "Request Number" field MUST be set to 0.

This TLV has the following format:



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0018 for "PW-RED Synchronization Data TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Request Number

  2 octets. Unsigned integer identifying the Request Number from the "PW-RED Synchronization Request TLV" that solicited this synchronization data response.

- Flags

  2 octets. Response flags encoded as follows:

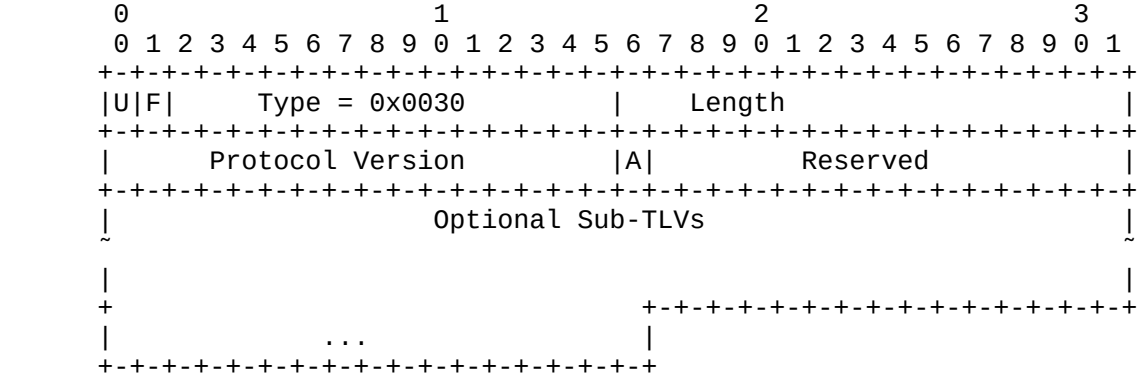
  0x00 Synchronization Data Start
  0x01 Synchronization Data End

7.2. Multi-Chassis LACP (mLACP) Application TLVs

This section discusses the "ICCP TLVs" for Ethernet attachment circuit redundancy using the multi-chassis LACP (mLACP) application.

7.2.1. mLACP Connect TLV

This TLV is included in the "RG Connect" message to signal the establishment of an mLACP Application Connection.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0030 for "mLACP Connect TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Protocol Version

  The version of this particular protocol for the purposes of ICCP. This is set to 0x0001.

- A-bit

  Acknowledgement bit. Set to 1 if the sender has received an "mLACP Connect TLV" from the recipient. Otherwise, set to 0.

- Reserved

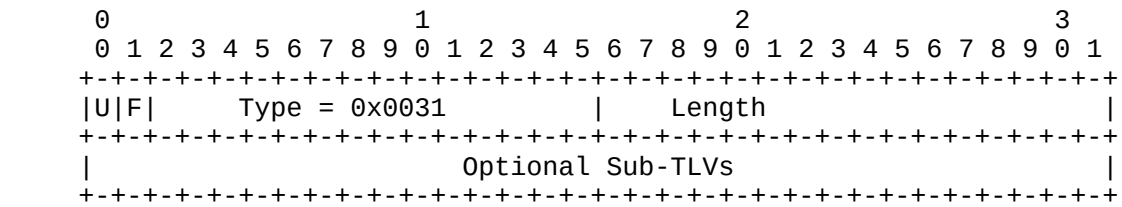
  Reserved for future use.

- Optional Sub-TLVs

  There are no optional sub-TLVs defined for this version of the protocol.

7.2.2. mLACP Disconnect TLV

This TLV is used in an "RG Disconnect" message to indicate that the connection for the mLACP application is to be terminated.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0031 for "mLACP Disconnect TLV".

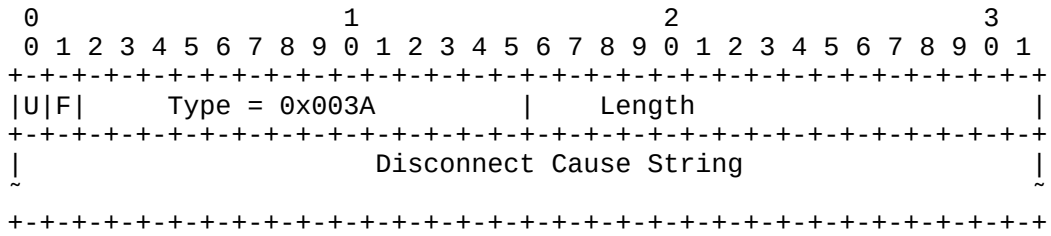
- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Optional Sub-TLVs

  The only optional sub-TLV defined for this version of the protocol is the "mLACP Disconnect Cause TLV" defined in Section 7.2.2.1.

7.2.2.1. mLACP Disconnect Cause TLV



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x003A for "mLACP Disconnect Cause TLV".

- Length

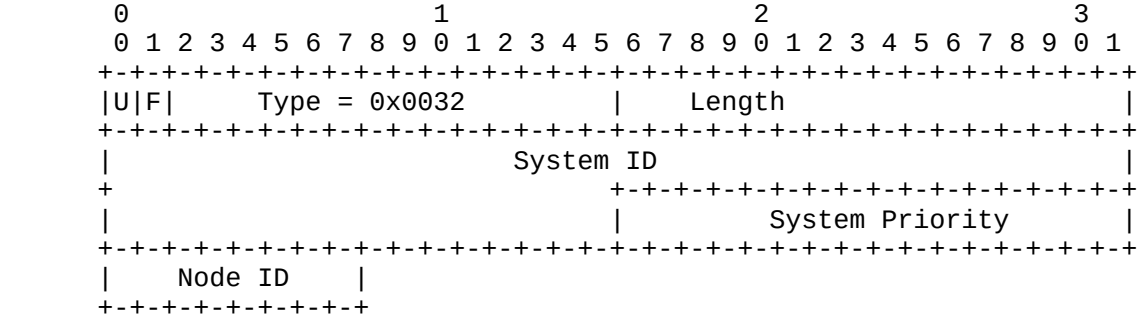
  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Disconnect Cause String

  Variable-length string specifying the reason for the disconnect. Used for network management.

7.2.3. mLACP System Config TLV

The "mLACP System Config TLV" is sent in the "RG Application Data" message. This TLV announces the local node's LACP system parameters to the RG peers.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0032 for "mLACP System Config TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- System ID

  6-octet field encoding the System ID used by LACP, as specified in [IEEE-802.1AX], Section 5.3.2.

- System Priority

  2 octets encoding the LACP System Priority, as defined in [IEEE-802.1AX], Section 5.3.2.

- Node ID

  1 octet. LACP Node ID. Used to ensure that the LACP Port Numbers are unique across all devices in an RG. Valid values are in the range 0-7. Uniqueness of the LACP Port Numbers across RG members is ensured by encoding the Port Numbers as follows:

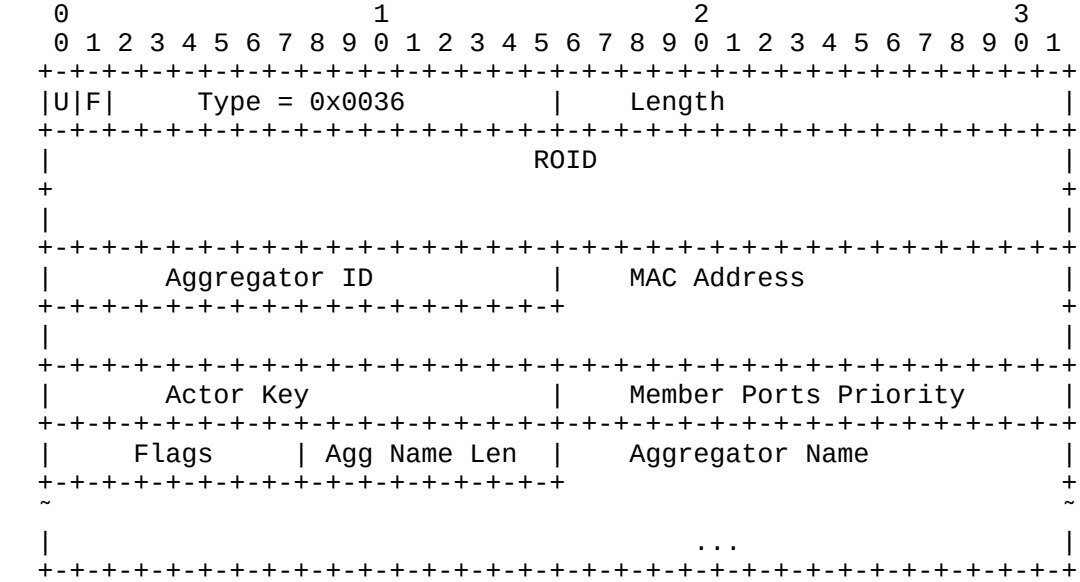
  - Most significant bit always set to 1

  - The next 3 most significant bits set to Node ID

  - Remaining 12 bits freely assigned by the system

7.2.4. mLACP Aggregator Config TLV

The "mLACP Aggregator Config TLV" is sent in the "RG Application Data" message. This TLV is used to notify RG peers about the local configuration state of an Aggregator.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0036 for "mLACP Aggregator Config TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- ROID

  Defined in Section 6.1.3 above.

- Aggregator ID

  2 octets. LACP Aggregator Identifier, as specified in [IEEE-802.1AX], Section 5.4.6.

- MAC Address

  6 octets encoding the Aggregator Media Access Control (MAC) address.

- Actor Key

  2 octets. LACP Actor Key for the corresponding Aggregator, as specified in [IEEE-802.1AX], Section 5.3.5.

- Member Ports Priority

  2 octets. LACP administrative port priority associated with all interfaces bound to the Aggregator. This field is valid only when the "Flags" field has "Priority Set" asserted.

- Flags

  Valid values are as follows:

  i. Synchronized (0x01)

     Indicates that the sender has concluded transmitting all Aggregator configuration information.

  ii. Purge Configuration (0x02)

     Indicates that the Aggregator is no longer configured for mLACP operation.

  iii. Priority Set (0x04)

     Indicates that the "Member Ports Priority" field is valid.

- Agg Name Len

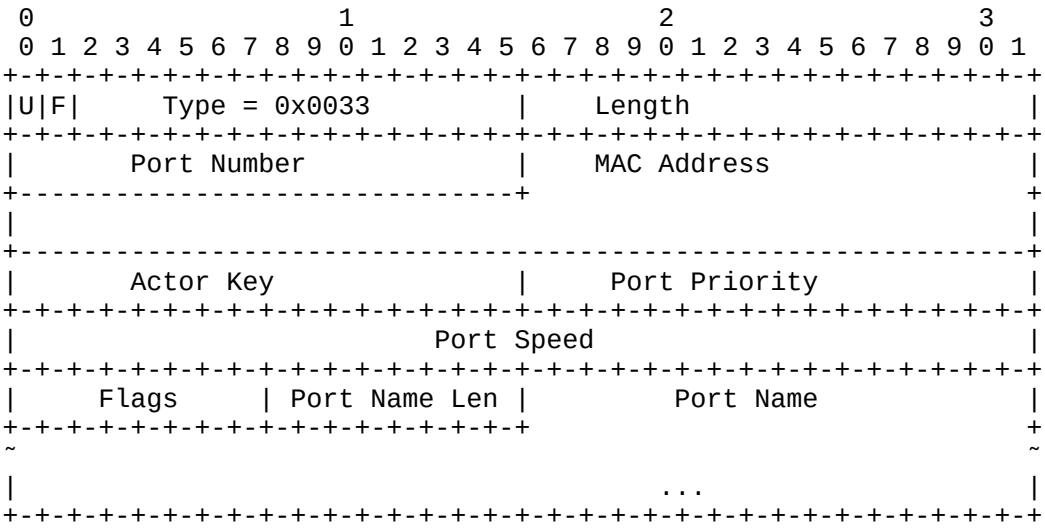
  1 octet. Length of the "Aggregator Name" field in octets.

- Aggregator Name

  Aggregator name, encoded in UTF-8 format, up to a maximum of 20 octets. Used for ease of management. The string does not include a terminating null character.

7.2.5. mLACP Port Config TLV

The "mLACP Port Config TLV" is sent in the "RG Application Data" message. This TLV is used to notify RG peers about the local configuration state of a port.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0033 for "mLACP Port Config TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Port Number

  2 octets. LACP Port Number for the corresponding interface, as specified in [IEEE-802.1AX], Section 5.3.4. The Port Number MUST be encoded with the Node ID, as discussed above.

- MAC Address

  6 octets encoding the port MAC address.

- Actor Key

  2 octets. LACP Actor Key for the corresponding interface, as specified in [IEEE-802.1AX], Section 5.3.5.

- Port Priority

  2 octets. LACP administrative port priority for the corresponding interface, as specified in [IEEE-802.1AX], Section 5.3.4. This field is valid only when the "Flags" field has "Priority Set" asserted.

- Port Speed

  4-octet integer encoding the port's current bandwidth in units of 1,000,000 bits per second. This field corresponds to the ifHighSpeed object of the IF-MIB [RFC2863].

- Flags

  Valid values are as follows:

  i. Synchronized (0x01)

     Indicates that the sender has concluded transmitting all member link port configurations for a given Aggregator.

  ii. Purge Configuration (0x02)

     Indicates that the port is no longer configured for mLACP operation.

  iii. Priority Set (0x04)

     Indicates that the "Port Priority" field is valid.

- Port Name Len

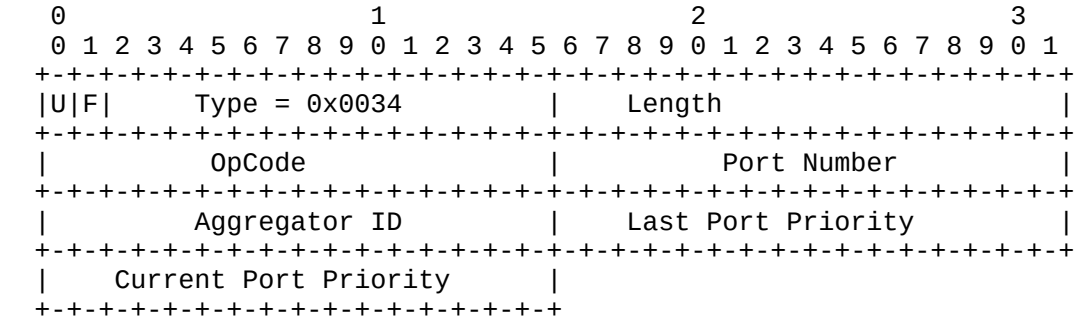
  1 octet. Length of the "Port Name" field in octets.

- Port Name

  Corresponds to the ifName object of the IF-MIB [RFC2863]. Encoded in UTF-8 format and truncated to 20 octets. Port Name does not include a terminating null character.

7.2.6. mLACP Port Priority TLV

The "mLACP Port Priority TLV" is sent in the "RG Application Data" message. This TLV is used by a device to either advertise its operational Port Priority to other members in the RG or authoritatively request that a particular member of an RG change its port priority.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0034 for "mLACP Port Priority TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- OpCode

  2 octets identifying the operational code point for the TLV, encoded as follows:

  0x00 Local Priority Change Notification
  0x01 Remote Request for Priority Change

- Port Number

  2-octet field representing the LACP Port Number, as specified in [IEEE-802.1AX], Section 5.3.4. When the value of this field is 0, it denotes all ports bound to the Aggregator specified in the "Aggregator ID" field. When non-zero, the Port Number MUST be encoded with the Node ID, as discussed above.

- Aggregator ID

  2 octets. LACP Aggregator Identifier, as specified in [IEEE-802.1AX], Section 5.4.6.

- Last Port Priority

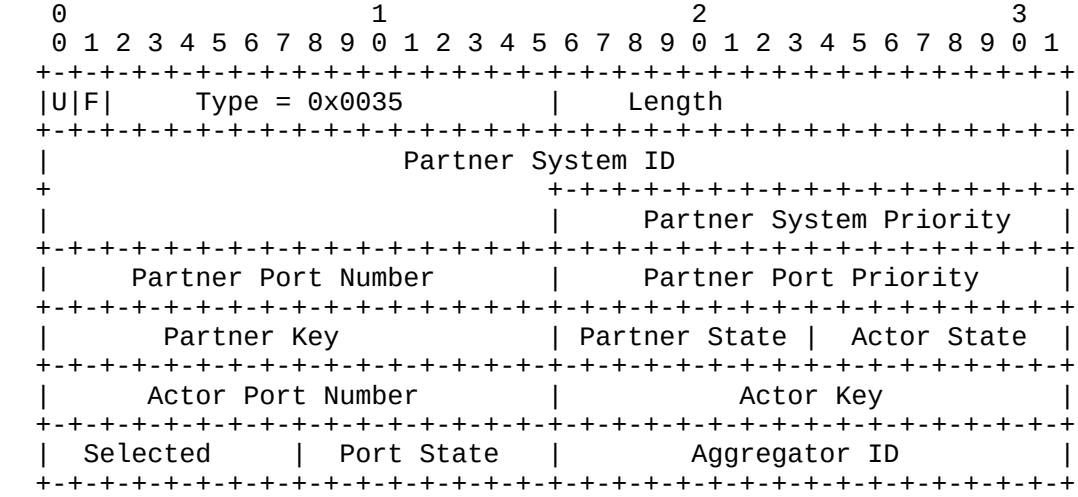
  2 octets. LACP port priority for the corresponding interface, as specified in [IEEE-802.1AX], Section 5.3.4. For local ports, this field encodes the previous operational value of port priority. For remote ports, this field encodes the operational port priority last known to the PE via notifications received from its peers in the RG.

- Current Port Priority

  2 octets. LACP port priority for the corresponding interface, as specified in [IEEE-802.1AX], Section 5.3.4. For local ports, this field encodes the new operational value of port priority being advertised by the PE. For remote ports, this field specifies the new port priority being requested by the PE.

7.2.7. mLACP Port State TLV

The "mLACP Port State TLV" is used in the "RG Application Data" message. This TLV is used by a device to report its LACP port status to other members in the RG.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0035 for "mLACP Port State TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Partner System ID

  6 octets. The LACP Partner System ID for the corresponding interface, encoded as a MAC address as specified in [IEEE-802.1AX], Section 5.4.2.2, item r.

- Partner System Priority

  2-octet field specifying the LACP Partner System Priority, as specified in [IEEE-802.1AX], Section 5.4.2.2, item q.

- Partner Port Number

  2 octets encoding the LACP Partner Port Number, as specified in [IEEE-802.1AX], Section 5.4.2.2, item u. The Port Number MUST be encoded with the Node ID, as discussed above.

- Partner Port Priority

  2-octet field encoding the LACP Partner Port Priority, as specified in [IEEE-802.1AX], Section 5.4.2.2, item t.

- Partner Key

  2-octet field representing the LACP Partner Key, as defined in [IEEE-802.1AX], Section 5.4.2.2, item s.

- Partner State

  1-octet field encoding the LACP Partner State Variable, as defined in [IEEE-802.1AX], Section 5.4.2.2, item v.

- Actor State

  1 octet encoding the LACP Actor State Variable for the port, as specified in [IEEE-802.1AX], Section 5.4.2.2, item m.

- Actor Port Number

  2-octet field representing the LACP Actor Port Number, as specified in [IEEE-802.1AX], Section 5.3.4. The Port Number MUST be encoded with the Node ID, as discussed above.

- Actor Key

  2-octet field encoding the LACP Actor Operational Key, as specified in [IEEE-802.1AX], Section 5.3.5.

- Selected

  1 octet encoding the LACP "Selected" variable, defined in [IEEE-802.1AX], Section 5.4.8 as follows:

  0x00 SELECTED
  0x01 UNSELECTED
  0x02 STANDBY

- Port State

  1 octet encoding the operational state of the port as follows:

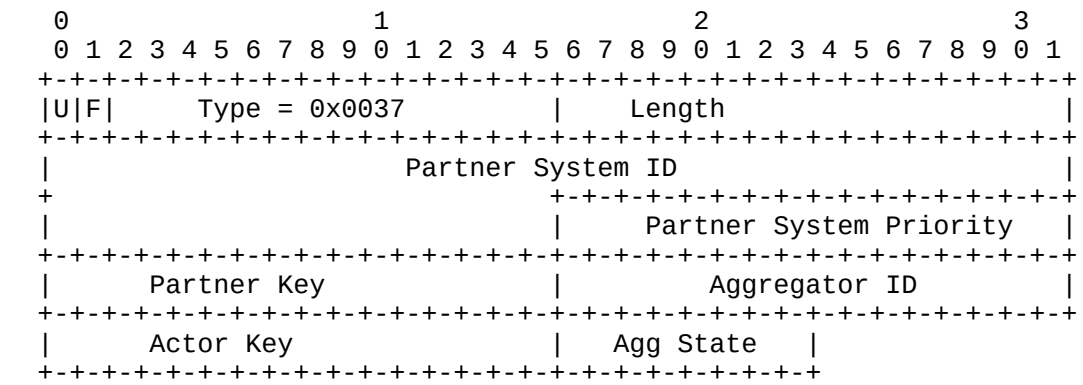
  0x00 Up
  0x01 Down
  0x02 Administratively Down
  0x03 Test (e.g., IEEE 802.3ah OAM Intrusive Loopback mode)

- Aggregator ID

  2 octets. LACP Aggregator Identifier to which this port is bound based on the outcome of the LACP selection logic.

7.2.8. mLACP Aggregator State TLV

The "mLACP Aggregator State TLV" is used in the "RG Application Data" message. This TLV is used by a device to report its Aggregator status to other members in the RG.



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0037 for "mLACP Aggregator State TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Partner System ID

  6 octets. The LACP Partner System ID for the corresponding interface, encoded as a MAC address as specified in [IEEE-802.1AX], Section 5.4.2.2, item r.

- Partner System Priority

  2-octet field specifying the LACP Partner System Priority, as specified in [IEEE-802.1AX], Section 5.4.2.2, item q.

- Partner Key

  2-octet field representing the LACP Partner Key, as defined in [IEEE-802.1AX], Section 5.4.2.2, item s.

- Aggregator ID

  2 octets. LACP Aggregator Identifier, as specified in [IEEE-802.1AX], Section 5.4.6.

- Actor Key

  2-octet field encoding the LACP Actor Operational Key, as specified in [IEEE-802.1AX], Section 5.3.5.

- Agg State

  1 octet encoding the operational state of the Aggregator as follows:

  0x00 Up
  0x01 Down
  0x02 Administratively Down
  0x03 Test (e.g., IEEE 802.3ah OAM Intrusive Loopback mode)

7.2.9. mLACP Synchronization Request TLV

The "mLACP Synchronization Request TLV" is used in the "RG Application Data" message. This TLV is used by a device to request that its peer retransmit configuration or operational state. The following information can be requested:

- system configuration and/or state

- configuration and/or state for a specific port

- configuration and/or state for all ports with a specific LACP Key

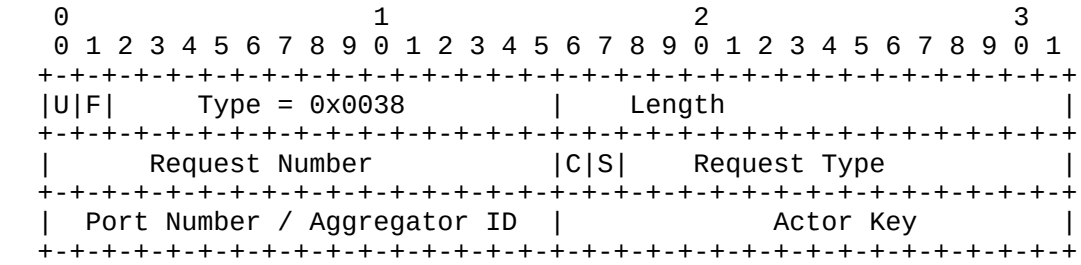
- configuration and/or state for all mLACP ports

- configuration and/or state for a specific Aggregator

- configuration and/or state for all Aggregators with a specific LACP Key

- configuration and/or state for all mLACP Aggregators

The format of the TLV is as follows:



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0038 for "mLACP Synchronization Request TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Request Number

  2 octets. Unsigned integer uniquely identifying the request. Used to match the request with a response. The value of 0 is reserved for unsolicited synchronization and MUST NOT be used in the "mLACP Synchronization Request TLV".

- C-bit

  Set to 1 if the request is for configuration data. Otherwise, set to 0.

- S-bit

  Set to 1 if the request is for running state data. Otherwise, set to 0.

- Request Type

  14 bits specifying the request type, encoded as follows:

0x00	Request System Data
0x01	Request Aggregator Data
0x02	Request Port Data
0x3FFF	Request All Data

- Port Number / Aggregator ID

  2 octets. When the "Request Type" field is set to "Request Port Data", this field encodes the LACP Port Number for the requested port. When the "Request Type" field is set to "Request Aggregator Data", this field encodes the Aggregator ID of the requested Aggregator. When the value of this field is 0, it denotes that information for all ports (or Aggregators) whose LACP Key is specified in the "Actor Key" field is being requested.

- Actor Key

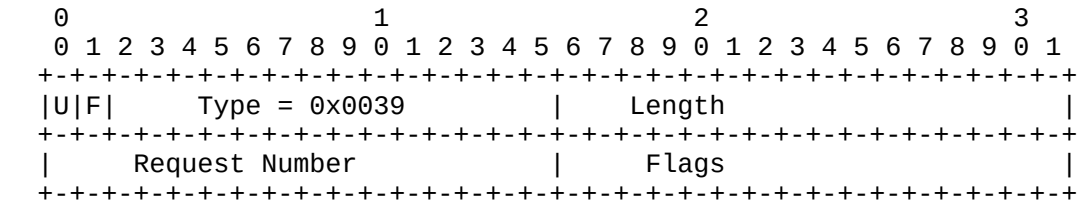
  2 octets. LACP Actor Key for the corresponding port or Aggregator. When the value of this field is 0 (and the Port Number / Aggregator ID field is 0 as well), it denotes that information for all ports or Aggregators in the system is being requested.

7.2.10. mLACP Synchronization Data TLV

The "mLACP Synchronization Data TLV" is used in the "RG Application Data" message. A pair of these TLVs is used by a device to delimit a set of TLVs that are being transmitted in response to an "mLACP Synchronization Request TLV". The delimiting TLVs signal the start and end of the synchronization data and associate the response with its corresponding request via the "Request Number" field.

The "mLACP Synchronization Data TLVs" are also used for unsolicited advertisements of complete mLACP configuration and operational state data. The "Request Number" field MUST be set to 0 in this case. For such unsolicited synchronization, the PE MUST advertise all system, Aggregator, and port information, as done during the initialization sequence.

This TLV has the following format:



- U-bit and F-bit

  Both are set to 0.

- Type

  Set to 0x0039 for "mLACP Synchronization Data TLV".

- Length

  Length of the TLV in octets, excluding the "U-bit", "F-bit", "Type", and "Length" fields.

- Request Number

  2 octets. Unsigned integer identifying the Request Number from the "mLACP Synchronization Request TLV" that solicited this synchronization data response.

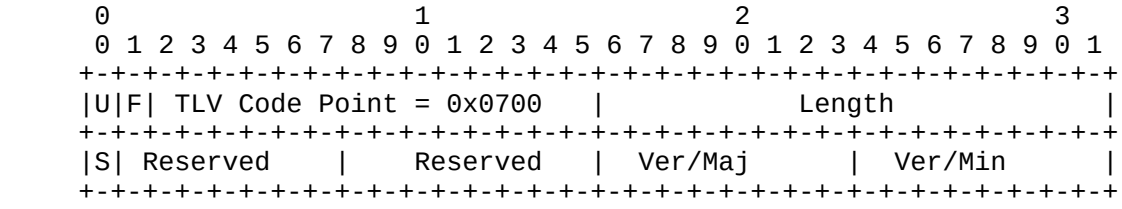
- Flags

  2 octets. Response flags, encoded as follows:

  0x00 Synchronization Data Start
  0x01 Synchronization Data End

8. LDP Capability Negotiation

As required in [RFC5561], the following TLV is defined to indicate the ICCP capability:



- U-bit

  SHOULD be 1 (ignore if not understood).

- F-bit

  SHOULD be 0 (don't forward if not understood).

- TLV Code Point

  The TLV type, which identifies a specific capability. The ICCP code point is listed in Section 12 below.

- S-bit

  State bit. Indicates whether the sender is advertising or withdrawing the ICCP capability. The State bit is used as follows:

  1 - The TLV is advertising the capability specified by the TLV Code Point.

  0 - The TLV is withdrawing the capability specified by the TLV Code Point.

- Ver/Maj

  The major version revision of ICCP. This document specifies 1.0, and so this field is set to 1.

- Ver/Min

  The minor version revision of ICCP. This document specifies 1.0, and so this field is set to 0.

ICCP capability is advertised to an LDP peer if there is at least one RG enabled on the local PE.

9. Client Applications

9.1. Pseudowire Redundancy Application Procedures

This section defines the procedures for the Pseudowire Redundancy (PW-RED) application.

It should be noted that the PW-RED application SHOULD NOT be enabled together with an AC redundancy application for the same service instance. This simplifies the operation of the multi-chassis redundancy solution (Figure 1) and eliminates the possibility of deadlock conditions between the AC and PW redundancy mechanisms.

9.1.1. Initial Setup

When an RG is configured on a system and multi-chassis pseudowire redundancy is enabled in that RG, the PW-RED application MUST send an "RG Connect" message with a "PW-RED Connect TLV" to each PE that is a member of the same RG. The sending PE MUST set the A-bit to 1 if it has already received a "PW-RED Connect TLV" from its peer; otherwise, the PE MUST set the A-bit to 0. If a PE that has sent the TLV with the A-bit set to 0 receives a "PW-RED Connect TLV" from a peer, it MUST repeat its advertisement with the A-bit set to 1. The PW-RED Application Connection is considered to be operational when both PEs have sent and received "PW-RED Connect TLVs" with the A-bit set to 1. Once the Application Connection becomes operational, the two devices can start exchanging "RG Application Data" messages for the PW-RED application.

If a system receives an "RG Connect" message with a "PW-RED Connect TLV" that has a different Protocol Version, it must follow the procedures outlined in Section 4.4.1 above.

When the PW-RED application is disabled on the device or is unconfigured for the RG in question, the system MUST send an "RG Disconnect" message with a "PW-RED Disconnect TLV".

9.1.2. Pseudowire Configuration Synchronization

A system MUST advertise its local PW configuration to other PEs that are members of the same RG. This allows the PEs to build a view of the redundant nodes and pseudowires that are protecting the same service instances. The advertisement MUST be initiated when the PW-RED Application Connection first comes up. To that end, the system sends "RG Application Data" messages with "PW-RED Config TLVs"

as part of an unsolicited synchronization. A PE MUST use a pair of "PW-RED Synchronization Data TLVs" to delimit the set of TLVs that are being sent as part of this unsolicited advertisement.

In the case of a configuration change, a PE MUST re-advertise the most up-to-date information for the affected pseudowires.

As part of the configuration synchronization, a PE advertises the ROID associated with the pseudowire. This is used to correlate the pseudowires that are protecting each other on different PEs. A PE also advertises the configured PW redundancy mode. This can be one of the following four options: Master Mode, Slave Mode, Independent Mode, or Independent Mode with Request Switchover. If the received redundancy mode does not match the locally configured mode for the same ROID, then the PE MUST respond with an "RG Notification" message to reject the "PW-RED Config TLV". The PE MUST disable the associated local pseudowire until a satisfactory "PW-RED Config TLV" is received from the peer. This guarantees that device misconfiguration does not lead to network-wide problems (e.g., by creating forwarding loops). The PE SHOULD also raise an alarm to alert the operator. If a PE receives a "NAK TLV" for an advertised "PW-RED Config TLV", it MUST disable the associated pseudowire and SHOULD raise an alarm to alert the operator.

Furthermore, a PE advertises in its "PW-RED Config TLVs" a priority value that is used to determine the precedence of a given pseudowire to assume the active role in a redundant setup. A PE also advertises a Service Name that is global in the context of an RG and is used to identify which pseudowires belong to the same service. Finally, a PE also advertises the pseudowire identifier as part of this synchronization.

9.1.3. Pseudowire Status Synchronization

PEs that are members of an RG synchronize pseudowire status for the purpose of identifying, on a per-ROID basis, which pseudowire will be actively used for forwarding and which pseudowire(s) will be placed in standby state.

Synchronization of pseudowire status is done by sending the "PW-RED State TLV" whenever the pseudowire state changes on a PE. This includes changes to the local end as well as the remote end of the pseudowire.

A PE may request that its peer retransmit previously advertised PW-RED state. This is useful, for instance, when the PE is recovering from a soft failure. To request such a retransmission, a PE MUST send a set of one or more "PW-RED Synchronization Request TLVs".

A PE MUST respond to a "PW-RED Synchronization Request TLV" by sending the requested data in a set of one or more "PW-RED TLVs" delimited by a pair of "PW-RED Synchronization Data TLVs". The TLVs comprising the response MUST be ordered such that the "Synchronization Response TLV" with the "Synchronization Data Start" flag precedes the various other "PW-RED TLVs" encoding the requested data. These, in turn, MUST precede the "Synchronization Data TLV" with the "Synchronization Data End" flag. It is worth noting that the response may span multiple "RG Application Data" messages; however, the above TLV ordering MUST be retained across messages, and only a single pair of "Synchronization Data TLVs" must be used to delimit the response across all "Application Data" messages.

A PE MAY re-advertise its PW-RED state in an unsolicited manner. This is done by sending the appropriate Config and State TLVs delimited by a pair of "PW-RED Synchronization Data TLVs" and using a "Request Number" of 0.

While a PE has a pending synchronization request for a pseudowire or a service, it SHOULD silently ignore all TLVs for said pseudowire or service that are received prior to the synchronization response and that carry the same type of information being requested. This saves the system from the burden of updating state that will ultimately be overwritten by the synchronization response. Note that TLVs pertaining to other pseudowires or services are to continue to be processed per normal procedures in the interim.

If a PE receives a synchronization request for a pseudowire or service that doesn’t exist or is not known to the PE, then it MUST trigger an unsolicited synchronization of all pseudowire information (i.e., replay the initialization sequence).

In the subsections that follow, we describe the details of pseudowire status synchronization for each of the PW redundancy modes defined in [RFC6870].

9.1.3.1. Independent Mode

This section covers the operation in Independent Mode with or without Request Switchover capability.

In this mode, the operator must ensure that for a given RO the PW Priority values configured for all associated pseudowires on a given PE are collectively higher (or lower) than those configured on other PEs in the same RG. If this condition is not satisfied after the PEs have exchanged "PW-RED State TLVs", a PE MUST disable the associated pseudowire(s) and SHOULD raise an alarm to alert the operator. Note that the PW Priority MAY be the same as the PW Precedence as defined in [RFC6870].

For a given RO, after all of the PEs in an RG have exchanged their "PW-RED State TLVs", the PE with the best PW Priority (i.e., least numeric value) advertises active Preferential Forwarding status in LDP on all of its associated pseudowires, whereas all other PEs in the RG advertise standby Preferential Forwarding status in LDP on their associated pseudowires.

If the service is VPWS, then only a single pseudowire per service will be selected for forwarding. This is the pseudowire that is independently advertised with active Preferential Forwarding status on both endpoints, as described in [RFC6870].

If the service is VPLS, then one or multiple pseudowires per service will be selected for forwarding. These are the pseudowires that are independently advertised with active Preferential Forwarding status on both PW endpoints, as described in [RFC6870].

9.1.3.2. Master/Slave Mode

In this mode, the operator must ensure that for a given RO the PW Priority values configured for all associated pseudowires on a given PE are collectively higher (or lower) than those configured on other PEs in the same RG. If this condition is not satisfied after the PEs have exchanged "PW-RED State TLVs", a PE MUST disable the associated pseudowire(s) and SHOULD raise an alarm to alert the operator. Note that the PW Priority MAY be the same as the PW Precedence as defined in [RFC6870]. In addition, the operator must ensure that for a given RO all of the PEs in the RG are consistently configured as Master or Slave.

In the context of a given RO, if the PEs in the RG are acting as Master, then the PE with the best PW Priority (i.e., least numeric value) advertises active Preferential Forwarding status in LDP on

only a single pseudowire, following the procedures in Sections 5.2 and 6.2 of [RFC6870], whereas all of the other pseudowires on other PEs in the RG are advertised with standby Preferential Forwarding status in LDP.

9.1.4. PE Node Failure or Isolation

When a PE node detects that a remote PE that is a member of the same RG is no longer reachable (using the mechanisms described in Section 5), the local PE determines if it has redundant PWs for the affected services. If the local PE has the highest priority (after the failed PE), then it becomes the active node for the services in question and subsequently activates its associated PW(s).

9.2. Attachment Circuit Redundancy Application Procedures

9.2.1. Common AC Procedures

This section describes generic procedures for AC redundancy applications, independent of the type of the AC (ATM, FR, or Ethernet).

9.2.1.1. AC Failure

When the AC redundancy mechanism on the active PE detects a failure of the AC, it should send an ICCP "Application Data" message to inform the redundant PEs of the need to take over. The AC failures can be categorized into the following scenarios:

- Failure of CE interface connecting to PE

- Failure of CE uplink to PE

- Failure of PE interface connecting to CE

9.2.1.2. Remote PE Node Failure or Isolation

When a PE node detects that a remote PE that is a member of the same RG is no longer reachable (using the mechanisms described in Section 5), the local PE determines if it has redundant ACs for the affected services. If the local PE has the highest priority (after the failed PE), then it becomes the active node for the services in question and subsequently activates its associated ACs.

9.2.1.3. Local PE Isolation

When a PE node detects that it has been isolated from the core network (i.e., all core-facing interfaces/links are not operational), then it should ensure that its AC redundancy mechanism will change the status of any active ACs to standby. The AC redundancy application SHOULD then send ICCP "Application Data" messages in order to trigger failover to a standby PE. Note that this works only in the case of dedicated interconnect (Sections 3.2.1 and 3.2.3), since ICCP will still have a path to the peer, even though the PE is isolated from the MPLS core network.

9.2.1.4. Determining Pseudowire State

If the PEs in an RG are running an AC redundancy application over ICCP, then the Independent Mode of PW redundancy, as defined in [RFC6870], MUST be used. On a given PE, the Preferential Forwarding status of the PW (active or standby) is derived from the state of the associated AC(s). This simplifies the operation of the multi-chassis redundancy solution (Figure 1) and eliminates the possibility of deadlock conditions between the AC and PW redundancy mechanisms. The rules by which the PW status is derived from the AC status are as follows:

- VPWS

  For VPWS, there’s a single AC per service instance. If the AC is active, then the PW status should be active. If the AC is standby, then the PW status should be standby.

- VPLS

  For VPLS, there could be multiple ACs per service instance (i.e., Virtual Switch Instance (VSI) [RFC4026]). If AT LEAST ONE AC is active, then the PW status should be active. If ALL ACs are standby, then the PW status should be standby.

In this case, the PW-RED application is not used to synchronize PW status between PEs. Rather, the AC redundancy application should synchronize AC status between PEs, in order to establish which AC (and subsequently which PE) is active or standby for a given service. When that is determined, each PE will then derive its local PW’s state according to the rules described above. The Preferential Forwarding status bit, described in [RFC6870], is used to advertise PW status to the remote peers.

9.2.2. Multi-Chassis LACP (mLACP) Application Procedures

This section defines the procedures that are specific to the multi-chassis LACP (mLACP) application, which is applicable for Ethernet ACs.

9.2.2.1. Initial Setup

When an RG is configured on a system and mLACP is enabled in that RG, the mLACP application MUST send an "RG Connect" message with an "mLACP Connect TLV" to each PE that is a member of the same RG. The sending PE MUST set the A-bit to 1 in said TLV if it has received a corresponding "mLACP Connect TLV" from its peer PE; otherwise, the sending PE MUST set the A-bit to 0. If a PE receives an "mLACP Connect TLV" from its peer after sending said TLV with the A-bit set to 0, it MUST resend the TLV with the A-bit set to 1. A system considers the mLACP Application Connection to be operational when it has sent and received "mLACP Connect TLVs" with the A-bit set to 1. When the mLACP Application Connection between a pair of PEs is operational, the two devices can start exchanging "RG Application Data" messages for the mLACP application. This involves having each PE advertise its mLACP configuration and operational state in an unsolicited manner. A PE SHOULD use the following sequence when advertising its mLACP state upon initial Application Connection setup:

- Advertise system configuration

- Advertise Aggregator configuration

- Advertise port configuration

- Advertise Aggregator state

- Advertise port state

A PE MUST use a pair of "mLACP Synchronization Data TLVs" to delimit the entire set of TLVs that are being sent as part of this unsolicited advertisement.

If a system receives an "RG Connect" message with an "mLACP Connect TLV" that has a different Protocol Version, it MUST follow the procedures outlined in Section 4.4.1 above.

After the mLACP Application Connection has been established, every PE MUST communicate its system-level configuration to its peers via the use of the "mLACP System Config TLV". This allows every PE to discover the Node ID and the locally configured System ID and System Priority values of its peers.

If a PE receives an "mLACP System Config TLV" from a remote peer advertising the same Node ID value as the local system, then the PE MUST respond with an "RG Notification" message to reject the "mLACP System Config TLV". The PE MUST suspend the mLACP application until a satisfactory "mLACP System Config TLV" is received from the peer. It SHOULD also raise an alarm to alert the operator. Furthermore, if a PE receives a "NAK TLV" for an "mLACP System Config TLV" that it has advertised, the PE MUST suspend the mLACP application and SHOULD raise an alarm to alert the network operator of potential device misconfiguration.

If a PE receives an "mLACP System Config TLV" from a new peer advertising the same Node ID value as another existing peer with which the local system has an established mLACP Application Connection, then the PE MUST respond to the new peer with an "RG Notification" message to reject the "mLACP System Config TLV" and MUST ignore the offending TLV.

If the Node ID of a particular PE changes due to administrative configuration action, the PE MUST then inform its peers to purge the configuration of all previously advertised ports and/or Aggregators and MUST replay the initialization sequence by sending an unsolicited synchronization of the system configuration, Aggregator configuration, port configuration, Aggregator state, and port state.

It is necessary for all PEs in an RG to agree upon the System ID and System Priority values to be used ubiquitously. To achieve this, every PE MUST use the values for the two parameters that are supplied by the PE with the numerically lowest value (among RG members) of System Aggregation Priority. This guarantees that the PEs always agree on uniform values that yield the highest System Priority.

When the mLACP application is disabled on the device or is unconfigured for the RG in question, the system MUST send an "RG Disconnect" message with an "mLACP Disconnect TLV".

9.2.2.2. mLACP Aggregator and Port Configuration

A system MUST synchronize the configuration of its mLACP-enabled Aggregators and ports with other RG members. This is achieved via the use of "mLACP Aggregator Config TLVs" and "mLACP Port Config TLVs", respectively. An implementation MUST advertise the configuration of Aggregators prior to advertising the configuration of any of their associated member ports.

The PEs in an RG MUST all agree on the MAC address to be associated with a given Aggregator. It is possible to achieve this via consistent configuration on member PEs. However, in order to protect against possible misconfiguration, a system MUST use, for any given Aggregator, the MAC address supplied by the PE with the numerically lowest System Aggregation Priority in the RG.

A system that receives an "mLACP Aggregator Config TLV" with an ROID-to-Key association that is different from its local association MUST reject the corresponding TLV and disable the Aggregator with the same ROID. Furthermore, it SHOULD raise an alarm to alert the operator. Similarly, a system that receives a "NAK TLV" in response to a transmitted "mLACP Aggregator Config TLV" MUST disable the associated Aggregator and SHOULD raise an alarm to alert the network operator.

A system MAY enforce a restriction that all ports that are to be bundled together on a given PE share the same Port Priority value. If so, the system MUST advertise this common priority in the "mLACP Aggregator Config TLV" and assert the "Priority Set" flag in that TLV. Furthermore, the system in this case MUST NOT advertise individual Port Priority values in the associated "mLACP Port Config TLVs" (i.e., the "Priority Set" flag in these TLVs should be 0).

A system MAY support individual Port Priority values to be configured on ports that are to be bundled together on a PE. If so, the system MUST advertise the individual Port Priority values in the appropriate "mLACP Port Config TLVs" and MUST NOT assert the "Priority Set" flag in the corresponding "mLACP Aggregator Config TLV".

When the configurations of all ports for member links associated with a given Aggregator have been sent by a device, it asserts that fact by setting the "Synchronized" flag in the last port’s "mLACP Port Config TLV". If an Aggregator doesn’t have any candidate member ports configured, this is indicated by asserting the "Synchronized" flag in its "mLACP Aggregator Config TLV".

Furthermore, for a given port/Aggregator, an implementation MUST advertise the port/Aggregator configuration prior to advertising its state (via the "mLACP Port State TLV" or "mLACP Aggregator State

TLV"). If a PE receives an "mLACP Port State TLV" or "mLACP Aggregator State TLV" for a port or Aggregator that it had not previously learned via an appropriate "Port Config TLV" or "Aggregator Config TLV", then the PE MUST request synchronization of the configuration and state of all mLACP ports as well as all mLACP Aggregators from its respective peer. During a synchronization (solicited or unsolicited), if a PE receives a "State TLV" for a port or Aggregator that it has not learned before, then the PE MUST send a "NAK TLV" for the offending TLV. The PE MUST NOT request resynchronization in this case.

When mLACP is unconfigured on a port/Aggregator, a PE MUST send a "Port/Aggregator Config TLV" with the "Purge Configuration" flag asserted. This allows receiving PEs to purge any state maintained for the decommissioned port/Aggregator. If a PE receives a "Port/Aggregator Config TLV" with the "Purge Configuration" flag asserted and the PE is not maintaining any state for that port/Aggregator, then it MUST silently discard the TLV.

9.2.2.3. mLACP Aggregator and Port Status Synchronization

PEs within an RG need to synchronize their state machines for proper mLACP operation with a multi-homed device. This is achieved by having each system advertise its Aggregators and ports running state in "mLACP Aggregator State TLVs" and "mLACP Port State TLVs", respectively. Whenever any LACP parameter for an Aggregator or a port -- whether on the Partner (i.e., multi-homed device) side or the Actor (i.e., PE) side -- is changed, a system MUST transmit an updated TLV for the affected Aggregator and/or port. Moreover, when the administrative or operational state of an Aggregator or port changes, the system MUST transmit an updated Aggregator or Port State TLV to its peers.

If a PE receives an Aggregator or Port State TLV where the Actor Key doesn’t match what was previously received in a corresponding "Aggregator Config TLV" or "Port Config TLV", the PE MUST then request synchronization of the configuration and state of the affected Aggregator or port. If such a mismatch occurs between the Config and State TLVs as part of a synchronization (solicited or unsolicited), then the PE MUST send a "NAK TLV" for the "State TLV". Furthermore, if a PE receives a "Port State TLV" with the "Aggregator ID" set to a value that doesn’t map to some Aggregator that the PE had learned via a previous "Aggregator Config TLV", then the PE MUST request synchronization of the configuration and state of all Aggregators and ports. If the above anomaly occurs during a synchronization, then the PE MUST send a "NAK TLV" for the offending "Port State TLV".

A PE MAY request that its peer retransmit previously advertised state. This is useful, for example, when the PE is recovering from a soft failure and attempting to relearn state. To request such retransmissions, a PE MUST send a set of one or more "mLACP Synchronization Request TLVs".

A PE MUST respond to an "mLACP Synchronization Request TLV" by sending the requested data in a set of one or more mLACP TLVs delimited by a pair of "mLACP Synchronization Data TLVs". The TLVs comprising the response MUST be ordered in the "RG Application Data" message(s) such that the "Synchronization Response TLV" with the "Synchronization Data Start" flag precedes the various other mLACP TLVs encoding the requested data. These, in turn, MUST precede the "Synchronization Data TLV" with the "Synchronization Data End" flag. Note that the response may span multiple "RG Application Data" messages -- for example, when MTU limits are exceeded; however, the above ordering MUST be retained across messages, and only a single pair of "Synchronization Data TLVs" MUST be used to delimit the response across all "Application Data" messages.

A PE device MAY re-advertise its mLACP state in an unsolicited manner. This is done by sending the appropriate Config and State TLVs delimited by a pair of "mLACP Synchronization Data TLVs" and using a "Request Number" of 0.

While a PE has a pending synchronization request for a system, Aggregator, or port, it SHOULD silently ignore all TLVs for said system, Aggregator, or port that are received prior to the synchronization response and that carry the same type of information being requested. This saves the system from the burden of updating state that will ultimately be overwritten by the synchronization response. Note that TLVs pertaining to other systems, Aggregators, or ports are to continue to be processed per normal procedures in this case.

If a PE receives a synchronization request for an Aggregator, port, or key that doesn’t exist or is not known to the PE, then it MUST trigger an unsolicited synchronization of all system, Aggregator, and port information (i.e., replay the initialization sequence).

If a PE learns, as part of a synchronization operation from its peer, that the latter is advertising a Node ID value that is different from the value previously advertised, then the PE MUST purge all Port/Aggregator data previously learned from that peer prior to the last synchronization.

9.2.2.4. Failure and Recovery

When a PE that is active for a multi-chassis link aggregation group encounters a core isolation fault, it SHOULD attempt to fail over to a peer PE that hosts the same RO. The default failover procedure is to have the failed PE bring down the link or links towards the multi-homed CE (e.g., by bringing down the line protocol). This will cause the CE to fail over to the other member link or links of the bundle that are connected to the other PE(s) in the RG. Other procedures for triggering failover are possible; such procedures are outside the scope of this document.

Upon recovery from a previous fault, a PE MAY reclaim the active role for a multi-chassis link aggregation group if configured for revertive protection. Otherwise, the recovering PE may assume the standby role when configured for non-revertive protection. In the revertive scenario, a PE SHOULD assume the active role within the RG by sending an "mLACP Port Priority TLV" to the currently active PE, requesting that the latter change its port priority to a value that is lower (i.e., numerically larger) for the Aggregator in question.

If a system is operating in a mode where different ports of a bundle are configured with different Port Priorities, then the system MUST NOT advertise or request changes of Port Priority values for aggregated ports collectively (i.e., by using a "Port Number" of 0 in the "mLACP Port Priority TLV"). This is to avoid ambiguity in the interpretation of the "Last Port Priority" field.

If a PE receives an "mLACP Port Priority TLV" requesting a priority change for a port or Aggregator that is not local to the device, then the PE MUST re-advertise the local configuration of the system, as well as the configuration and state of all of its mLACP ports and Aggregators.

If a PE receives an "mLACP Port Priority TLV" in which the remote system is advertising priority change for a port or Aggregator that the local PE had not previously learned via an appropriate "Port Config TLV" or "Aggregator Config TLV", then the PE MUST request synchronization of the configuration and state of all mLACP ports as well as all mLACP Aggregators from its respective peer.

10. Security Considerations

ICCP SHOULD only be used in well-managed and highly monitored networks. It ought not be deployed on or over the public Internet. ICCP is not intended to be applicable when the Redundancy Group spans PEs in different administrative domains.

The security considerations described in [RFC5036] and [RFC4447] that apply to the base LDP specification and to the PW LDP control protocol extensions apply to the capability mechanism described in this document. In particular, ICCP implementations MUST provide a mechanism to select to which LDP peers the ICCP capability will be advertised, and from which LDP peers the ICCP messages will be accepted. Therefore, an incoming ICCP connection request MUST NOT be accepted unless its source IP address is known to be the source of an "eligible" ICCP peer. The set of eligible peers could be preconfigured (as a list of either IP addresses or address/mask combinations), or it could be discovered dynamically via some secure discovery protocol. The TCP Authentication Option (TCP-AO), as defined in [RFC5925], SHOULD be used. This provides integrity and authentication for the ICCP messages and eliminates the possibility of source address spoofing. However, for backwards compatibility and/or to accommodate the ease of migration, the LDP MD5 authentication key option, as described in Section 2.9 of [RFC5036], MAY be used instead.

The security framework and considerations for MPLS in general, and LDP in particular, as described in [RFC5920] apply to this document. Moreover, the recommendations of [RFC6952] and mechanisms of [LDP-CRYPTO] aimed at addressing LDP’s vulnerabilities are applicable as well.

Furthermore, activity on the attachment circuits may cause security threats or be exploited to create denial-of-service attacks. For example, a malicious CE implementation may trigger continuously varying LACP messages that lead to excessive ICCP exchanges. Also, excessive link bouncing of the attachment circuits may lead to the same effect. Similar arguments apply to the inter-PE MPLS links. Implementations SHOULD provide mechanisms to perform control-plane policing and mitigate these types of attacks.

11. Manageability Considerations

Implementations SHOULD generally minimize the number of parameters required to configure ICCP in order to help make ICCP easier to use. Implementations SHOULD allow the user to control the RGID via configuration, as this is required to support flexible grouping of PEs in RGs. Furthermore, implementations SHOULD provide mechanisms to troubleshoot the correct operation of ICCP; this includes providing mechanisms to diagnose ICCP connections as well as Application Connections. Implementations MUST provide a means for the user to indicate the IP addresses of remote PEs that are to be members of a given RG. Automatic discovery of RG membership MAY be supported; this topic is outside the scope of this specification.

12. IANA Considerations

12.1. Message Type Name Space

This document uses several new LDP message types. IANA maintains the "Message Type Name Space" registry as defined by [RFC5036]. The following values have been assigned:

Message Type	Description
0x0700	RG Connect Message
0x0701	RG Disconnect Message
0x0702	RG Notification Message
0x0703	RG Application Data Message
0x0704-0x070F	Reserved for future ICCP use

12.2. TLV Type Name Space

This document uses a new LDP TLV type. IANA maintains the "TLV Type Name Space" registry as defined by [RFC5036]. The following value has been assigned:

TLV Type	Description
0x0700	ICCP capability TLV

12.3. ICC RG Parameter Type Space

IANA has created a registry called "ICC RG Parameter Types", within the "Pseudowire Name Spaces (PWE3)" registry. ICC RG parameter types are 14-bit values. Parameter Type values 1 through 0x003A are specified in this document. Parameter Type values 0x003B through 0x1FFF are to be assigned by IANA, using the "Expert Review" policy defined in [RFC5226]. Parameter Type values 0x2000 through 0x2FFF, 0x3FFF, and 0 are to be allocated using the "IETF Review" policy defined in [RFC5226]. Parameter Type values 0x3000 through 0x3FFE are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in [RFC5226].

Initial ICC parameter type space value allocations are specified below:

Parameter Type	Description
0x0001	ICC Sender Name
0x0002	NAK TLV
0x0003	Requested Protocol Version TLV
0x0004	Disconnect Code TLV
0x0005	ICC RG ID TLV
0x0006-0x000F	Reserved
0x0010	PW-RED Connect TLV
0x0011	PW-RED Disconnect TLV
0x0012	PW-RED Config TLV
0x0013	Service Name TLV
0x0014	PW ID TLV
0x0015	Generalized PW ID TLV
0x0016	PW-RED State TLV
0x0017	PW-RED Synchronization Request TLV
0x0018	PW-RED Synchronization Data TLV
0x0019	PW-RED Disconnect Cause TLV
0x001A-0x002F	Reserved
0x0030	mLACP Connect TLV
0x0031	mLACP Disconnect TLV
0x0032	mLACP System Config TLV
0x0033	mLACP Port Config TLV
0x0034	mLACP Port Priority TLV
0x0035	mLACP Port State TLV
0x0036	mLACP Aggregator Config TLV
0x0037	mLACP Aggregator State TLV
0x0038	mLACP Synchronization Request TLV
0x0039	mLACP Synchronization Data TLV
0x003A	mLACP Disconnect Cause TLV

12.4. Status Code Name Space

This document uses several new Status codes. IANA maintains the "Status Code Name Space" registry as defined by [RFC5036]. The following values have been assigned; the "E" column is the required setting of the Status Code E-bit.

Range/Value	E	Description
0x00010001	0	Unknown ICCP RG
0x00010002	0	ICCP Connection Count Exceeded
0x00010003	0	ICCP Application Connection Count Exceeded
0x00010004	0	ICCP Application not in RG
0x00010005	0	Incompatible ICCP Protocol Version
0x00010006	0	ICCP Rejected Message
0x00010007	0	ICCP Administratively Disabled
0x00010010	0	ICCP RG Removed
0x00010011	0	ICCP Application Removed from RG

13. Acknowledgments

The authors wish to acknowledge the important contributions of Dennis Cai, Neil McGill, Amir Maleki, Dan Biagini, Robert Leger, Sami Boutros, Neil Ketley, and Mark Christopher Sains.

The authors also thank Daniel Cohn, Lizhong Jin, and Ran Chen for their valuable input, discussions, and comments.

14. References

14.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, October 2007.

[RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.

[RFC4447] Martini, L., Ed., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[IEEE-802.1AX]
IEEE Std. 802.1AX-2008, "IEEE Standard for Local and metropolitan area networks--Link Aggregation", IEEE Computer Society, November 2008.

[RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB", RFC 2863, June 2000.

[RFC6870] Muley, P., Ed., and M. Aissaoui, Ed., "Pseudowire Preferential Forwarding Status Bit", RFC 6870, February 2013.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, May 2013.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

14.2. Informative References

[RFC2922] Bierman, A. and K. Jones, "Physical Topology MIB", RFC 2922, September 2000.

[RFC4026] Andersson, L. and T. Madsen, "Provider Provisioned Virtual Private Network (VPN) Terminology", RFC 4026, March 2005.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, RFC 3629, November 2003.

[LDP-CRYPTO]
Zheng, L., Chen, M., and M. Bhatia, "LDP Hello Cryptographic Authentication", Work in Progress, June 2014.

Authors' Addresses

Luca Martini
Cisco Systems, Inc.
9155 East Nichols Avenue, Suite 400
Englewood, CO 80112
United States
EMail: lmartini@cisco.com

Samer Salam
Cisco Systems, Inc.
595 Burrard Street, Suite 2123
Vancouver, BC V7X 1J1
Canada
EMail: ssalam@cisco.com

Ali Sajassi
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
United States
EMail: sajassi@cisco.com

Matthew Bocci
Alcatel-Lucent
Voyager Place
Shoppenhangers Road
Maidenhead
Berks, SL6 2PJ
UK
EMail: matthew.bocci@alcatel-lucent.com

Satoru Matsushima
Softbank Telecom
1-9-1, Higashi-Shinbashi, Minato-ku
Tokyo 105-7304
Japan
EMail: satoru.matsushima@g.softbank.co.jp

Thomas Nadeau
Brocade
EMail: tnadeau@brocade.com