SFF-8601

Specification for

Speed Negotiation for Ethernet Drives

Rev 0.7 November 17, 2015

Secretariat: SFF TA TWG

Abstract: The initial introduction of HDDs supporting object storage solutions included those with an Ethernet interface. The initial offerings utilized the 1Gbps SGMII Ethernet variant. To enable enhanced cold storage applications, to expanded usage of Ethernet object drives beyond cold storage, and to provide interface speeds that do not present a bottleneck to data availability from HDDs, a standard, robust speed negotiation process is needed to provide a common method to move beyond 1Gbps SGMII while maintaining backwards compatibility to existing applications.

This specification provides a common reference for systems manufacturers, system integrators, and suppliers.

This specification defines the speed negotiation process for Ethernet drives, and the LLDP based port auto-negotiation protocol for Ethernet drives that do not support IEEE 802.3(tm)-2012 Clause 73 auto-negotiation.

This specification is made available for public review, and written comments are solicited from readers. Comments received by the members will be considered for inclusion in future revisions of this specification.

POINTS OF CONTACT:

Alvin Cox
Seagate Technology
10321 West Reno Avenue
Oklahoma City, OK 73127

Ph: 405-206-4809
Email: alvin.cox@seagate.com

Chairman SFF TA TWG
Email: SFF-Chair@snia.org

Change History:

Rev 0.0 September 9, 2015
- Cover page indicating the purpose of this specification.

Rev 0.1 September 14, 2015
- Initial draft with technical content.

Rev 0.2 October 2, 2015
- Updated Figure 5-1 and OUI value.

Rev 0.3 October 10, 2015
- Editorial update to current template style and fonts

Rev 0.4 October 16, 2015
- Significant editorial updates based on SSWG discussions. Changes are not tracked to preserve readability. All figures have minor test updates.

Rev 0.5 October 20, 2015
- Minor editorial updates and spelling corrections.

Rev 0.6 November 13, 2015
- Minor editorial updates to Figures 5-5 and 5-6.
- Updated Section 5.6 to include information of Section 5.7 and define the missing TLV format details.

Rev 0.7 November 17, 2015
- Removed minor editorial updates to Figures 5-5 and 5-6 of Rev 0.6.
- Added increment of failure counters to Figure 5-1 and Figure 5-4
- Editorial update to Table 5-4.

Foreword

The development work on this specification was done by the SNIA SFF TWG, an industry group. Since its formation as the SFF Committee in August 1990, the membership has included a mix of companies which are leaders across the industry.

When 2 1/2" diameter disk drives were introduced, there was no commonality on external dimensions e.g. physical size, mounting locations, connector type, connector location, between vendors. The SFF Committee provided a forum for system integrators and vendors to define the form factor of disk drives.

During their definition, other activities were suggested because participants in SFF faced more challenges than the form factors. In November 1992, the charter was expanded to address any issues of general interest and concern to the storage industry. The SFF Committee became a forum for resolving industry issues that are either not addressed by the standards process or need an immediate solution.

In July 2016, the SFF Committee transitioned to SNIA (Storage Networking Industry Association), as a TA (Technology Affiliate) TWG (Technical Work Group).

Industry consensus is not a requirement to publish a specification because it is recognized that in an emerging product area, there is room for more than one approach. By making the documentation on competing proposals available, an integrator can examine the alternatives available and select the product that is felt to be most suitable.

SFF meets during the T10 (see www.t10.org) and T11 (see www.t11.org) weeks, and SSWGs (Specific Subject Working Groups) are held at the convenience of the participants. Material presented to SFF becomes public domain, and there are no restrictions on the open mailing of the presented material by Members.

Many of the specifications developed by SFF have either been incorporated into standards or adopted as standards by ANSI, EIA, JEDEC and SAE.

For those who wish to participate in the activities of the SFF TWG, the signup for membership can be found at:
http://www.snia.org/sff/join

The complete list of specifications which have been completed or are currently being worked on by SFF can be found at:
http://www.snia.org/sff/specifications/SFF-8000.TXT

If you wish to know more about the SFF TWG, the principles which guide the activities can be found at:
http://www.snia.org/sff/specifications/SFF-8032.PDF

Suggestions for improvement of this specification will be welcome, they should be submitted to: http://www.snia.org/feedback

CONTENTS

FIGURES

TABLES

1. Scope

This specification defines the speed negotiation process for Ethernet drives, and the LLDP based port auto-negotiation protocol (AP) for Ethernet drives that do not support IEEE 802.3(tm)-2012 Clause 73 auto-negotiation.

1.1 Copyright

The SNIA hereby grants permission for individuals to use this document for personal use only, and for corporations and other business entities to use this document for internal use only (including internal copying, distribution, and display) provided that:

1. Any text, diagram, chart, table or definition reproduced shall be reproduced in its entirety with no alteration, and,
2. Any document, printed or electronic, in which material from this document (or any portion hereof) is reproduced shall acknowledge the SNIA copyright on that material, and shall credit the SNIA for granting permission for its reuse.

Other than as explicitly provided above, there may be no commercial use of this document, or sale of any part, or this entire document, or distribution of this document to third parties. All rights not explicitly granted are expressly reserved to SNIA.

Permission to use this document for purposes other than those enumerated (Exception) above may be requested by e-mailing copyright_request@snia.org. Please include the identity of the requesting individual and/or company and a brief description of the purpose, nature, and scope of the requested use. Permission for the Exception shall not be unreasonably withheld. It can be assumed permission is granted if the Exception request is not acknowledged within ten (10) business days of SNIA's receipt. Any denial of permission for the Exception shall include an explanation of such refusal.

1.2 Disclaimer

The information contained in this publication is subject to change without notice. The SNIA makes no warranty of any kind with regard to this specification, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The SNIA shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance, or use of this specification.

Suggestions for revisions should be directed to http://www.snia.org/feedback/

2. References

2.1 Industry Documents

- IEEE 802.1AB(tm)-2009 - Station and Media Access Control Connectivity Discovery
  referred to herein as 802.1AB
- IEEE Std 802.3(tm)-2012 - IEEE Standard for Ethernet
  referred to herein as 802.3
- IEEE 802.3(tm)-2012 Clause 73 Auto Negotiation
  referred to herein as CL73
- IEEE 802.3(tm)-2012 Clause 79
  referred to herein as CL79
- Serial-GMII Specification Revision 1.7, July 20, 2001

2.2 Sources

There are several projects active within the SFF TWG. The complete list of specifications which have been completed or are still being worked on are listed in http://www.snia.org/sff/specifications/SFF-8000.TXT

Copies of ANSI standards may be purchased from the InterNational Committee for Information Technology Standards (http://www.techstreet.com/incitsgate.tmpl).

2.3 Conventions

The dimensioning conventions are described in ANSI-Y14.5M, Geometric Dimensioning and Tolerancing. All dimensions are in millimeters, which are the controlling dimensional units (if inches are supplied, they are for guidance only).

The ISO convention of numbering is used i.e., the thousands and higher multiples are separated by a space and a period is used as the decimal point. This is equivalent to the English/American convention of a comma and a period.

American	French	ISO
0.6	0,6	0.6
1,000	1 000	1 000
1,323,462.9	1 323 462,9	1 323 462.9

2.4 Definitions and Abbreviations

For the purpose of this SFF Specifications, the following definitions and abbreviations apply:

AP: Auto-negotiation protocol.

HCD: Highest common denominator.

LLDP: Link Layer Discovery Protocol.

Local port: The designator for a port establishing a point of reference when describing interaction between partner ports. Either port may be described as the local port, depending on context.

Partner ports: Two ports at opposite ends of a link. When describing interaction between the partner ports, one is referred to as the local port and the other, the remote port.

PDU: Protocol Data Unit (i.e., protocol payload).

Remote Port: The designator for a port establishing a point of reference when describing interaction between partner ports. Either port may be described as the remote port, depending on context.

SGMII: (Serial Gigabit Media Independent Interface.

SFF 1G: See Serial-GMII Specification Revision 1.7, July 20, 2001.

SFF 2.5G: A scaled version of SFF 1G at 2.5 times the specified rate. Encoding is 8b/10b with no scrambling and transmitter training is not supported.

SFF 5G: A scaled version of 10GBASE-KR at 0.5 times the rate and corresponding channel characteristics. The encoding is 64b/66b with scrambling and transmitter training is optional.

SW: Software

TLV: Type-Length-Value.

UCT: Unconditional transition.

3. General Description

This specification defines the speed negotiation process for Ethernet drives, and the LLDP based port auto-negotiation protocol for Ethernet drives that do not support CL73 auto-negotiation.

4. Speed Negotiation Sequence

The speed negotiation sequence shall follow the sequence defined in this section. Refer to definitions for a description of the speed references SFF 1G, SFF 2.5G, and SFF 5G.

1. At power up, ports enter CL73 auto-negotiation. If not implemented, then proceed to step 3.

2. The Ethernet PHYs start CL73 auto-negotiation and complete with a mutual common denominator. Link training runs (if applicable) and end-to-end link is established.

    Note 1: If the link is not established (as per respective 802.3 standard, "link status=fail"), then a different set of capabilities may be advertised and the negotiation process repeated. This is currently done by some implementations and is a behavior outside of the standard, but is allowed.

    Note 2: If the Ethernet PHY enters auto-negotiation and the link remote port does not (and sends an SFF 1G signal), then auto-negotiation allows for "parallel detect" on its receiver to support Non-CL73 auto-negotiation operation.

3. Non-CL73 auto-negotiation operation: If CL73 auto-negotiation is not supported or fails, then initiate connection at SFF 1G from both ports. All SFF ports shall support SFF 1G.

4. Non-CL73 auto-negotiation operation: If a port is capable of higher speeds/protocols, then send a capabilities packet via Link Layer Discovery Protocol (LLDP) specified in CL79 and 802.1AB. LLDP uses simple Type-Length-Value (TLV) format on a well-known protocol format to indicate end-point capabilities (or anything else). In this case, alternate speeds in addition to SFF 1G may be indicated through this protocol (i.e., SFF 2.5G and SFF 5G) and allow for auto-negotiation of a speed higher than SFF 1G, but less than 10GBASE-KR.

5. Non-CL73 auto-negotiation operation: If both ports are capable of higher speeds/protocols, then disconnect the SFF 1G link, and reconnect at the highest common denominator (HCD).

6. Non-CL73 auto-negotiation operation: If the ports fail to sync at the HCD, then try at the next lower supported common denominator.

Section 5 of this specification defines information required for support of LLDP-based port auto-negotiation.

5. LLDP-Based Port Auto-Negotiation.

5.1 High Level Flow

Figure 5-1 illustrates the high level flow of the LLDP-based port auto-negotiation process between the local port and the remote port. The process is separated into three parts, A, B, and C. Part A illustrates initialization of the connection at SFF 1G and then enabling LLDP. Parts B and C illustrate the determination of the HCD and the process of establishing the link at the fastest common rate that the system supports. Figure 5-1 shows the overall flow including Part A, Part B, and Part C while Figure 5-2, Figure 5-3, and Figure 5-4 provide additional detail of Part A, Part B, and Part C, respectively.

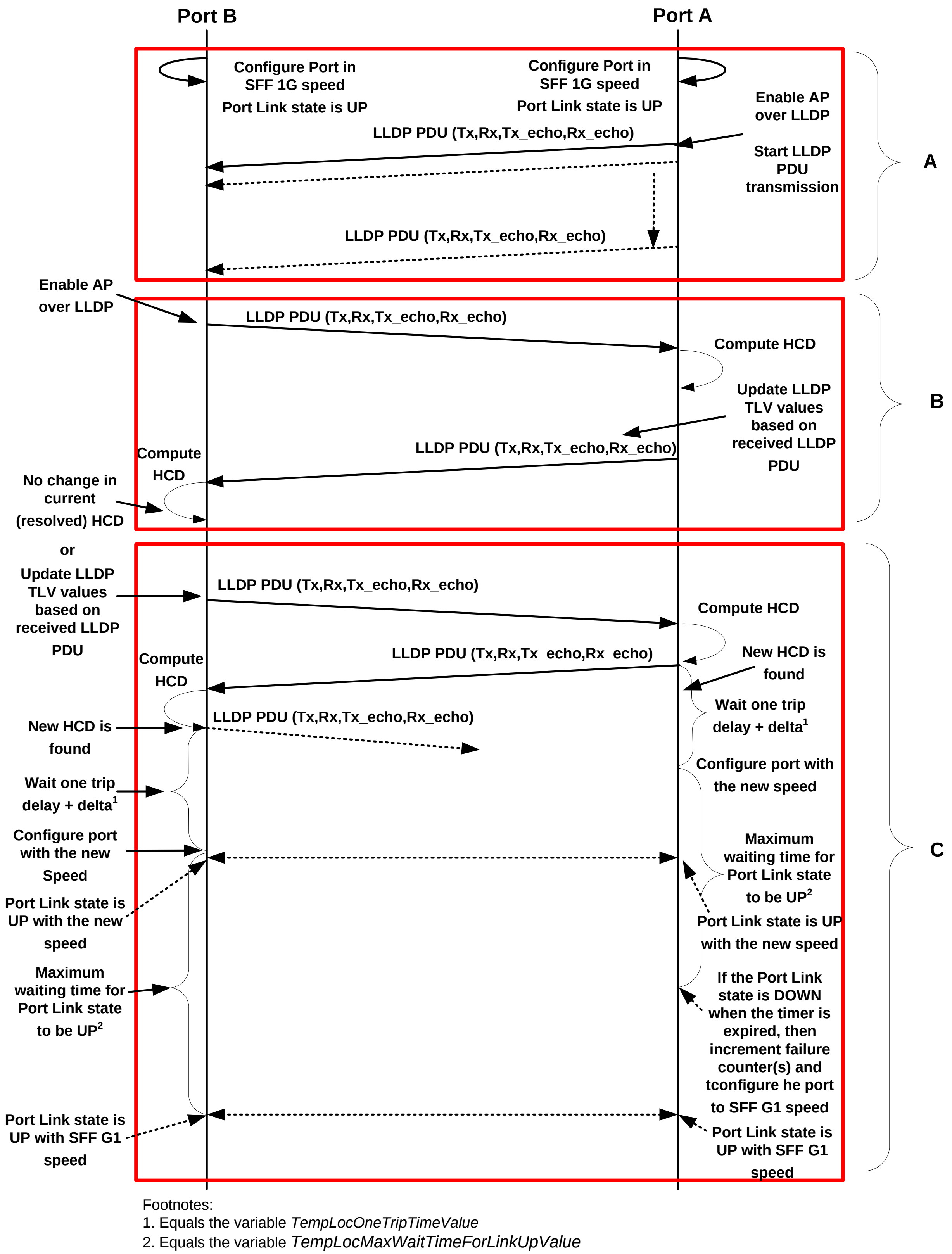

FIGURE 5-1 HIGH LEVEL FLOW

Part A

1. Both the local port and remote ports are configured to default values with a port speed of SFF 1G.
2. LLDP is enabled on one of the partner ports (Figure 5-2, port A) and that port starts sending LLDP PDUs to its remote port at the default speed of SFF 1G. Since LLDP is not enabled in the remote port (Figure 5-2, port B), no negotiation is performed, and both ports' speeds remain with the default configuration (SFF 1G).

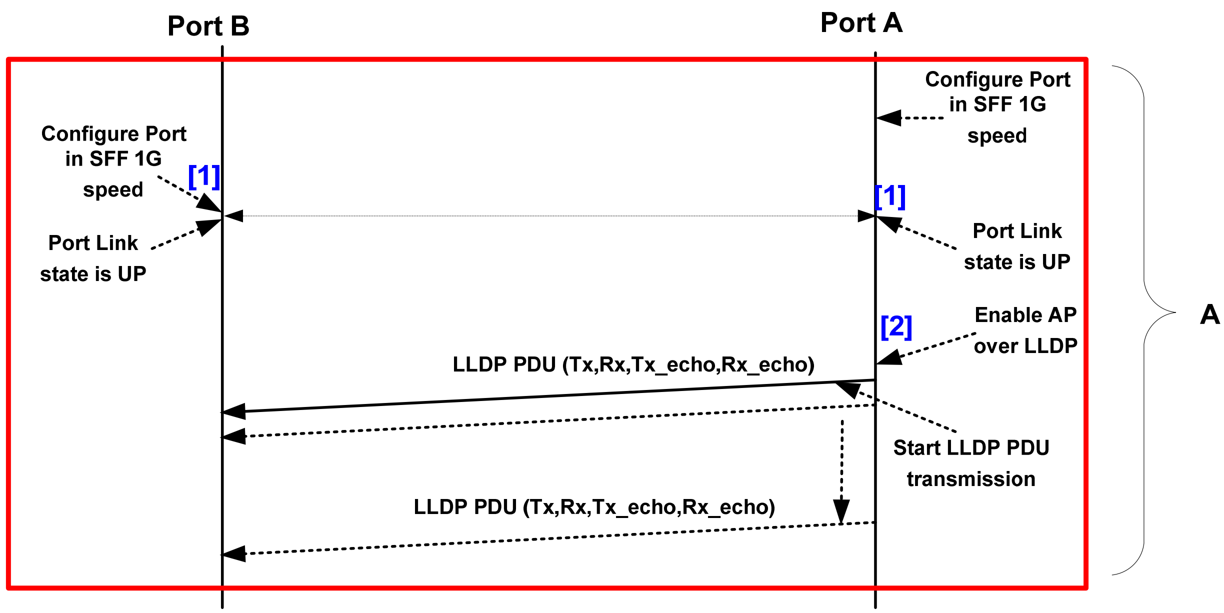

FIGURE 5-2 HIGH LEVEL FLOW PART A

Part B

3. LLDP is enabled on both port A and port B in
4. Figure 5-3. One or more LLDP PDUs are sent and received in each port. The frequency is specified in
   802.1AB Section 9. The default period is every 2 seconds.
5. In every received LLDP PDU, each port computes the HCD. The example in
6. Figure 5-3 indicates that the current (resolved) HCD and computed HCD are the same in both sides.
7. LLDP TLV parameters values are updated in transmitted PDUs based on values in received PDUs.

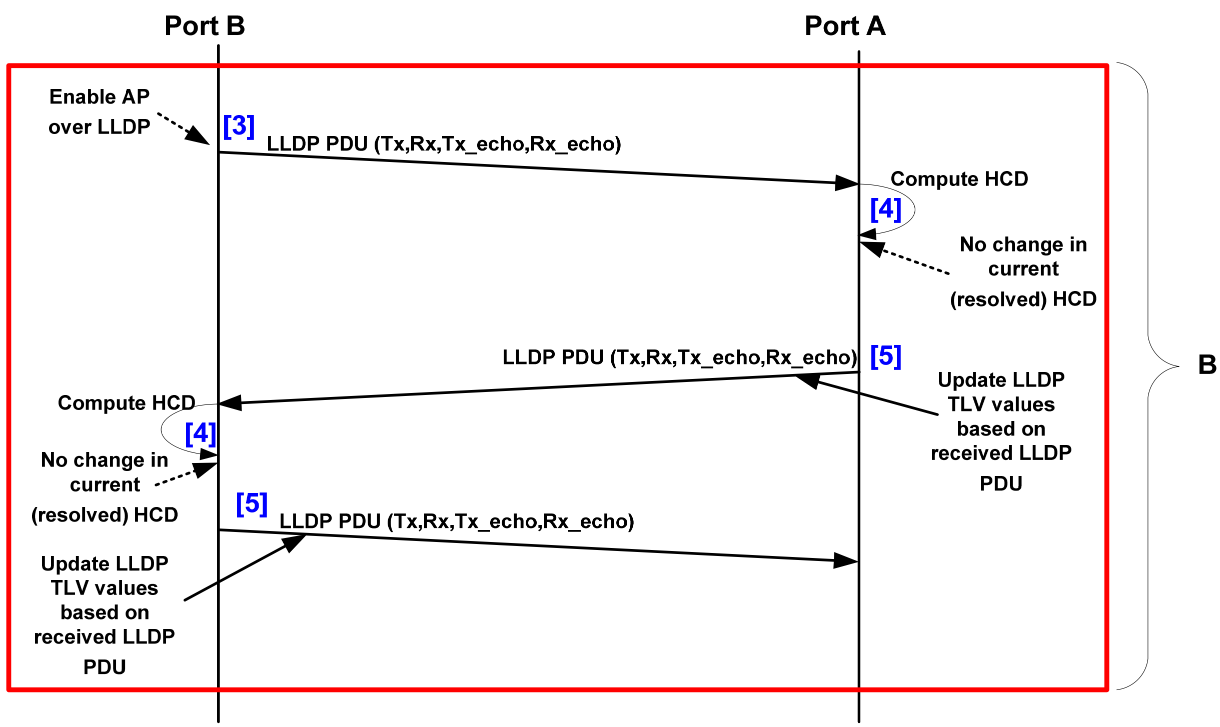

FIGURE 5-3 HIGH LEVEL FLOW PART B

Part C

8. Each port checks independently if a new HCD was found. If yes, then another LLDP PDU is sent to its partner port and the local port waits *TempLocOneTripTimeValue* (see Table 5-1) to ensure that the transmitted PDU is received by its remote port.
9. Each port is configured with the new HCD capabilities. If the Port Link status is DOWN, the port waits *TempLocMaxWaitTimeForLinkUpValue* (see Table 5-1) until its partner port configures its port with the same new port capabilities.
10. The Port Link state is UP in both ports with the new speed mode (e.g., SFF 5G). When done, goes to Part B step 3.
11. If the Port Link state is DOWN when the timer (i.e., *TempLocMaxWaitTimeForLinkUpValue)* is expired, then update the failure counter(s) and the port returns to the default configuration with a speed of SFF 1G and goes to Part B step 3.

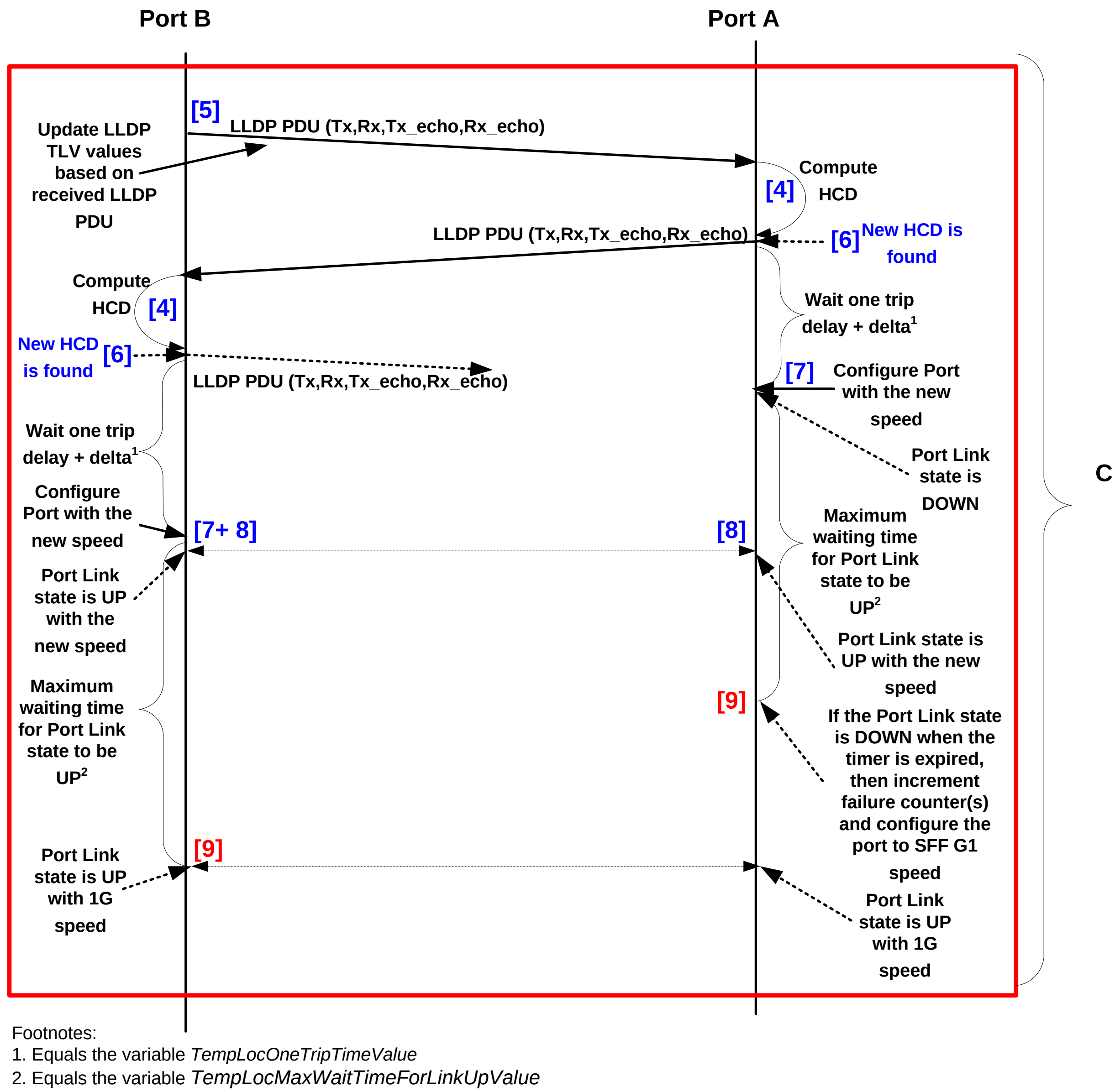

FIGURE 5-4 HIGH LEVEL FLOW PART C

5.2 LLDP Parameters, Variables, and Constants

Table 5-1 lists the LLDP parameters, variables, and Constants.

TABLE 5-1 LLDP PARAMETERS, VARIABLES, AND CONSTANTS

Parameter	Description
AdvertisementAbility	A bit array that contains the Advertised Abilities of the link. This bit array includes · Technology ability (25-bits) - Define the local port mode capabilities as defined in CL73 with taking reserved bits for SFF 1G, SFF 2.5G, and SFF 5G · Pause capability (2-bits) · FEC capabilities (2-bits)
TLV Parameters	**Description**
PortAdvValue	Indicates the value of *AdvertisementAbility* that the local port supports
PortAdvValueEcho	Indicates the value of receive *PortAdvValue* echoed back by the port
FailureIndication	Indicates a failure value of the local port to the remote port
Internal Variables	**Description**
LocPortAdvValue	Indicates the value of *PortAdvValue* of the local port
LocPortAdvValueEcho	Indicates the value of receive *PortAdvValueEcho* echoed back by the local port
LocFailureIndication	Indicates a failure value of the local port sent to the remote port
RemPortAdvValue	Indicates the value of *PortAdvValue* of the remote port
RemPortAdvValueEcho	Indicates the value of receive *PortAdvValueEcho* echoed back by the remote port
RemFailureIndication	Indicates a failure value *FailureIndication* sent by the remote port
TempLocPortAdvValue	Variable used to store value of *LocPortAdvValue*
TempLocPortAdvValueEcho	Variable used to store value of *LocPortAdvValueEcho*
TempRemPortAdvValue	Variable used to store value of *RemPortAdvValue*
TempRemPortAdvValueEcho	Variable used to store value of *RemPortAdvValueEcho*
LocOneTripTimeValue	Indicates the one trip time (latency) of single packet to be sent from the local port and being received in the remote port
LocMaxWaitTimeForLinkUpValue	Indicates the maximum time the local port waits for the ports to be in the Port Link state is UP
TempLocOneTripTimeValue	Variable used to store a value of time
TempLocMaxWaitTimeForLinkUpValue	Variable used to store a value of time
LocResolvedHCDValue	Indicates the current HCD used between the two partner ports
LocComputedHCDValue	Indicates the computed HCD based the latest parameters
RxChanged	Boolean Variable to indicate any change found in the Rx state machine
Constants	**Description**
DEFAULT TX VALUE	Indicates the default Tx value of the port (SFF 1G)
DEFAULT RX VALUE	Indicates the default Rx value of the port (SFF 1G)

5.3 Rx State Machine

Figure 5-5 defines the Rx state machine.

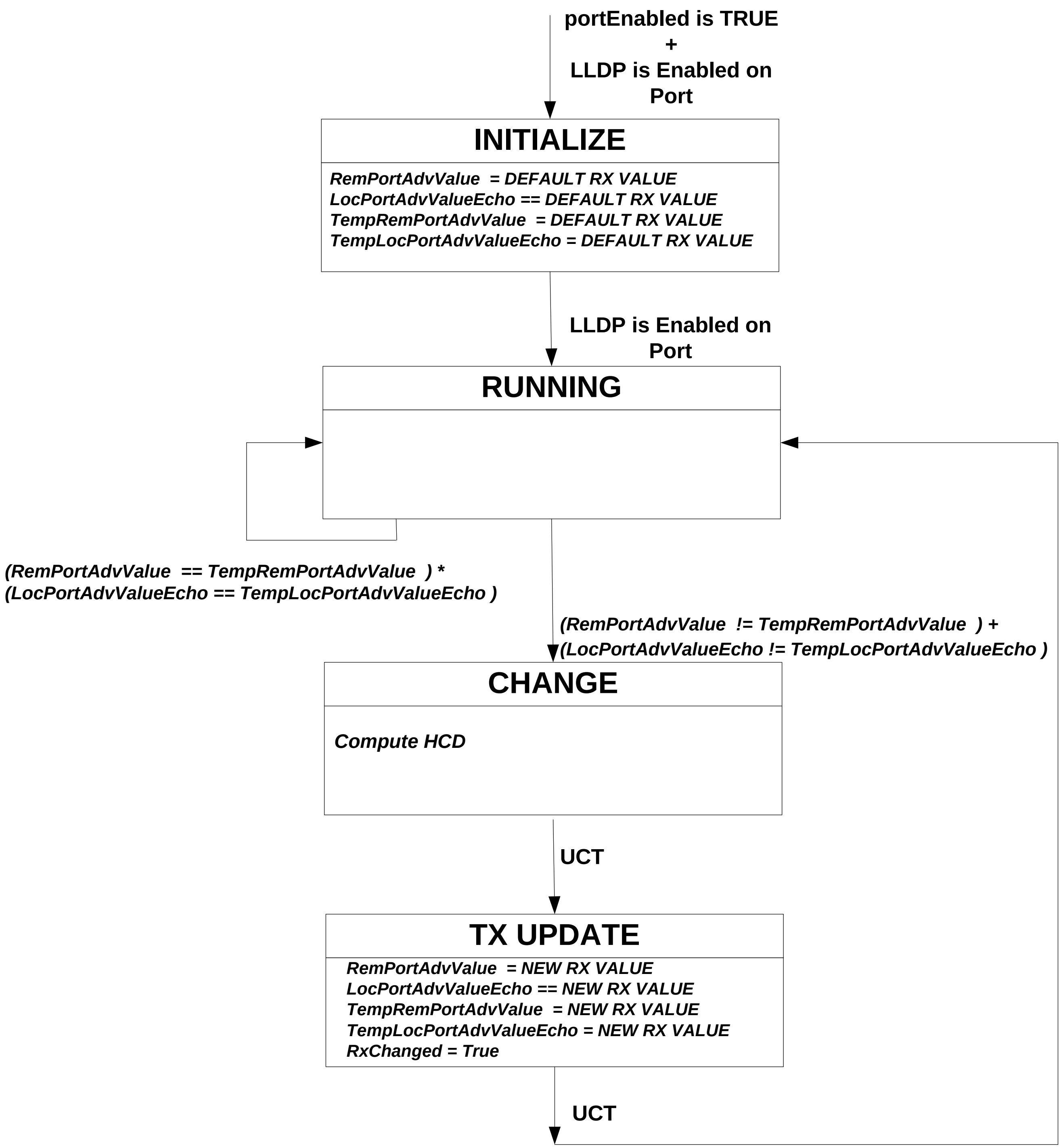

FIGURE 5-5 RX STATE MACHINE

INITIALIZE state
- Initialize all variables to default
- When LLDP is enabled, enter RUNNING state

RUNNING state

- When a new LLDP PDU is received, check the LLDP TLV values
- If no CHANGE, then stay in RUNNING state; otherwise, enter CHANGE state

CHANGE state

- Compute the new HCD based the new incoming LLDP PDU values, and enter TX_UPDATE state

TX_UPDATE state

- Update the Tx parameters values used in LLDP transmission PDU
- Trigger TX state machine regarding the new computed HCD
- Enter RUNNING state

5.4 TX State Machine

Figure 5-6 defines the Tx state machine.

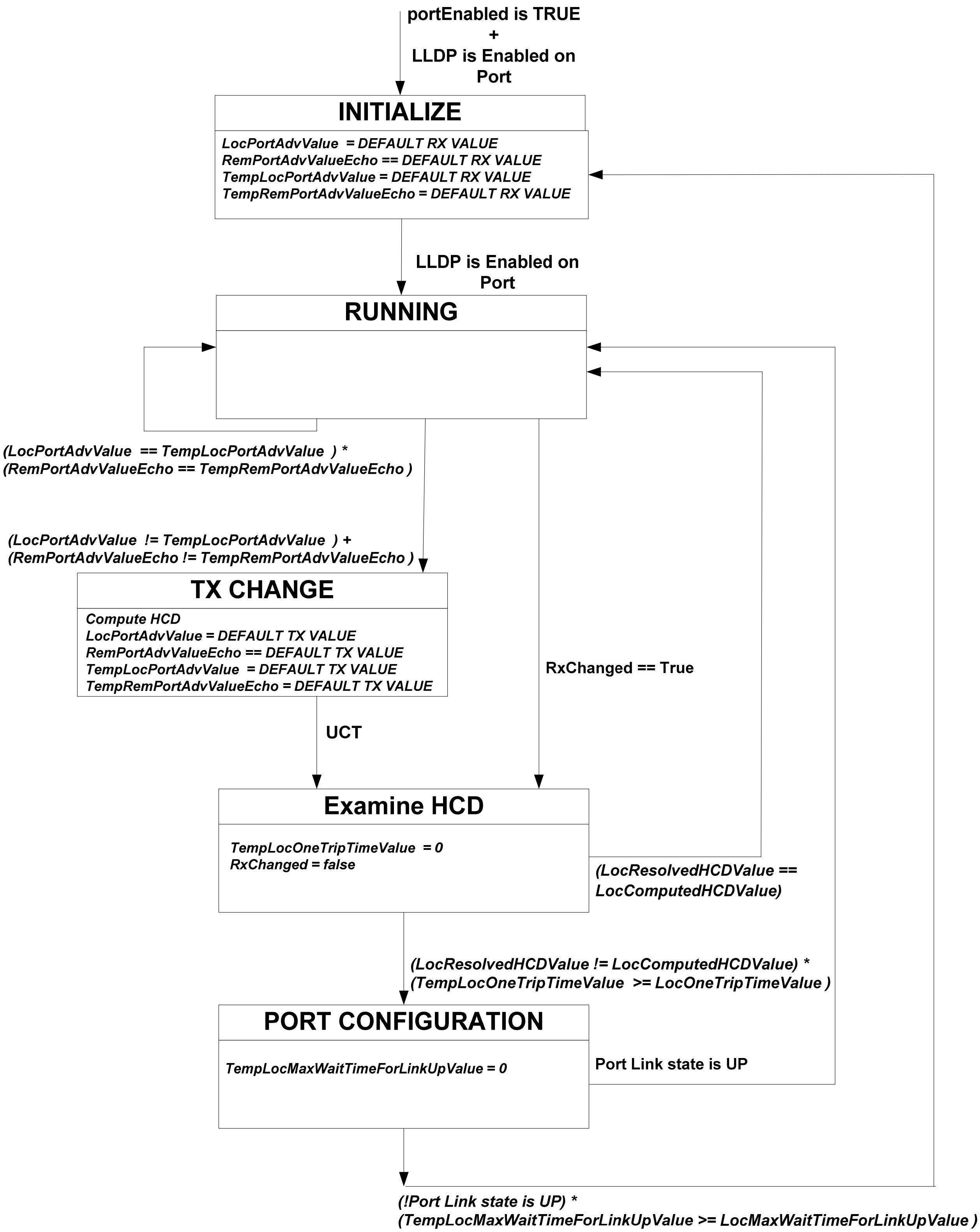

FIGURE 5-6 TX STATE MACHINE

INITIALIZE state
- Initialize all variables to default
- When LLDP is enabled, enter RUNNING state

RUNNING state
- Check if a new HCD was computed by the Rx state machine (i.e., RxChange = true). If yes, then enter EXAMINE_HCD state
- Check if port capabilities were changed. If yes, then enter TX_CHANGE state and then enter EXAMINE_HCD state

EXAMINE_HCD state
- Check if the new HCD is different from the current (resolved) HCD. If yes, then enter PORT_CONFIGURATION state; otherwise enter RUNNING state

PORT_CONFIGURATION state
- Update the local port with the new HCD capabilities and wait until the remote port does the same
- On Port Link state is UP enter RUNNING state, otherwise enter INITIALIZE state and configure the local port to the default configuration with a speed of SFF 1G.

5.5 Trip Delay

5.5.1 One Trip Delay Measurement

Indicates the one trip time (latency) of single packet to be sent from the local port and being received by the remote port.

5.5.2 System Configuration

The one trip time parameter value is configured explicitly as part of whole related system parameters. Default value is 30 msec.

5.6 SFF LLDP Specific TLV

5.6.1 Port Auto-Negotiation - TLV Format

Port Auto-Negotiation is used to exchange information about the port capabilities. Figure 5-7 shows the format of this TLV

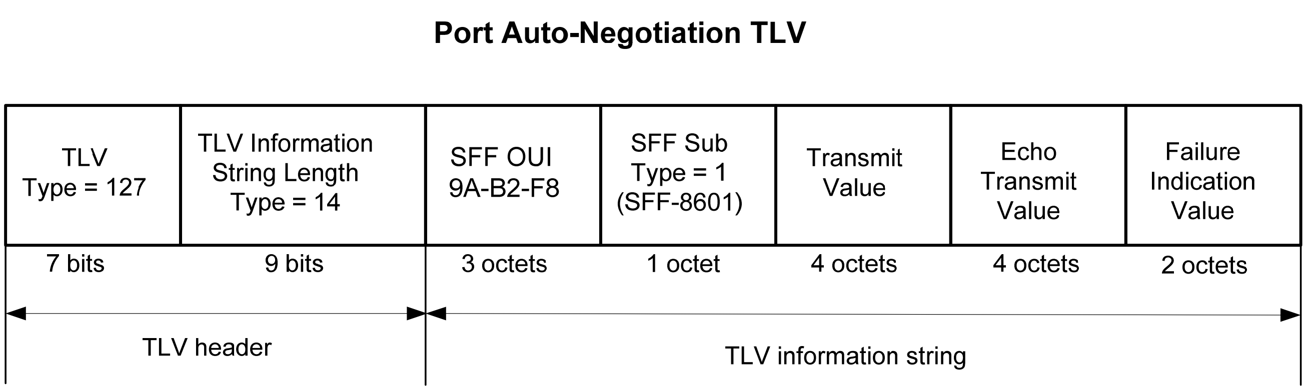

FIGURE 5-7 PORT AUTO-NEGOTIATION TLV

5.6.2 TLV Type and Subtype

Table 5-2 defines the TLV subtype.

TABLE 5-2 TLV SUBTYPE

TLV Type	SFF subtype	TLV name
127	1 (SFF-8601)	Port Auto-Negotiation

TLV Transmit Value

Table 5-3 defines the port auto-negotiation TLV Transmit Value field.

The TLV Transmit Value contains the supported transmit and receive capabilities (*PortAdvValue*) and the current speed settings of the port. The current speed settings of the port reflect the initial speed of SFF 1G or the HCD determined during the auto-negotiation process. The Failure Indication Value information is combined with the *PortAdvValue* and current port speed settings to determine the HCD.

TABLE 5-3 PORT AUTO-NEGOTIATION TLV TRANSMIT VALUE

Bit	Parameter	Description
0-1	*PortAdvValue*	This value indicates the SFF speeds supported by the port transmitter: 00 = Not valid 01 = SFF 1G only 10 = SFF 2.5G and SFF 1G 11 = SFF 5G, SFF 2.5G, and SFF 1G
2-7	Reserved	
8-9	*PortAdvValue*	This value indicates the SFF speeds supported by the port receiver: 00 = Not valid 01 = SFF 1G only 10 = SFF 2.5G and SFF 1G 11 = SFF 5G, SFF 2.5G, and SFF 1G
10-15	Reserved	
16-17	Current transmitter speed setting	This value indicates the current SFF speed setting of the port transmitter: 00 = Not valid 01 = SFF 1G 10 = SFF 2.5G 11 = SFF 5G
18-23	Reserved	
24-25	Current receiver speed setting	This value indicates the current SFF speed setting of the port receiver: 00 = Not valid 01 = SFF 1G 10 = SFF 2.5G 11 = SFF 5G
26-31	Reserved	

5.6.3 TLV Echo Transmit Value

Table 5-4 defines the port auto-negotiation TLV Echo Transmit Value field.

The TLV Echo Transmit Value contains the transmitter and receiver capabilities (*PortAdvValue*) and the current speed settings values of the remote port received by the local port to be transmitted back to the remote port.

TABLE 5-4 PORT AUTO-NEGOTIATION TLV ECHO TRANSMIT VALUE

Bit	Parameter	Description
0-1	*PortAdvValueEcho*	This value indicates the received SFF speeds supported by the remote port transmitter: 00 = No data received 01 = SFF 1G only 10 = SFF 2.5G and SFF 1G 11 = SFF 5G, SFF 2.5G, and SFF 1G
2-7	Reserved	
8-9	*PortAdvValueEcho*	This value indicates the received SFF speeds supported by the remote port receiver: 00 = No data received 01 = SFF 1G only 10 = SFF 2.5G and SFF 1G 11 = SFF 5G, SFF 2.5G, and SFF 1G
10-15	Reserved	
16-17	*PortAdvValueEcho* (Tx_echo)	This value indicates the received SFF speed setting of the remote port transmitter: 00 = No data received 01 = SFF 1G 10 = SFF 2.5G 11 = SFF 5G
18-23	Reserved	
24-25	*PortAdvValueEcho* (Rx_echo)	This value indicates the received SFF speed setting of the remote port receiver: 00 = No data received 01 = SFF 1G 10 = SFF 2.5G 11 = SFF 5G
26-31	Reserved	

5.6.4 Failure Indication Value

Table 5-5 defines the port auto-negotiation Failure Indication Value field.

The Failure Indication Value provides a record of occurrence and a count of auto-negotiation failures. The failure indications and counters are reset to 0 at power up or at the beginning of an auto-negotiation sequence.

TABLE 5-5 PORT AUTO-NEGOTIATION FAILURE PARAMETERS

Bit	Parameter	Description
0	*PortLinkUpFailureIn1G*	Indicates a hardware link up failure in SFF 1G
1	*PortLinkUpFailureIn2.5G*	Indicates a hardware link up failure in SFF 2.5G
2-4	*PortLinkUpFailureCntrIn2.5G*	3-bit binary counter that indicates the number of hardware link up failures in SFF 2.5G
5	*PortLinkUpFailureIn5G*	Indicates a hardware link up failure in SFF 5G
6-8	*PortLinkUpFailureCntrIn5G*	3-bit binary counter that indicates the number of hardware link up failures in SFF 5G
9-11	*MaxWaitTimeForLinkUpExceededCntr*	3-bit binary counter that indicates the number of times the *LocMaxWaitTimeForLinkUp* timer was exceeded
12-15	Reserved	