Intel® Ethernet 25G Media Guide

Application Note

NEX Cloud Networking Group (NCNG)

May 2023

Revision 1.2
630739-001

Revision History

Revision	Date	Comments
1.2	May 24, 2023	Initial public release.

Contents

NOTE: ***This page intentionally left blank.***

1.0 Introduction to 25G Interoperability

This document is intended to provide engineering insight regarding 25G PHY configurations as noted by the IEEE standard, and to address any observed interoperability issues related to the use and deployment of 25G media in enterprise ecosystems based on customer use cases. This application note focuses specifically on the 25G PHY link modes associated with our Intel® Ethernet Network Interface Card (NIC), System on Chip (SoC), and LAN on Motherboard (LOM) Portfolios, as well as the common media used for 25G Ethernet applications; 25G Twinaxial passive Direct Attach Copper (DAC) cables, 25G optical transceivers, and 25G backplane.

Understanding 25G interoperability requires familiarity with the concept of a PHY link mode as it is defined in the 25G IEEE standard and the 25G/50G Consortium Specification. PHY link modes are sometimes referred to as the PHY mode or link mode in some 25G Ethernet arenas.

The IEEE 802.3-2018 standard defines categories for PHY link modes that did not exist in prior Ethernet generations. It also mandates certain sets of capabilities required by each link partner to guarantee correct operation with various 25G media. The focus on the updated clauses is mainly on the different Forward Error Correction (FEC) modes required, and whether Auto-Negotiation (Clause 73) is required for the type of media used in 25G applications.

Most 25G switches and link partners in the ecosystem are defined by the degree to which they comply to these standards and specifications. Some support IEEE 802.3 only, while others only adhere to the Consortium specifications. For example, the support of 25G Auto-Negotiation and the various FEC modes among existing switches is mixed and highly dependent on when the switch was manufactured. Some of the 25G switches entered the ecosystem before the IEEE standard added these clauses, and as such, they do not support some of these features.

The main sections of this document describe:

- 25G Auto-Negotiation
- 25G FEC Clauses
- 25G PHY Link Modes
- 25G Interoperability Debug Flows

1.1 References

- *IEEE 802.3-2018 standard*
  - Table 110-10, “Cable Assembly Characteristic Summary”, section 110.10
  - Table 110C-1, “Host and Cable Assembly Combination Annex”, section 110C
- *25/50 Gigabit Ethernet Consortium Specification*, Rev 1.6
- *MSA/SFF-8024*, Rev 4.6
- *MSA/SFF-8636*, Rev 2.10a
- *MSA/SFF-8472*, Rev 12.3

2.0 25G Auto-Negotiation

Auto-negotiation (AN) is the mechanism by which two networking devices can advertise their capabilities for the purpose of determining link partner capabilities, and for resolving speed and FEC options. Auto-Negotiation is performed by the advertisement of Differential Manchester Encoding (DME) pages.

The capabilities advertised are:

- Supported link speeds for each device. For example: 25G, 10G, 2.5G, 1G.
- Supported FEC modes for each device. For example: no FEC, BASE-R FEC, RS-528 FEC.
- Requested FEC mode for each device.

For the purposes of this document, IEEE 802.3-2018, Section 5, Clause 73 defines Auto-Negotiation as specified for use with:

- 25G Ethernet PHYs operating over a backplane (25GBASE-KR or 25GBASE-KR-S)
- 25G Ethernet PHYs operating over a copper cable assembly (25GBASE-CR or 25GBASE-CR-S)

Note: When deploying configurations for switch link partners using copper cable assemblies, enable Auto-Negotiation if the switch supports it.

2.1 25G FEC Clauses

Forward Error Correction (FEC) is a coding technique that enhances data reliability by inserting redundant bits into the data stream to drastically reduce the probability of the actual data bits getting corrupted due to channel noise or other factors. FEC is absolutely necessary to achieve a lower Bit Error Ratio (BER) on a given communication channel. In the context of 25G Ethernet, the channel refers to the total length of the media to the link partner plus the length of the trace on the board of the component itself.

In 25G Ethernet, there are two primary FEC algorithms in use:

- BASE-R FEC (FC-FEC, Fire-Code FEC, Clause 74)
- RS-FEC (Reed Solomon FEC, RS528-FEC, Clause 108)

BASE-R FEC is a moderately capable encoding that aims to fix burst errors that are caused by Distributed Forwarding Engine (DFE) errors. However, the random error correction capability of BASE-R FEC is comparatively weaker than other FEC algorithm implementations.

RS-FEC is a much more capable FEC implementation than BASE-R FEC due to its 528-bit cyclic encoding.

The different FEC algorithms and their encodings as defined in IEEE 802.3 offer the following options:

- Lower capability to fix random errors with a lower latency
- Higher capability to fix random errors with a higher latency

Regardless of the FEC implementation, for 25G Ethernet PHYs, any additional encoding and decoding of the data stream will always cause network latency:

- Latency for BASE-R FEC is 80 ns.
- Latency for RS-FEC is 250 ns.

BASE-R FEC (or even no FEC) is used if the BER is sufficiently low. RS-FEC is used if there are channel impairments such as channel length, channel loss, or crosstalk.

If the channel quality is such that the BER is equal to or higher than $5x10^{-5}$, both link partners must use RS-FEC for reliable operation. Alternatively, if the channel quality introduces BER higher than $5x10^{-8}$, BASE-R FEC must be used by both link partners. If the BER can be established to be lower than 10^{-12}, FEC is not required.

2.2 25G PHY Link Modes

The following 25G PHY link modes are implemented in our 25G and 100G Ethernet product families and are defined in IEEE 802.3-2018, Section 7.

2.2.1 25GBASE-CR/25GBASE-CR-S

The PHY link modes 25GBASE-CR and 25GBASE-CR-S are intended for operation with passive DAC cables and differ only in their support for RS-FEC. As shown in Figure 1, Auto-Negotiation is required for 25GBASE-CR/CR-S PHYs.

Table 110–1—Physical Layer clauses associated with the 25GBASE-CR and 25GBASE-CR-S PMDs

Associated clause	25GBASE-CR	25GBASE-CR-S
106—RS	Required	Required
106—25GMII[a]	Optional	Optional
107—PCS	Required	Required
74—BASE-R FEC[b]	Required	Required
108—RS-FEC[b]	Required	N/A
109—PMA	Required	Required
109A—25GAUI C2C	Optional	Optional
73—Auto-Negotiation	Required	Required
78—Energy Efficient Ethernet	Optional	Optional

[a]The 25GMII is an optional interface. However, if the 25GMII is not implemented, a conforming implementation must behave functionally as though the RS and 25GMII were present.
[b]FEC sublayers can be enabled or disabled according to the FEC mode (see 110.6).

Figure 1. Table 110-1 (IEEE 802.3-2018, Section Seven, Page 627)

Figure 2 demonstrates that 25GBASE-CR/CR-S PHYs operate over three different types of cable assemblies: CA-25G-N (CA-N), CA-25G-S (CA-S), and CA-25G-L (CA-L).

Each 25GBASE-CR/CR-S PHY and copper cable assembly supports different FEC modes.

Table 110–2—Cable assembly types supported by each PHY type

PHY type	CA-25G-N	CA-25G-S	CA-25G-L
25GBASE-CR	Yes	Yes	Yes
25GBASE-CR-S	Yes	Yes	No

Figure 2. Table 110-2 (IEEE 802.3-2018, Section Seven, Page 627)

Table 1 has a listed summarization of the various cable assembly types alongside their supported FEC modes.

Table 1. 25G Host/Cable Assembly Combinations (IEEE 802.3-2018, Annex 110C)

Cable Assembly Form Factor	Cable Assembly Type	First Cage Connector	Second Cage Connector	Length[1]	FEC Modes Supported[2]
SFP28 to SFP28 (see 110C.3.1)	CA-25G-L	SFP28 form factor (see 110C.2.1)	SFP28 form factor (see 110C.2.1)	3-5 m	RS-FEC
	CA-25G-S			2-3 m	RS-FEC, BASE-R FEC
	CA-25G-N			1-2 m	RS-FEC, BASE-R FEC, no FEC
QSFP28 to QSFP28 (see 110C.3.2)	CA-25G-L	QSFP28 form factor (see 110C.2.2)	QSFP28 form factor (see 110C.2.2)	1-5 m	RS-FEC
QSFP28 to 4×SFP28 (see 110C.3.3)	CA-25G-L	QSFP28 form factor (see 110C.2.2)	SFP28 form factor (see 110C.2.1)	3-5 m	RS-FEC
	CA-25G-S			2-3 m	RS-FEC, BASE-R FEC
	CA-25G-N			1-2 m	RS-FEC, BASE-R FEC, no FEC

1. Indicates the average and achievable length of compliant cable assemblies. It might be possible to construct compliant cable assemblies longer than indicated. Length of the cable assembly does not imply compliance to specifications.
2. FEC mode is selected through Auto-Negotiation (Clause 73). See 110.6 IEEE 802.3-2018.

The IEEE standard defines three categories of passive DAC cables based on the electrical properties of these cables. Listed above are the typical length of the DAC for each of the three categories of cable assembly alongside their supported FEC mode required for proper operation.

Note: The length of the DAC is typical only to the availability of the cable assembly in the ecosystem and is not to be taken literally when determining FEC requirements or electrical operation.

For example, it is possible to find a 3 m DAC with the designation CA-25G-L, the category for the lowest electrical performance, and another 3 m DAC with the designation CA-25G-N, which is the category for the best electrical performance. The wire gauge used in manufacturing these DAC range from 22 to 32 AWG, which explains why two DAC of the same length might have different electrical performance. Determining which DAC has which FEC mode support is one of the many common causes for issues with interoperability in 25G networks. FEC requirements did not exist in the 10G network ecosystem where, for the most part, cables were identified just by length.

Note: For 25G DAC, the real difference between these cables is in the insertion loss limit measured and correlated to each cable length.

Table 2 demonstrates the insertion loss and minimum differential return loss per cable assembly:

Table 2. Insertion Loss and Min Differential Return Loss

Category Assembly Type	Length (meters)	Insertion Loss @ 12.899 GHz		Min Differential Return loss @ 12.89 GHz (dB)	FEC Modes Supported	EEPROM Code[1]
		Max	Min			
CA-25G-L	5	22.48	8	6	RS-FEC	0BH
CA-25G-S	3	16.48	8	6	RS-FEC, BASE-R FEC	0CH
CA-25G-N	3	15.5	8	6	RS-FEC, Base-R FEC, No FEC	0DH

1. Compliance code was read through SFF-8472 A0 page, Byte 36.

2.2.2 25GBASE-KR/25GBASE-KR-S

The PHY link modes 25GBASE-KR and 25GBASE-KR-S are intended for operation within 25G Ethernet Backplanes. A backplane is a specifically defined connector for chip-to-chip applications; whether it be between two board segments or for connecting various daughter-cards or separate I/O boards together. Similar to 25GBASE-CR and 25GBASE-CR-S, 25GBASE-KR and 25GBASE-KR-S differ in their support of RS-FEC. Figure 3 demonstrates these differences.

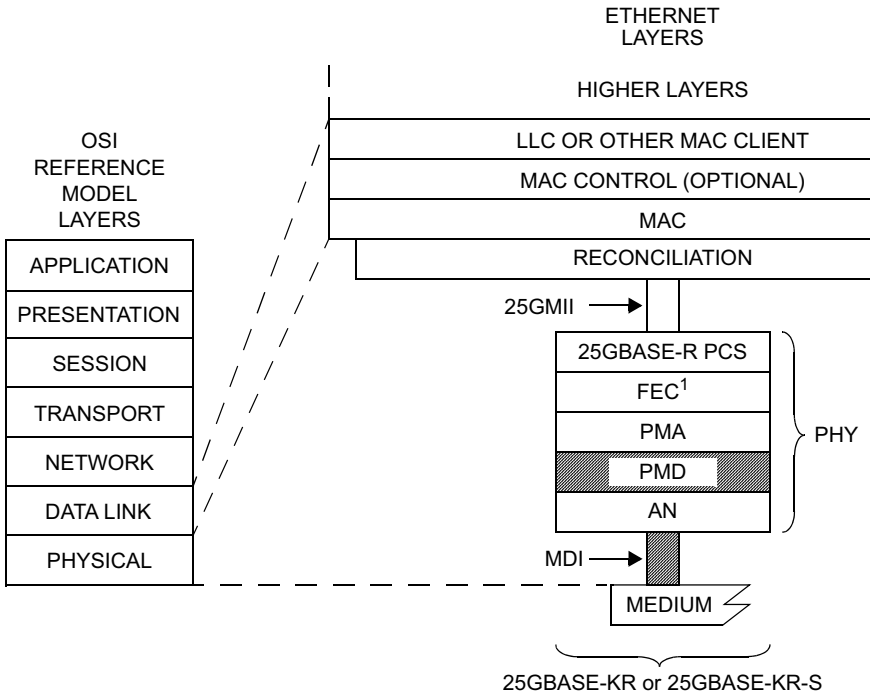
Note: For KR and KR-S PHY link modes, Auto-Negotiation is also required. Notice that AN is explicitly defined in Figure 4 for 25GBASE-KR and 25GBASE-KR-S PHYs.

Table 111–1—Physical Layer clauses associated with the 25GBASE-KR and 25GBASE-KR-S PMDs

Associated clause	25GBASE-KR	25GBASE-KR-S
106—RS	Required	Required
106—25GMII[a]	Optional	Optional
107—PCS	Required	Required
74—BASE-R FEC[b]	Required	Required
108—RS-FEC[b]	Required	N/A
109—PMA	Required	Required
109A—25GAUI C2C	Optional	Optional
73—Auto-Negotiation	Required	Required
78—Energy Efficient Ethernet	Optional	Optional

[a]The 25GMII is an optional interface. However, if the 25GMII is not implemented, a conforming implementation must behave functionally as though the RS and 25GMII were present.
[b]FEC sublayers can be enabled or disabled according to the FEC mode (see 111.6).

Figure 3. Table 111-1 (IEEE 802.3-2018, Section Seven, Page 656)



25GMII = 25 GIGABIT MEDIA INDEPENDENT INTERFACE
AN = AUTO-NEGOTATION
FEC = FORWARD ERROR CORRECTION
LLC = LOGICAL LINK CONTROL
MAC = MEDIA ACCESS CONTROL
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT

NOTE 1—CONDITIONAL BASED ON PHY TYPE

Figure 111–1—25GBASE-KR and 25GBASE-KR-S relationship to the ISO/IEC Open Systems Interconnection (OSI) reference model and the IEEE 802.3 Ethernet model

Figure 4. Figure 111-1 (IEEE 802.3-2018, Section Seven, Page 657)

2.2.3 25G-AUI-C2M/25GBASE-SR/25GBASE-LR

25G-AUI-C2M (chip-to-module) is the PHY mode defined to support optical media and be compatible with 25G-SR and 25G-LR discreet optical transceivers. This PHY link mode is specific to the channel between the PHY and module only. IEEE 802.3 defines 25GBASE-SR and 25GBASE-LR as the PHY modes used between the two ends of the optical transceiver.

For optical media, the channel seen by the link partner's PHY is only the trace from the pin of the device to the Small Form-Factor Pluggable (SFP) cage or Quad Small Form-Factor Pluggable (QSFP) on the board. This is because an active circuit inside the module regenerates the signal transmitted by each link partner before it goes on the wire.

For 25G-SR and 25G-LR optical modules, the regenerated electrical signal undergoes electrical-to-optical conversion prior to transmission over fiber. The signal regeneration reduces channel impairment, and hence the signal quality is much less dependent on the transmitter drive strength of the PHYs in each device. However, the need for FEC is stipulated by the BER limits of the optical modules in order to achieve the necessary signal quality for a healthy link. This is expressed in Table 3.

Table 3. 25G-AUI-C2M Map with Required FEC Based on BER

PHY Type	Module Type	FEC Mode	EEPROM Code[1]
25G-AUI-C2M (25GBASE-SR)	25G-SR with worst BER of 5×10^{-5}	RS-FEC required	02H
25GAUI C2M (25GBASE-SR)	25G-SR with worst BER of 10^{-12} or below	No FEC	02H
25GAUI C2M (25GBASE-LR)	25G-LR with worst BER of 5×10^{-5}	RS-FEC required	03H
25G-AUI-C2M (25GBASE-LR)	25G-LR with worst BER of 10^{-12} or below	No FEC	03H
25G-AUI-C2M	25G AOC with worst BER of 5×10^{-5}	RS-FEC required	01H
25G-AUI-C2M	25G AOC with worst BER of 10^{-12} or below	No FEC	18H

1. Compliance code was read through SFF-8472 A0 page, Byte 36.

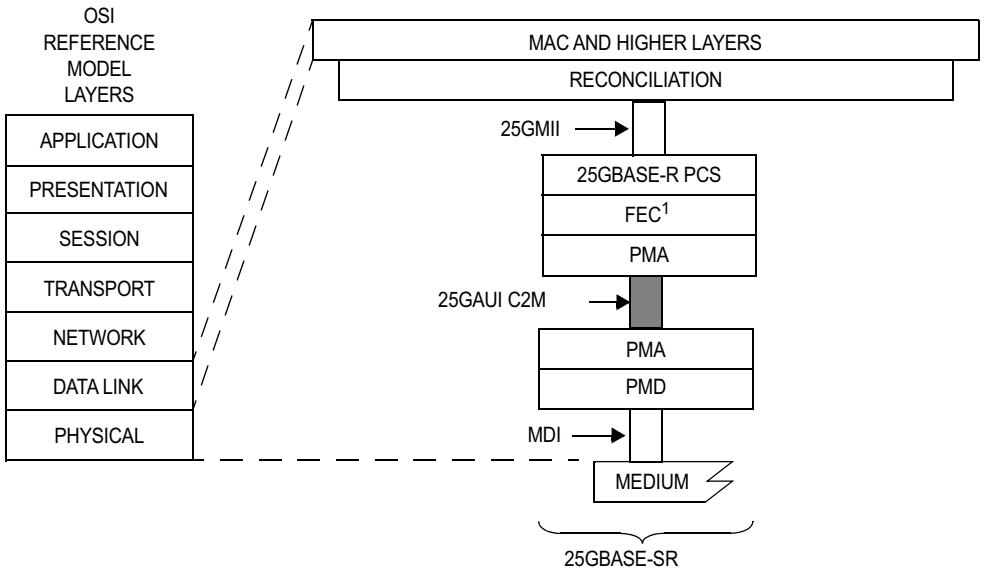
Auto-Negotiation is not possible via optical or active media and is designed to work only on passive channels. The IEEE standard does not support advertising capabilities for optical or active configurations, and in fact the optical channels cannot support the Auto-Negotiation process. This is due to the bandwidth of the active circuitry, which is not designed to pass the low frequency DME signaling of Auto-Negotiation, as well as the Link Training equalization tuning process, which occurs directly after Auto-Negotiation and is not supported through active connections.

As can be seen in the PMD definitions in Section 7 of the IEEE 802.3-2018 specification for 25G-AUI-C2M, 25GBASE-SR, and 25GBASE-LR PHYs, AN is explicitly not defined in their OSI reference diagrams.

Chip-to-module 25 Gigabit Attachment Unit Interface (25GAUI C2M)

109B.1 Overview

This annex defines the functional and electrical characteristics for the optional chip-to-module 25 Gigabit Attachment Unit Interface (25GAUI C2M). Figure 109B–1 shows the relationship of the 25GAUI C2M interface to the ISO/IEC Open System Interconnection (OSI) reference model. The 25GAUI C2M interface provides electrical characteristics and associated compliance points that can optionally be used when designing systems with pluggable module interfaces.

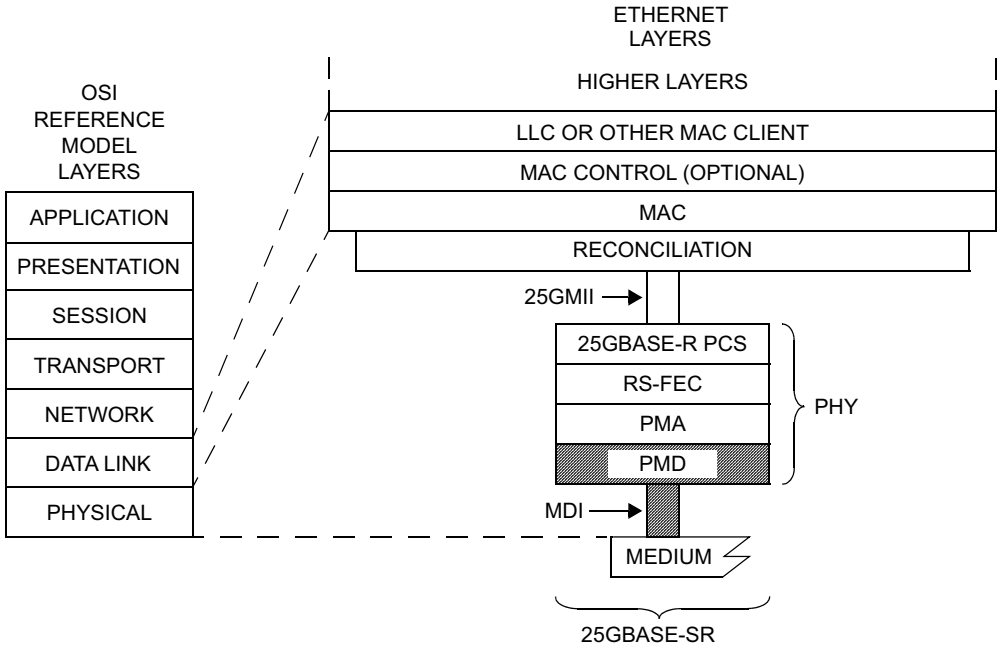


25GAUI = 25 GIGABIT ATTACHMENT UNIT INTERFACE
25GMII = 25 GIGABIT MEDIA INDEPENDENT INTERFACE
FEC = FORWARD ERROR CORRECTION
MAC = MEDIA ACCESS CONTROL
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT

NOTE 1—CONDITIONAL BASED ON PHY TYPE

Figure 109B–1—Example 25GAUI C2M relationship to the ISO/IEC Open System Interconnection reference model and the IEEE 802.3 Ethernet model

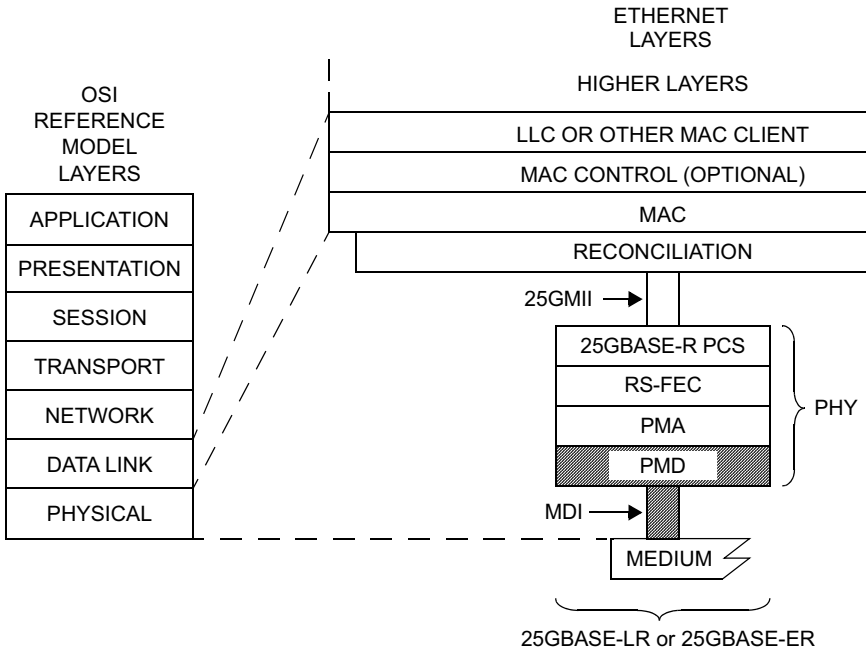
Figure 5. Figure 109B-1 (IEEE 802.3-2018, Section Seven, Annex 109B, Page 961)



25GMII = 25 GIGABIT MEDIA INDEPENDENT INTERFACE
LLC = LOGICAL LINK CONTROL
MAC = MEDIA ACCESS CONTROL
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
RS-FEC = REED-SOLOMON FORWARD ERROR CORRECTION
SR = PMD FOR MULTIMODE FIBER

Figure 112–1—25GBASE-SR PMD relationship to the ISO/IEC Open Systems Interconnection (OSI) reference model and the IEEE 802.3 Ethernet model

Figure 6. Figure 112-1 (IEEE 802.3-2018, Section Seven, Page 676)



25GMII = 25 GIGABIT MEDIA INDEPENDENT INTERFACE
LLC = LOGICAL LINK CONTROL
MAC = MEDIA ACCESS CONTROL
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
RS-FEC = REED-SOLOMON FORWARD ERROR CORRECTION
LR = PMD FOR SINGLE-MODE FIBER - 10 km
ER = PMD FOR SINGLE-MODE FIBER - 40 km

Figure 114–1—25GBASE-LR and 25GBASE-ER PMDs relationship to the ISO/IEC Open Systems Interconnection (OSI) reference model and the IEEE 802.3 Ethernet model

Figure 7. Figure 114-1 (IEEE 802.3-2018, Section Seven, Page 817)

Section 112.1 also implicitly defines the lack of AN support for 25GBASE-SR PMDs.

Table 112–1—Physical Layer clauses associated with the 25GBASE-SR PMD

Associated clause	25GBASE-SR
106—RS	Required
106—25GMII[a]	Optional
107—PCS for 25GBASE-R	Required
108—RS-FEC[b]	Required
109—PMA for 25GBASE-R	Required
109A—25GAUI C2C	Optional
109B—25GAUI C2M	Optional
78—Energy Efficient Ethernet	Optional

[a]The 25GMII is an optional interface. However, if the 25GMII is not implemented, a conforming implementation must behave functionally as though the RS and 25GMII were present.
[b]The option to bypass the Clause 108 RS-FEC correction function is not supported.

Figure 8. Table 112-1 (IEEE 802.3-2018, Section Seven, Page 817)

2.3 25G Interoperability Debug Flows

Intel recommends the following 25G interoperability debug flows:

- Section 2.3.1, "Passive Direct Attach Copper (DAC) Cables"
- Section 2.3.2, "25GBASE-SR/25GBASE-LR Optics and Active Optical Cables (AOC)"

2.3.1 Passive Direct Attach Copper (DAC) Cables

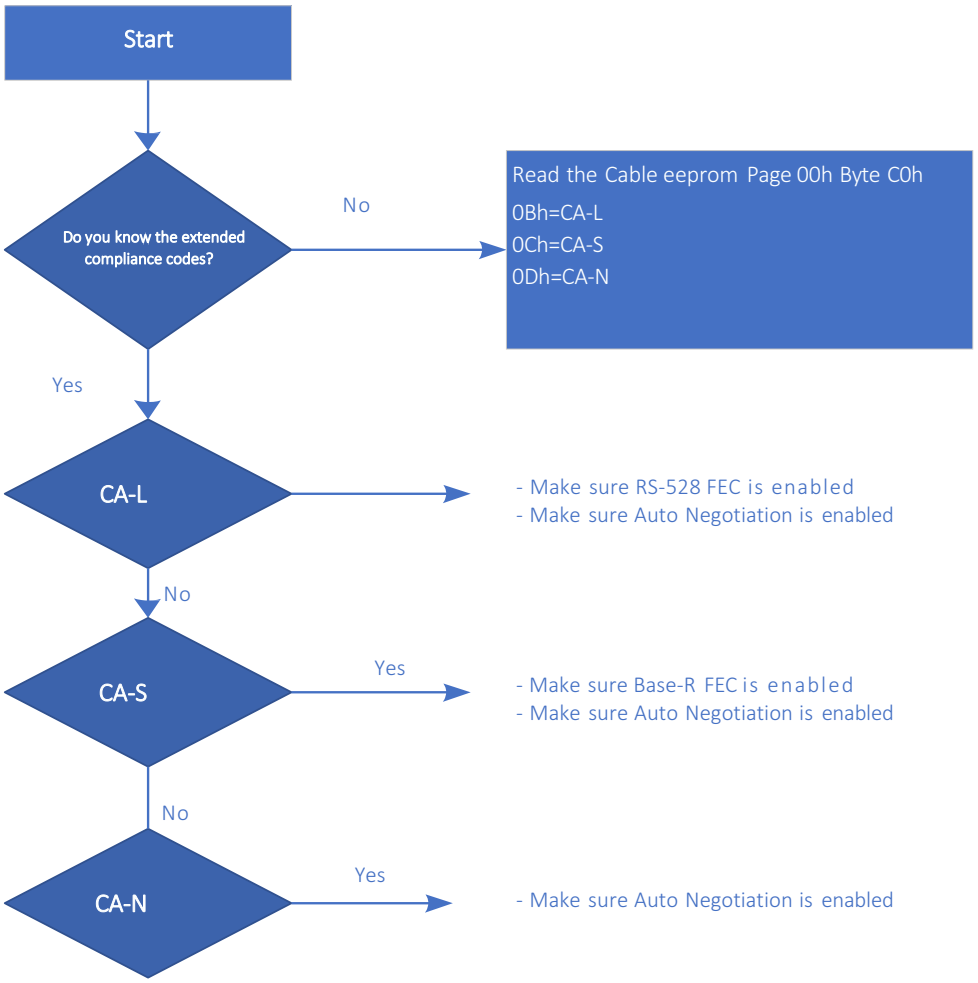


Figure 9. Passive DAC Cables

2.3.2 25GBASE-SR/25GBASE-LR Optics and Active Optical Cables (AOC)

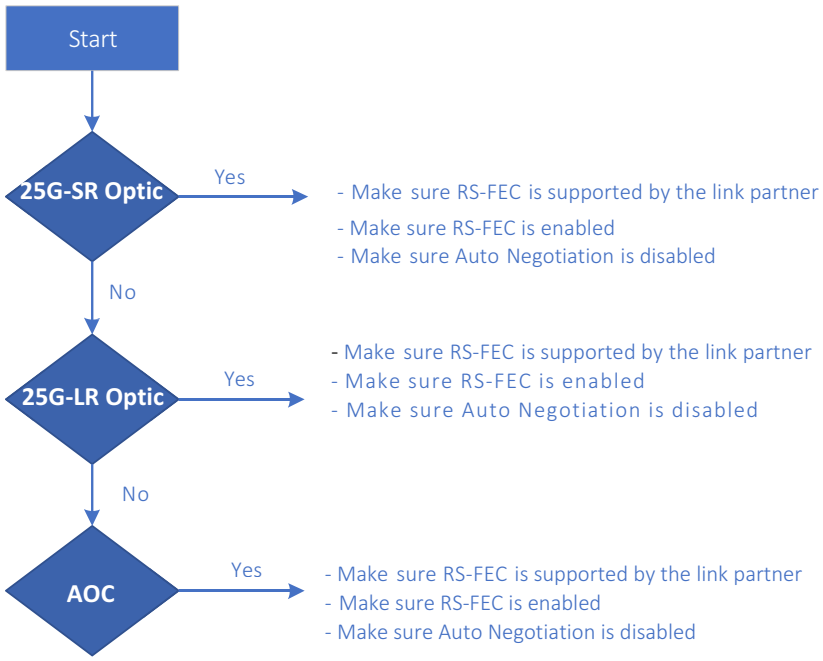


Figure 10. 25GBASE-SR/LR Optics and AOCs

Note: When debugging a link, ensure the media has the correct FEC mode enabled appropriate for the BER. For instance, if your media does not require RS-FEC due to its own FEC implementation or if the BER is less that 10^{-12}, some of the steps above can safely be skipped.

LEGAL