1G/10G/25G/40G/100G Optical Transceiver Troubleshooting Manual

I、Document Background

As the digital transformation of enterprises accelerates, the demand for network bandwidth continues to grow steadily, driven by cloud computing, video streaming, 5G networks, and medium-sized data centers. The standard rate optical transceiver, with its mature transmission performance—based on NRZ or PAM4 modulation and supporting per-channel data rates of 25G/50Gbps—has become a key component in building modern network architectures. It is widely deployed in intra-data center interconnects, enterprise core networks, and edge computing nodes.

Despite the relative maturity of 100G optical module technologies, fault diagnostics must address the following core challenges:

1. Physical Layer Stability:

NRZ modulation has a lower tolerance for signal noise, while PAM4 introduces timing jitter and amplitude noise, which can lead to increased bit error rates (BER). Therefore, strict control over optical power budget, fiber attenuation, and connector cleanliness is essential.

2. Module Integration and Reliability:

Most modules incorporate TOSA/ROSA components, conventional DML lasers, and basic DSP chips. Although packaging density is moderate, prolonged operation can lead to performance degradation due to uneven heat dissipation or PCB aging.

3. Interoperability Across Vendors:

Mainstream form factors such as SFP+、SFP28、QSFP、QSFP28 and CFP2 must maintain compatibility across different switch vendors. Compliance with standards such as IEEE 802.3bm and 100G-LR4/ER4 is required. Failed protocol negotiation (e.g., Auto-Negotiation) can lead to link anomalies.

4. Maintenance Tool Compatibility:

Potential failures include optical link attenuation, signal distortion, or inadequate FEC correction. Multi-layer diagnostics require tools like optical power meters, BERTs, and eye diagram analyzers.

This guide provides FS technical engineers with a standardized troubleshooting procedure for standard rate optical modules, covering common failure scenarios (e.g., port not coming UP, intermittent packet loss, module overtemperature alarms, etc.). By integrating hands-on case studies and diagnostic flowcharts, it enhances network operation efficiency and ensures transmission reliability in high-load scenarios.

II、Switch System Overview

Switches on the market are generally categorized into white-box switches and black-box switches. White-box switches are open network devices with decoupled hardware and software, offering greater flexibility. Black-box switches, come with vendor-specific integrated software, where the underlying system logic varies between manufacturers. As a result, features, functionalities, and troubleshooting methods also differ significantly. Unlike black-box systems, white-box switches typically run open-source network operating systems, such as Picos, Cumulus, and SONiC.

For users deploying the same open-source system, the command-line interface and troubleshooting processes remain largely consistent, regardless of the hardware vendor.

As a global leader in accelerated computing and data center interconnect technologies, **NVIDIA** continues to push the boundaries of AI training, inference, and real-time data analytics through its end-to-end ecosystem built with GPU clusters, InfiniBand networks, and the Spectrum series of Ethernet switches. Against this backdrop, systems based on the NVIDIA architecture—namely **Cumulus** and **MLNX-OS**—have been widely adopted and have received positive feedback from the market regarding their user experience.

This guide focuses on troubleshooting issues in these two types of systems: the first involves analyzing problems that occur when using devices running the open-source **Cumulus** system; the second involves analyzing issues encountered when using devices installed with the **NVIDIA MLNX-OS** system.

III、Troubleshooting Guide to Cumulus Systems

1. Introduction to Physical Layer Interface Parameters

This chapter introduces the key parameters of optical modules at the physical layer (Layer 1). By understanding these core parameters, readers can better identify potential physical layer issues that may affect the ability of switch ports to connect to the network.

1.1 Specification Standards

Before familiarizing yourself with troubleshooting physical layer interface errors, it is important to understand the relevant specification protocols.

- General Management Interface Specifications
- SFF Module Management Reference Supplementary Protocol
- Institute of Electrical and Electronics Engineers (IEEE) International Standards Organization Protocol: IEEE Standards Association

1.2 Form Factors

- **SFP**: Mainly used for 1Gbps (Gigabit Ethernet, 1G Fibre Channel); some modules support 2.5Gbps or 4Gbps (Fibre Channel)
- **SFP+**: Mainly used for 10Gbps (10G Ethernet, 8G/10G/16G Fibre Channel). Physically compatible with SFP, but supports higher electrical rates.
- **SFP28**: Uses NRZ (Non-Return-to-Zero) modulation, evolved from 10G NRZ (SFP+), designed for 25Gbps per lane.
- **QSFP28:** Four-Channel Small Form-Factor Pluggable Module (Single-Channel Rate 25G/50Gbps, Supporting NRZ/PAM4 Modulation)
- **CFP2:** Second-Generation 100-Gigabit Pluggable Module (Supports 4×25G or 10×10G link aggregation, Commonly used in telecom backbone networks)

- **Core Features of the Form Factor:**

  All form factors are equipped with embedded **EEPROM memory**, which stores the manufacturer's information, specifications (such as wavelength, transmission distance, power consumption, etc.), and real-time status data (including temperature, voltage, and optical power). In a Cumulus system, you can use the command "ethtool -m swp" to parse the EEPROM content and retrieve detailed configuration and operational metrics of the module.

1.3 EEPROM Information

As a component operating at the first layer of the physical layer, an optical transceivers must have EEPROM information that complies with relevant specifications in order to function properly. The content and structure of this information are defined under the corresponding version of the CMIS. The following section describes the key information bytes in detail.

1.4 Analysis of Form Factor Information

In the module's EEPROM, the first byte (Identifier value) defines the form factor type, following the **SFF-8636 Table 4-1** standard. The following is an interpretation of the identifier values related to 100G optical module form factors:

Identifier Value	Form Factor Types	Technical Specifications and Typical Applications
0x03	**SFP**	Small form-factor pluggable (SFP) module supporting 1G/2.5G data rates (NRZ modulation), using the SFF-8472 management interface. Commonly used in Gigabit Ethernet and Fibre Channel applications, with interface types such as 1000BASE-SX/LX.
0x0D	**SFP+**	Enhanced SFP form factor (SFP+), supporting 10G data rates (NRZ modulation), compliant with the SFF-8431 management protocol. Widely used in access-layer applications of data centers, such as 10G-SR/LR and 10GBase-T.
0x0D	**SFP+**	Same physical size as SFP+, supports 25G data rates (NRZ modulation), with a management interface compatible with SFF-8431. Mainly used for 25G server access and 5G fronthaul networks (25G eCPRI).
0x0D	**QSFP28**	Four-channel small form-factor pluggable module, supporting 4×25G NRZ or 2×50G PAM4 modulation, uses the SFF-8636 management interface, and is suitable for short-distance data center interconnects such as 100G-SR4/LR4.

0x12	**CFP2**	Second-generation 100-Gigabit pluggable module, supporting 4×25G or 10×10G link aggregation, compatible with long-distance telecom transmission scenarios such as 100G-ER4/ZR4.
0x13	**CFP4**	Compact CFP form factor (half the size of CFP2), supports 4×25G or 2×50G configurations, suitable for high-density 100G-LR4/ER4 deployments.

1.5 Module Coding Information

The encoding values are primarily used for the host interface, indicating the encoding methods supported by the host device. Common encoding methods include 8B/10B, 64B/66B, 254B/256B, NRZ, and PAM4. This section mainly introduces NRZ and PAM4. The module's EEPROM information specifies the encoding method used by the module. If the encoding method differs from the one supported by the switch, it may lead to errors during the signal transmission process.

For common encoding methods in the optical module EEPROM, please refer to SFF-8024 Table 4-2 Encoding Values.

- **NRZ Encoding**: NRZ (also known as PAM2) uses two voltage (or laser) levels to represent "0" and "1". Negative voltage level = 0, positive voltage = 1. It is typically used in Ethernet technologies with channel speeds lower than 50Gbps.

The above shows the optical eye diagram output results of the NRZ optical module from the FS laboratory.

- **PAM4 Encoding**: PAM4 encoding uses four signal levels to represent two bits (00, 01, 10, 11). These levels are closer together compared to the levels in NRZ/PAM2, making signal integrity a more significant concern. PAM4 encoding is used for products with channel speeds of ≥ 50Gbps.

The above shows the optical eye diagram output results of the PAM4 optical module from the FS laboratory.

1.6 Meida Type, Interface Type, Vendor OUI, VendorSN, etc.

To allow the port to recognize the characteristics of the inserted module, the EEPROM of a 100G module contains a set of standardized data

to describe the module's features. These values can be seen in the output of the command "ethtool -m swp".

The **Meida Type** describes the type of Ethernet technology implemented by the module. The 100G Type can be referenced as described below:

Table 8-16 Media Type Encodings

Code	Media Type	Associated Interface ID Table
00h	Undefined	None, not applicable
01h	Optical Interfaces: MMF	[5] Table "850 nm MM media interface codes"
02h	Optical Interfaces: SMF	[5] Table "SM media interface codes"
03h	Passive Copper Cables	[5] Table "Passive Copper Cable interface codes"
04h	Active Cables	[5] Table "Active Cable assembly interface codes"
05h	BASE-T	[5] Table "BASE-T media interface codes"
06h-3Fh	-	**Reserved**
40h-8Fh	-	**Custom**
90h-FFh	-	**Reserved**

1.7 Digital Diagnostic Monitoring Function (DDM/DOM)

DDM/DOM is a crucial parameter for monitoring and troubleshooting optical modules. It provides feedback on the current temperature, voltage, current, transmit and receive power of the port module. Each of these parameters is extremely important as they reflect the current operational status of the module. These parameter values can be seen in the output of the command "ethtool -m swp".

1.7.1 Temperature

- **Module operating temperature:**

Optical modules are temperature-sensitive devices and must operate within a reasonable temperature range to avoid performance degradation or device damage. Temperature grades are divided into two categories:

- **Commercial grade temperature**: **0℃ ~ 70℃** (typical data center/enterprise network environments).
- **Industrial grade temperature**: **-45℃ ~ 80℃** (extreme environments, such as outdoor base stations or industrial settings).

Due to higher power consumption, 100G modules generate significant heat during operation. Optical modules are temperature-sensitive devices, and when they overheat, the performance metrics of the module degrade significantly, which can affect the quality of signal transmission.

1.7.2 Operating voltage

- **Voltage range:**

The module power supply needs to meet the range of **3.135V ~ 3.465V** (default **3.3V**). Insufficient or fluctuating voltage may result in:

- **Startup timing issues**: Module initialization fails, preventing the module from entering the operational state.
- **Logic errors**: The MCU (Microcontroller) or DSP (Digital Signal Processor) malfunctions, causing DDM/DOM functionality to fail.
- **Performance degradation**: Unstable output from the laser driver circuit, leading to fluctuations in optical power.

1.7.3 Bias current

- **Definition**：The current required to drive the Transmitter Optical Sub-Assembly (TOSA) reflects the operating state of the laser (not the overall power consumption of the module) .
- **Typical scenarios:**
  - **Normal range**: Varies depending on the module type (e.g.: 100GBASE-SR4: 30~50mA)
  - **Abnormal judgment**:
    0mA: The laser is not powered on, or there is hardware damage (such as a TOSA failure).
    Above threshold: If the current reaches **1.5 times** the laser's threshold current (e.g., 60mA when the threshold is 40mA), it indicates the end of the laser's lifespan (due to aging or degradation).

1.7.4 Transmit (TX) and Receive (RX) optical power

- **Parameter meaning**:
  - **TX power**: The optical power output from the module's transmit (TX) end, which must remain within the device's allowable range (e.g.: **100GBASE-SR4:** -6.5 ~ -1.0 dBm)
  - **RX power**: The optical power received by the module's receive (RX) end, which should be higher than the sensitivity threshold and lower than the overload point (e.g.: **100GBASE-SR4:** -11.1 ~ +1.5 dBm)
- **Alarm mechanism:**

When the TX/RX power exceeds the threshold, it triggers the **High/Low Warning/Alarm** flags, for example:

```
ethtool -m swp1...
Laser output power : 1.2 mW / 0.8 dBm
Receiver signal average power: 0.6 mW / -2.2 dBm
Transmit avg power high alarm : Off
Transmit avg power low alarm : Off
Receive avg power high alarm : Off
Receive avg power low alarm : Off
```

1.7.5 DDM/DOM fieldstandardization

- **Function description:**

Digital Diagnostic Monitoring (DDM/DOM) stores real-time operating parameters of the optical module (such as temperature, voltage, optical power, etc.) in the EEPROM.

- **View command:**
  - **Cumulus Linux**：nv show platform transceiver interface
  - **General Linux system**: ethtool -m swp
- **Alarm flags:**

  When parameters exceed the limit, the corresponding alarm flags (e.g., High Alarm/Low Warning) switch from "Off" to "On". Troubleshooting the link issues requires referring to the threshold table.

Appendix: Typical Parameter Threshold Table

Module Type	TX Power Range（dBm）	RX Sensitivity（dBm）	Maximum Bias Current（mA）
1000BASE-SX	-9.5~-4.0	≤-17.0	45
1000BASE-LX	-11.0~-3.0	≤-19.0	60
10GBASE-SR	-7.3~-1.0	≤-11.1	65
10GBASE-LR	-8.2~0.5	≤-12.6	75
25GBASE-SR	-8.4~2.4	≤-10.3	70
25GBASE-LR	-8.4~4.5	≤-11.2	90
40GBASE-SR4	-7.3~-1.0	≤-11.1	150
40GBASE-LR4	-8.2~1.5	≤-14.0	180
100GBASE-SR4	-8.4~2.4	≤-10.3	50
100GBASE-LR4	-8.2 ~ +4.5	≤-14.0	100
100GBASE-ER4	-5.0 ~ +5.0	≤-21.0	120

1.8 Configuration Negotiation on Cumulus

1.8.1 FEC Settings and Negotiation

FEC (Forward Error Correction) is an error control method where signals are encoded with a specific algorithm before being transmitted through the communication channel. Redundant bits that carry the characteristics of the signal itself are added. At the receiver end, the signal is decoded using the corresponding algorithm to detect and correct any errors that occurred during transmission. For high-speed modules, the protocol specifies a bit error rate threshold of 2.4E(-4). In practice, manufacturers design their modules to operate at much lower error rates to ensure error-free data transmission during the link process. To ensure transmission quality, FEC is typically enabled.

> ⓘ Please note that when FEC is enabled, the FEC type on both ends of the devices must be consistent. In certain speed settings, due to issues related to encoding changes over time, there may be FEC compatibility issues between different versions of the same product.

- Reed-Solomon RS-FEC(544,514) algorithm adds 30 bits of overhead to correct 14+ bit errors in every 514 bits. The 50G (PAM4) channel requires FEC RS; this applies to all 800G, 400G, 200G, 100G-CR2, and 50G-CR interfaces.
- Reed-Solomon RS-FEC(528,514) algorithm adds 14 bits of encoding information to a 514-bit stream. It replaces and uses the same overhead as the 64B/66B encoding, so the bit rate is unaffected. This algorithm can correct 7-bit errors in a 514-bit stream. RS(528,514) is used for 25G (NRZ) channels, including 25G, 50G-CR2, and 100G-SR4/CR4 interfaces.
- Base-R (also known as FireCode/FC) FEC adds 32 bits for every 32 64B/66B blocks to correct 11 bits out of every 2048 bits. It replaces one bit per block, so its overhead is the same as the 64B/66B encoding. It is used exclusively for 25G interfaces. The execution speed of this algorithm is faster than that of the RS-FEC algorithm, resulting in reduced latency. Both RS-FEC and Base-R FEC are implemented in hardware.
- None/Off: FEC is optional and typically useful for 25G channels, including 100G-SR4/CR4 and 50G-CR2 links. If the cable quality is sufficient to achieve a BER of 10^-12 without using FEC, there is no reason to enable it. 10G/40G links never require FEC. If a 10G/40G link experiences errors, replace the faulty cable or module.
- Auto: FEC can be automatically negotiated between two devices. When auto-negotiation is enabled, the default FEC setting is auto, allowing the neighbor to send and receive FEC feature information. The port's FEC active/operational setting is set to the negotiated result. Auto is the default setting on NVIDIA switches (with auto-negotiation enabled by default).
- If auto-negotiation is disabled on 100G and 25G interfaces, FEC must be set to OFF *, RS, or BaseR to match the neighbor. When auto-negotiation is disabled, the FEC default setting of *auto* will not link.

1.9 CLI Display Use Case in Cumulus System

Take the QSFP28-100G-LR4 as an example: please see the attached file for details.

```
root@SN3700:mgmt:~# ethtool -m swp1
 Identifier : 0x11 (QSFP28)
 Extended identifier : 0xce
 Extended identifier description : 3.5W max. Power consumption
 Extended identifier description : CDR present in TX, CDR present in RX
 Extended identifier description : 4.5W max. Power consumption, High Power Class (> 3.5 W) enabled
 Power set : Off
 Power override : On
 Connector : 0x07 (LC)
 Transceiver codes : 0x80 0x00 0x00 0x00 0x12 0x00 0x01 0xe0
 Transceiver type : 100G Ethernet: 100G Base-LR4
 Transceiver type : FC: long distance (L)
 Transceiver type : FC: Longwave laser (LC)
 Transceiver type : FC: Single Mode (SM)
 Transceiver type : FC: 1200 MBytes/sec
 Transceiver type : FC: 800 MBytes/sec
 Transceiver type : FC: 1600 MBytes/sec
 Encoding : 0x05 (64B/66B)
 BR, Nominal : 25500Mbps
 Rate identifier : 0x02
 Length (SMF,km) : 25km
 Length (OM3 50um) : 0m
 Length (OM2 50um) : 0m
 Length (OM1 62.5um) : 0m
 Length (Copper or Active cable) : 0m
 Transmitter technology : 0x40 (1310 nm DFB)
```

Laser wavelength : 1310.000nm
Laser wavelength tolerance : 2.245nm
Vendor name : FS
Vendor OUI : 00:00:00
Vendor PN : QSFP28-LR4-100G
Vendor rev : 01
Vendor SN : C2003030166
Date code : 20041619
Revision Compliance : SFF-8636 Rev 2.5/2.6/2.7
Module temperature : 35.32 degrees C / 95.58 degrees F
Module voltage : 3.3469 V
Alarm/warning flags implemented : Yes
Laser tx bias current (Channel 1) : 39.200 mA
Laser tx bias current (Channel 2) : 39.200 mA
Laser tx bias current (Channel 3) : 39.200 mA
Laser tx bias current (Channel 4) : 39.200 mA
Transmit avg optical power (Channel 1) : 1.4283 mW / 1.55 dBm
Transmit avg optical power (Channel 2) : 1.9447 mW / 2.89 dBm
Transmit avg optical power (Channel 3) : 2.1287 mW / 3.28 dBm
Transmit avg optical power (Channel 4) : 2.1196 mW / 3.26 dBm
Rcvr signal avg optical power(Channel 1) : 0.0001 mW / -40.00 dBm
Rcvr signal avg optical power(Channel 2) : 0.0001 mW / -40.00 dBm
Rcvr signal avg optical power(Channel 3) : 0.0001 mW / -40.00 dBm
Rcvr signal avg optical power(Channel 4) : 0.0001 mW / -40.00 dBm
Laser bias current high alarm (Chan 1) : Off
Laser bias current low alarm (Chan 1) : Off
Laser bias current high warning (Chan 1) : Off
Laser bias current low warning (Chan 1) : Off
Laser bias current high alarm (Chan 2) : Off
Laser bias current low alarm (Chan 2) : Off
Laser bias current high warning (Chan 2) : Off
Laser bias current low warning (Chan 2) : Off
Laser bias current high alarm (Chan 3) : Off
Laser bias current low alarm (Chan 3) : Off
Laser bias current high warning (Chan 3) : Off
Laser bias current low warning (Chan 3) : Off
Laser bias current high alarm (Chan 4) : Off
Laser bias current low alarm (Chan 4) : Off
Laser bias current high warning (Chan 4) : Off
Laser bias current low warning (Chan 4) : Off
Module temperature high alarm : Off
Module temperature low alarm : Off
Module temperature high warning : Off
Module temperature low warning : Off
Module voltage high alarm : Off
Module voltage low alarm : Off
Module voltage high warning : Off
Module voltage low warning : Off
Laser tx power high alarm (Channel 1) : Off
Laser tx power low alarm (Channel 1) : Off
Laser tx power high warning (Channel 1) : Off
Laser tx power low warning (Channel 1) : Off
Laser tx power high alarm (Channel 2) : Off
Laser tx power low alarm (Channel 2) : Off
Laser tx power high warning (Channel 2) : Off
Laser tx power low warning (Channel 2) : Off

```
Laser tx power high alarm (Channel 3) : Off
Laser tx power low alarm (Channel 3) : Off
Laser tx power high warning (Channel 3) : Off
Laser tx power low warning (Channel 3) : Off
Laser tx power high alarm (Channel 4) : Off
Laser tx power low alarm (Channel 4) : Off
Laser tx power high warning (Channel 4) : Off
Laser tx power low warning (Channel 4) : Off
Laser rx power high alarm (Channel 1) : Off
Laser rx power low alarm (Channel 1) : On
Laser rx power high warning (Channel 1) : Off
Laser rx power low warning (Channel 1) : On
Laser rx power high alarm (Channel 2) : Off
Laser rx power low alarm (Channel 2) : On
Laser rx power high warning (Channel 2) : Off
Laser rx power low warning (Channel 2) : On
Laser rx power high alarm (Channel 3) : Off
Laser rx power low alarm (Channel 3) : On
Laser rx power high warning (Channel 3) : Off
Laser rx power low warning (Channel 3) : On
Laser rx power high alarm (Channel 4) : Off
Laser rx power low alarm (Channel 4) : On
Laser rx power high warning (Channel 4) : Off
Laser rx power low warning (Channel 4) : On
Laser bias current high alarm threshold : 42.500 mA
Laser bias current low alarm threshold : 20.000 mA
Laser bias current high warning threshold : 41.500 mA
Laser bias current low warning threshold : 25.000 mA
 Laser output power high alarm threshold : 3.5481 mW / 5.50 dBm
 Laser output power low alarm threshold : 0.1862 mW / -7.30 dBm
 Laser output power high warning threshold : 2.8183 mW / 4.50 dBm
 Laser output power low warning threshold : 0.3715 mW / -4.30 dBm
 Module temperature high alarm threshold : 80.00 degrees C / 176.00 degrees F
 Module temperature low alarm threshold : -5.00 degrees C / 23.00 degrees F
 Module temperature high warning threshold : 75.00 degrees C / 167.00 degrees F
 Module temperature low warning threshold : 0.00 degrees C / 32.00 degrees F
 Module voltage high alarm threshold : 3.6000 V
 Module voltage low alarm threshold : 3.0500 V
 Module voltage high warning threshold : 3.4650 V
 Module voltage low warning threshold : 3.1350 V
 Laser rx power high alarm threshold : 3.5481 mW / 5.50 dBm
 Laser rx power low alarm threshold : 0.0430 mW / -13.67 dBm
 Laser rx power high warning threshold : 2.8183 mW / 4.50 dBm
 Laser rx power low warning threshold : 0.0870 mW / -10.60 dBm
root@SN3700:mgmt:~# ethtool -m swp7
Identifier : 0x11 (QSFP28)
Extended identifier : 0xce
Extended identifier description : 3.5W max. Power consumption
Extended identifier description : CDR present in TX, CDR present in RX
Extended identifier description : 4.5W max. Power consumption, High Power Class (> 3.5 W) enabled
Power set : Off
Power override : On
Connector : 0x07 (LC)
Transceiver codes : 0x80 0x00 0x00 0x00 0x12 0x00 0x01 0xe0
Transceiver type : 100G Ethernet: 100G Base-LR4
Transceiver type : FC: long distance (L)
```

Transceiver type : FC: Longwave laser (LC)
Transceiver type : FC: Single Mode (SM)
Transceiver type : FC: 1200 MBytes/sec
Transceiver type : FC: 800 MBytes/sec
Transceiver type : FC: 1600 MBytes/sec
Encoding : 0x05 (64B/66B)
BR, Nominal : 25500Mbps
Rate identifier : 0x02
Length (SMF,km) : 30km
Length (OM3 50um) : 0m
Length (OM2 50um) : 0m
Length (OM1 62.5um) : 0m
Length (Copper or Active cable) : 0m
Transmitter technology : 0x40 (1310 nm DFB)
Laser wavelength : 1310.000nm
Laser wavelength tolerance : 2.245nm
Vendor name : FS
Vendor OUI : 00:00:00
Vendor PN : QSFP28-LR4-100G
Vendor rev : 01
Vendor SN : C2003030166
Date code : 20041619
Revision Compliance : SFF-8636 Rev 2.5/2.6/2.7
Module temperature : 28.75 degrees C / 83.74 degrees F
Module voltage : 3.3557 V
Alarm/warning flags implemented : Yes
Laser tx bias current (Channel 1) : 39.424 mA
Laser tx bias current (Channel 2) : 40.320 mA
Laser tx bias current (Channel 3) : 38.752 mA
Laser tx bias current (Channel 4) : 39.200 mA
Transmit avg optical power (Channel 1) : 2.0331 mW / 3.08 dBm
Transmit avg optical power (Channel 2) : 2.8339 mW / 4.52 dBm
Transmit avg optical power (Channel 3) : 2.4186 mW / 3.84 dBm
Transmit avg optical power (Channel 4) : 2.4196 mW / 3.84 dBm
Rcvr signal avg optical power(Channel 1) : 0.0001 mW / -40.00 dBm
Rcvr signal avg optical power(Channel 2) : 0.0001 mW / -40.00 dBm
Rcvr signal avg optical power(Channel 3) : 0.0001 mW / -40.00 dBm
Rcvr signal avg optical power(Channel 4) : 0.0001 mW / -40.00 dBm
Laser bias current high alarm (Chan 1) : Off
Laser bias current low alarm (Chan 1) : Off
Laser bias current high warning (Chan 1) : Off
Laser bias current low warning (Chan 1) : Off
Laser bias current high alarm (Chan 2) : Off
Laser bias current low alarm (Chan 2) : Off
Laser bias current high warning (Chan 2) : Off
Laser bias current low warning (Chan 2) : Off
Laser bias current high alarm (Chan 3) : Off
Laser bias current low alarm (Chan 3) : Off
Laser bias current high warning (Chan 3) : Off
Laser bias current low warning (Chan 3) : Off
Laser bias current high alarm (Chan 4) : Off
Laser bias current low alarm (Chan 4) : Off
Laser bias current high warning (Chan 4) : Off
Laser bias current low warning (Chan 4) : Off
Module temperature high alarm : Off
Module temperature low alarm : Off

Module temperature high warning : Off
Module temperature low warning : Off
Module voltage high alarm : Off
Module voltage low alarm : Off
Module voltage high warning : Off
Module voltage low warning : Off
Laser tx power high alarm (Channel 1) : Off
Laser tx power low alarm (Channel 1) : Off
Laser tx power high warning (Channel 1) : On
Laser tx power low warning (Channel 1) : Off
Laser tx power high alarm (Channel 2) : Off
Laser tx power low alarm (Channel 2) : Off
Laser tx power high warning (Channel 2) : On
Laser tx power low warning (Channel 2) : Off
Laser tx power high alarm (Channel 3) : Off
Laser tx power low alarm (Channel 3) : Off
Laser tx power high warning (Channel 3) : Off
Laser tx power low warning (Channel 3) : Off
Laser tx power high alarm (Channel 4) : Off
Laser tx power low alarm (Channel 4) : Off
Laser tx power high warning (Channel 4) : Off
Laser tx power low warning (Channel 4) : Off
Laser rx power high alarm (Channel 1) : Off
Laser rx power low alarm (Channel 1) : On
Laser rx power high warning (Channel 1) : Off
Laser rx power low warning (Channel 1) : On
Laser rx power high alarm (Channel 2) : Off
Laser rx power low alarm (Channel 2) : On
Laser rx power high warning (Channel 2) : Off
Laser rx power low warning (Channel 2) : On
Laser rx power high alarm (Channel 3) : Off
Laser rx power low alarm (Channel 3) : On
Laser rx power high warning (Channel 3) : Off
Laser rx power low warning (Channel 3) : On
Laser rx power high alarm (Channel 4) : Off
Laser rx power low alarm (Channel 4) : On
Laser rx power high warning (Channel 4) : Off
Laser rx power low warning (Channel 4) : On
Laser bias current high alarm threshold : 42.500 mA
Laser bias current low alarm threshold : 20.000 mA
Laser bias current high warning threshold : 41.500 mA
Laser bias current low warning threshold : 25.000 mA
Laser output power high alarm threshold : 3.5481 mW / 5.50 dBm
Laser output power low alarm threshold : 0.1862 mW / -7.30 dBm
Laser output power high warning threshold : 2.8183 mW / 4.50 dBm
Laser output power low warning threshold : 0.3715 mW / -4.30 dBm
Module temperature high alarm threshold : 80.00 degrees C / 176.00 degrees F
Module temperature low alarm threshold : -5.00 degrees C / 23.00 degrees F
Module temperature high warning threshold : 75.00 degrees C / 167.00 degrees F
Module temperature low warning threshold : 0.00 degrees C / 32.00 degrees F
Module voltage high alarm threshold : 3.6000 V
Module voltage low alarm threshold : 3.0500 V
Module voltage high warning threshold : 3.4650 V
Module voltage low warning threshold : 3.1350 V
Laser rx power high alarm threshold : 3.5481 mW / 5.50 dBm
Laser rx power low alarm threshold : 0.0430 mW / -13.67 dBm

```
Laser rx power high warning threshold : 2.8183 mW / 4.50 dBm
Laser rx power low warning threshold : 0.0870 mW / -10.60 dBm
```

2. Physical Layer Troubleshooting and Issue Resolution

2.1 Classification of Cumulus Issues

Physical layer issues on the Cumulus system are mainly classified into three categories:

- Configuration issues: Incorrect configuration of one neighbor or another, or configuration mismatch between neighbors.
- Hardware issues: Faults in fiber optic links or modules, and (in rare cases) faults in the switch port.
- Errors when switching drivers for specific module types. These errors are rare and can usually be resolved.

For example, when the device is using 100G SR4, the signaling rate at the physical layer is based on PAM4 modulation. At this point, the switch's lower-level software logic supports 100G SR4. However, if this module is removed and replaced with a 100G coherent module, the modulation method for the 100G coherent module is DPSK-QAM. The switch needs to switch its lower-level logic accordingly. Sometimes, this switch fails, resulting in errors.

2.2 General Fault Troubleshooting Approach

Before addressing specific issues, it is necessary to rule out the impact of physical hardware by using the method of controlling variables for troubleshooting at the physical hardware level. The physical hardware includes the device, fiber patch cables, modules, or cables.

Common physical layer hardware errors:

- DDM parameter information error: Check the DDM parameters. "l1-show".
- Remote end is sending RX failure: Check if the neighbor is sending remote fault isolation information.
- PCS error count: If FEC is not required and errors appear on the link, check the "HwIfInErrors" counter by using "ethtool -S swp" to see if the error count increases over time.

Operation steps:

First, output the product's specific information through the "l1-show" command, especially the DDM information. Based on this information, make an initial judgment on the problematic board and troubleshoot hardware issues. Following the principle of simplicity, prioritize troubleshooting the easiest potential issue first.

When addressing specific problems, it's recommended to first observe the end faces of the optical modules or jumpers (except for DAC/AOC cables). Check for dirt, wear, or damage. If such issues are found, clean the end faces before using them. If a visual inspection tool is not available, clean multiple times to ensure proper cleanliness.

1. For module/cable issues, replace with the same type of product for interconnection, or substitute with a known good product of the same type.

2. For jumper issues, replace the jumper to resolve the problem. If the actual line in use is a long-distance cable, use OTDR to evaluate the overall loss of the cable.

3. Test by changing the device port or replacing the device.

4. Loopback test: Perform a loopback test either on the near-end or far-end device. This can be a self-loop on a single port or a self-loop connection between different ports on the same device.

2.3 Other Faults and Solutions

Other faults can be roughly classified into the following categories:

- Link Disconnection or Jitter
- Physical Layer Errors on the Link (PCS Errors), which are different from packet loss; packet loss is a Layer 2 or Layer 3 switching issue.
- Signal Integrity Issues, manifested as errors, link interruptions, or link jitter.
- High Power Module Issues
- I2C Issues

2.3.1 Link disconnection and flapping troubleshooting

When a Cumulus system device experiences link disconnection or jitter, the following situations may occur:

- "l1-show" returns link status: Kernel: Down and Hardware: Down for the operational state
- "ip link show " returns: NO-CARRIER,BROADCAST,MULTICAST,UP. An active uplink shows something like BROADCAST,MULTICAST,UP,LOWER_UP.
- "ip link show" output changes every 1–2 seconds, indicating the link is flapping.
- Log messages in "/var/log/linkstate" show that the carrier is going up or down.
- The switch does not receive any LLDP data, or the link is flapping.

To begin troubleshooting, use "l1-show" and check the output on both ends of the link whenever possible. The output includes all relevant information that can help identify and resolve link issues.

```
cumulus@switch~$ sudo l1-show swp10
Port: swp10
 Module Info
 Vendor Name: FS PN: QSFP-SR4-100G
 Identifier: 0x19 (QSFP) Type: 100G-SR4
 Configured State
 Admin: Admin Up Speed: 100G MTU: 9216
 Autoneg: On FEC: Auto
 Operational State
 Link Status: Kernel: Up Hardware: Up
 Speed: Kernel: 100G Hardware: 100G
 Autoneg: On (Autodetect enabld) FEC: Auto
 TX Power (mW): [0.5267]
 RX Power (mW): [0.5427]
 Topo File Neighbor: qct-ix8-51, swp3
 LLDP Neighbor: qct-ix8-51, swp3
 Port Hardware State:
 Compliance Code: 100G-SR4
 Cable Type: Optical Module (separated)
 Speed: 100G Autodetect: Enabled
 Eyes: 411 Grade: 41609
 Troubleshooting Info: No issue was observed.
```

2.3.1.1 Check the module information

```
Module Info
 Vendor Name: FS PN: QSFP-SR4-100G
 Identifier: 0x19 (QSFP) Type: 100G-SR4
```

Check the following information from the device output for accuracy.

- Does the module vendor name and part number match the module connected to the switch? Is this the correct port with the link issue? Is the correct module installed?
- Does the module type match the technology used for this link? For example, if it's a 100G DAC, is the type shown as 100GBASE-LR4? Refer to earlier sections for compliance codes, Ethernet type, Ethmode, and interface type.
- Does the remote device recognize the module as the same Ethernet type as identified by the local switch?

2.3.1.2 Check configuration status

```
Configured State
```

```
Admin: Admin Up Speed: 100G MTU: 9216
Autoneg: On FEC: Auto
```

- Admin: Is the port enabled? Has the link been configured and brought up?
- Speed: Is the configured speed correct? Does it match the configuration on the remote end?
- MTU: Are the MTU values on both ends matching? (Note: MTU mismatch won't stop the link from coming up, but it can affect traffic forwarding.)
- Autoneg: Does this setting align with your configuration and expectations? (Refer to the previous section on autonegotiation.)
- FEC: Is the FEC setting configured correctly?

2.3.1.3 Check operational status

```
Operational State
 Link Status: Kernel: Up Hardware: Up
 Speed: Kernel: 100G Hardware: 100G
 Autoneg: On (Autodetect enabld) FEC: Auto
 TX Power (mW): [0.5267]
 RX Power (mW): [0.5427]
 Topo File Neighbor: qct-ix8-51, swp3
 LLDP Neighbor: qct-ix8-51, swp3
```

- Link status (kernel and hardware): What is the current state of both?
  - Typically, the kernel and hardware status should be synchronized.
  - When troubleshooting link failure issues, one or both of these values will likely show "down" (usually both are "down").
  - For link jitter issues, one or both of these values might change every second or even more frequently, meaning the output at one moment may not reflect the next.
- Speed(kernel and hardware): Does the operating speed match the configured speed?
  - When the link is up, the operational values for both the kernel and hardware should be synchronized and match the configured speed.
  - When the link is down and auto-negotiation is enabled, the kernel value will show "unknown！" because the hardware is not synchronized to speed.
  - When the link is down and auto-negotiation is disabled, the kernel speed will display the configured value, while the "Hardware" field may show various values depending on the specific hardware interface implementation.
- Autoneg and Autodetect: Typically, the operational values should match the configured values.
- FEC: This field is for informational purposes only. The actual FEC is displayed only when the link is up.
  - When the link is down, the operational FEC will be "None".
  - When the link is up, this field displays the actual working FEC value on the link.
- TX Power and RX Power (Optical Modules / AOC with DDM functionality): If the module supports laser power DDM/DOM, are these values within the working range?
  - Use "ethtool -m swp" to check the TX High Alarm/TX Low Alarm and TX High Warning/TX Low Warning thresholds in the output to determine the expected low and high values. Refer to the technical specifications of the specific module to determine if a power alarm is generated.
  - A value of "0.0000" or "0.0" indicates that the module does not support DDM/DOM TX or RX power, or the module is not sending or receiving signals.
    If the TX power is "0.0000" or "0.0", it means that the module does not support TX DDM/DOM or the module's laser has been turned off for some reason.
    If the RX power is "0.0000" or "0.0", it means that the module does not support RX DDM/DOM or the module's receiver is not receiving a signal.
    A value of "0.0001" indicates that the module supports DDM/DOM, but the module has not sent or received a signal.

For QSFP modules, check the values for all four channels. Sometimes, if only one channel fails, the entire link can be down.

- Topo File Neighborptmd: If you have configured a topology file on the switch, you can identify the expected link neighbors.
- LLDP Neighbor: Does this match the expected neighbor and port?
  - This value is the neighbor and port reported by LLDP.
  - If the link is down, this value is usually blank.

2.3.1.4 Check port hardware status

```
Port Hardware State:
Compliance Code: 100G-SR4
Cable Type: Optical Module (separated)
Speed: 100G Autodetect: Enabled
Eyes: 411 Grade: 41609
Troubleshooting Info: No issue was observed.
```

- Compliance Code：
  - Does the firmware-recognized interface type match the installed module type? Is the module type correctly identified by the firmware?
- Cable Type：Does the cable type recognized by the firmware match the type of cable that is installed?
- Speed：
  - Does the speed match the expected speed?
  - If auto-negotiation is enabled and the link is down, it may show "N/A" or a speed different from what you expect.
- Autodetect：
  - If enabled, the link negotiation algorithm may fail with the partner device.
  - Try disabling auto-negotiation and setting a forced speed.
  - Refer to "Auto-Negotiation and Auto-Detection" for further guidance.
- Eyes and Grade：
  - If the link is down, all values will show as zero.
  - If the link is up, RX eye diagram (mV) and level values will be displayed.
  - Refer to "Eye" for more information.
- Troubleshooting Info：
  - What does the firmware assess the issue to be? Although this information is located at the end of the output, it is sometimes the first place to look for basic guidance.
  - Example:
    1. The port is closed by command. Please check that the interface is enabled. Configure the port to be in an administrative up state.
    2. The cable is unplugged. The firmware did not detect a module. Check if there is a module in this port or reinstall the module.
    3. Auto-negotiation no partner detected. The link is down because the neighbor is not visible. This alone doesn’t help much in pinpointing the cause.
    4. Force Mode no partner detected. Auto-negotiation or auto-detection is disabled, and the link is down because no neighbor was detected. This alone doesn’t help much in identifying the cause.
    5. Neighbor is sending remote faults. This link end is receiving data from the neighbor, but the neighbor is not receiving recognizable data from the local port. For more details, refer to "RX Fault" in the "Signal Integrity" section above. The local device is not transmitting data, and the remote receiver is either not receiving recognizable data or is receiving corrupted data.

Example 1: RX Signal Fault Example

Below is the output of a 100G SR4 showing an RX fault on channel 3 (on swp6). It indicates a fault on channel 3 of the local 100G module. If

the remote transmit power is normal, the issue is typically identified as a fault in the local module, and module replacement is recommended.

```
Port: swp6
Module Info
Vendor Name:FS PN: QSFP-SR4-100G
Identifier: 0x19 (QSFP) Type: 100G-SR4
Configured State
Admin: Admin Up Speed: 400G MTU: 9216
Autoneg: Off FEC: Off
Operational State
Link Status: Kernel: Down Hardware: Down <=Link is down, Kernel and Hardware
Speed: Kernel: 100G Hardware: 100G
Autoneg: Off FEC: None (down)
TX Power (mW): [1.1645, 1.171, 1.1155, 1.0945]
RX Power (mW): [0.159, 0.1732, 0.153, 0.0067] <=Low power on lane 3
Topo File Neighbor: switch_1, swp6
LLDP Neighbor: None, None
Port Hardware State:
Rx Fault: Local <=Local RX Failed Carrier Detect: no <=No bi-directional communication
Rx Signal: Detect: YYYY Signal Lock: YYYN <=No signal lock on lane 3
Ethmode Type: 400G-sr4 Interface Type: SR4
Speed: 400G Autoneg: Off
MDIX: ForcedNormal, Normal FEC: Off
Local Advrtsd: None Remote Advrtsd: None
Eyes: L: 357, R: 326, U: 211, D: 219, L: 328, R: 312, U: 206, D: 211,
L: 359, R: 343, U: 211, D: 200, L: 0, R: 0, U: 0, D: 0 <= No valid eye on lane 3
```

Below is the output of the "l1-show" command for a 100G SR4 module with faults on channels 0 and 1. Note that signal lock is flapping and occasionally shows as "Y". The module must be replaced.

```
Port: swp8
 Module Info
 Vendor Name: FS PN: OSFP-SR4-100G
 Identifier: 0x19 (QSFP) Type: 100G-SR4
 Configured State
 Admin: Admin Up Speed: 100G MTU: 9216
 Autoneg: Off FEC: Off
 Operational State
 Link Status: Kernel: Down Hardware: Down <=Link is down, Kernel and Hardware
 Speed: Kernel: 100G Hardware: 100G
 Autoneg: Off FEC: None (down)
 TX Power (mW): [1.1762, 1.1827, 1.1272, 1.1062]
 RX Power (mW): [0.0001, 0.0001, 0.5255, 0.64] <=Low power on lanes 0,1
 Topo File Neighbor: switch_2, swp10
 LLDP Neighbor: None, None
 Port Hardware State:
 Rx Fault: Local <=Local RX Failed Carrier Detect: no <=No bi-directional communication
 Rx Signal: Detect: YYYY Signal Lock: YNYY <=No lock on lane 1 at moment of capture
 Ethmode Type: 400G-SR4 Interface Type: SR4
 Speed: 400G Autoneg: Off
 MDIX: ForcedNormal, Normal FEC: Off
 Local Advrtsd: None Remote Advrtsd: None
 Eyes: L: 0, R: 0, U: 0, D: 0, L: 0, R: 0, U: 0, D: 0, <=No valid eyes on lanes 0,1
 L: 359, R: 359, U: 214, D: 226, L: 359, R: 359, U: 243, D: 264
```

2.3.2 Physical Link Errors and Troubleshooting

Physical link errors occur when there are signal integrity issues or the required FEC type is not configured on a specific module or cable type.

The target bit error rate (BER) for high-speed Ethernet is 2.4E(-4), but manufacturers' product specs usually exceed this threshold by far, with BER typically at (-7) or better. When the BER exceeds this value, configure the correct FEC settings or replace the optical module or patch cable. If the BER remains unacceptable with the correct FEC configuration, a hardware component in the link needs to be replaced to resolve the errors.

- To check the port's error counters, run the command "ethtool -S swp | grep Errors". If FEC is enabled, these counters only count errors that FEC fails to correct.
- On NVIDIA switches, to see the number of bit errors corrected by FEC on the link, run the command "sudo l1-show swp --pcs-errors".
- Since errors can occur during link up and down transitions, it's best to monitor error counters over time to ensure they increase regularly, rather than showing stale counts from the last link transition. The "/var/log/linkstate" log file shows the history of link up and down transitions on the switch.

2.3.3 Troubleshooting Signal Integrity Issues

Signal integrity issues are usually the root cause of different types of symptoms:

- If signal integrity is very poor or absent, the link will remain down.
- If signal integrity is poor, the link may be unstable whether FEC is enabled or not.
- If signal integrity is borderline, the link may show physical error counts. Depending on the link speed and cable type, the module or cable may have some signal integrity degradation. In these cases, FEC is used to correct errors and achieve the IEEE bit error rate (BER) target on the link.
- If FEC is enabled but the bitstream can't be restored to an acceptable level, the link will stay down. If signal integrity is borderline but too poor for FEC to correct to an acceptable error rate, the link will jitter when FEC signals a restart to try restoring the bitstream.
- To check the port's error counters, run "ethtool -S swp | grep Errors". If FEC is enabled, these counters only count errors FEC fails to correct.
- To see the number of bit errors corrected by FEC on the link, run "sudo l1-show swp --pcs-errors".
- Signal integrity issues are physical problems that usually require replacing some hardware components in the link to fix.
- In rare cases, signal integrity problems may be caused by the switch misidentifying the module type (active instead of passive). You need to check the product information in the EEPROM, including the module type, standard protocol, and other basic details.

2.3.4 High-Power Module Failures and Troubleshooting

The SFF MSA (CMIS) defines the power consumption for different modules. Please refer to the table below for the power specifications of 100G modules.

Table 8-27 Module Power Class and Max Power (Page 00h)

Byte	Bits	Field Name	Field Description	Type
200	7-5	ModulePowerClass[1]	000: Power class 1 001: Power class 2 010: Power class 3 011: Power class 4 100: Power class 5 101: Power class 6 110: Power class 7 111: Power class 8	RO Rqd.
	4-0	-	**Reserved**	RO
201	7-0	MaxPower	Maximum power consumption in multiples of 0.25 W rounded up to the next whole multiple of 0.25 W	RO Rqd.

Note 1: See relevant hardware specification for maximum power allowed in each Power class

- The following are the clear definitions for each power consumption class.

Class	1	2	3	4	5	6	7	8

Power Consumption	≤1.5W	≤2.0W	≤2.5W	≤3.5W	≤4.0W	≤4.5W	≤5.0W	＞5W

Switch ports have rated power support; some specific ports may also support a high-power mode. If the module requires high-power operation, this mode can be configured on the corresponding port. If the port supports it, the switch will approve the mode. To determine whether the switch supports higher power modes, refer to the hardware vendor’s specifications for power limitations.

The following section uses mid-to-low speed optical transceivers as examples to illustrate precautions when using high-power modes on device ports.

NVIDIA switches vary in their support for high-power modules. For example, on some NVIDIA Spectrum 1 switches, only the first and last QSFP ports support up to QSFP Power Level 6 (4.5W), and only the first and last SFP ports support SFP Power Level 3 (2.0W) modules. Other Spectrum 1 switches may not support high-power ports at all. For exact details on which ports support high-power modules, refer to the hardware manufacturer's specifications.

The total rated power is calculated by multiplying the default rated power of each port type (SFP: 1.5W, QSFP: 3.5W) by the number of ports of that type on the bus.

To check the request and enable status of high-power modules, review the output of "sudo ethtool -m". The following output is from a device with a power class level between 1 and 4 (1.5W to 3.5W). In this case, the module does not request a high power level, or the switch does not enable it.

```
cumulus@switch:mgmt:~# sudo ethtool -m swp53
 Identifier : 0x11 (QSFP)
 Extended identifier : 0x00
 Extended identifier description : 1.5W max. Power consumption <= ignore for high power modules
 Extended identifier description : No CDR in TX, No CDR in RX
 Extended identifier description : High Power Class (> 3.5 W) not enabled <= high power mode not requested or enabled
```

The following is the output from a Power Class 7 (5.0W) module. The module requests Power Class 7, but the switch does not support or enable it. The switch only supports up to Power Class 6 on this port.

```
cumulus@switch:mgmt:~# sudo ethtool -m swp49
[sudo] password for cumulus:
 Identifier : 0x11 (QSFP)
 Extended identifier : 0xcf
 Extended identifier description : 3.5W max. Power consumption <= ignore for high power modules
 Extended identifier description : CDR present in TX, CDR present in RX
 Extended identifier description : 5.0W max. Power consumption, High Power Class (> 3.5 W) not enabled <= Request 5.0W, not
enabled
```

The following is the output from a Power Class 6 (4.5W) module. The module requests Power Class 6, and the switch enables this feature.

```
cumulus@switch:mgmt:~# sudo ethtool -m swp3
 Identifier : 0x11 (QSFP)
 Extended identifier : 0xce
 Extended identifier description : 3.5W max. Power consumption <= ignore for high power modules
 Extended identifier description : CDR present in TX, CDR present in RX
 Extended identifier description : 4.5W max. Power consumption, High Power Class (> 3.5 W) enabled <= Request 4.5W, enabled
```

2.3.5 Troubleshooting I2C Issues

Ethernet switches contain multiple I2C buses, which are set up for the switch CPU to exchange low-speed control information with port modules, fans, and power supplies within the system.

2.3.5.1 Introduction to I2C Issues

In rare cases, a port module with a defective I2C component or firmware may malfunction and lock one or more I2C buses. Depending on the switch's specific hardware design and the nature of the failure, this issue may present various symptoms. Typically, traffic may continue to operate for some time, but the failure can sometimes lead to improper module configuration, resulting in link failures or increased error rates on the link. In the worst-case scenario, the switch may reboot or become unresponsive.

Since I2C issues occur in the module's low-speed control circuitry, high-speed traffic is not affected on the module's data side. Software bugs in Cumulus Linux do not cause these problems.

When an I2C bus experiences issues or becomes locked, installed port modules may no longer appear in the output of "sudo l1-show swp" or "sudo ethtool -m swp". The "/var/log/syslog" may contain a large number of "smbus", "i2c" or "EEPROM" read errors. Once one module locks the bus, some or all other modules may start to malfunction as well, making it nearly impossible to determine which module caused the failure. "EEPROM read/var/log/syslog"

The primary cause of I2C lock-ups is faulty I2C components or design flaws in port modules. Most failures are caused by low-cost vendor modules. However, even high-end, high-quality modules can fail—though at a much lower rate—due to their higher MTBF (Mean Time Between Failures) rating.

Removing port modules one by one until the problem is resolved can help identify the faulty module. However, clearing an I2C fault usually requires a reboot or power cycle, as bus lock-ups rarely clear on their own. While clearing the fault this way may work temporarily, edge I2C components can fail again after hours, days, or even months when conditions are right.

In the worst case, the switch may have multiple bad or defective I2C modules from the same vendor batch, making it difficult to pinpoint the exact faulty modules. Because I2C issues can be very serious and may reoccur long after initial resolution, it's important to address them quickly and decisively.

2.3.5.2 Troubleshooting Approach

To verify if an I2C failure has occurred, run "sudo tail -F /var/log/syslog" and look for continuous or sudden occurrences of "smbus/i2c/EEPROM" read errors.

Based on the severity of the issue when detected, decide whether to fix it immediately or wait until the scheduled maintenance window.

- If traffic or the switch is failing and you cannot reroute traffic through a redundant network, immediate action is required.
- If you can reroute traffic around the faulty switch and troubleshoot it without impact, continue rerouting traffic to find a suitable time for troubleshooting the switch.
- To troubleshoot and restore the switch, use the following options based on the urgency of the situation:
  - Remove port modules one by one to see if the issue clears. This has a lower chance of clearing the I2C fault but may have less impact on traffic. If successful, this method may identify the problematic module.
  - Restart the "switchd" process by running "sudo systemctl reset-failed ; sudo systemctl restart switchd". After the restart completes, verify whether the condition has cleared. This method has a moderate chance of clearing the I2C fault and causes a moderate impact on traffic. It does not help identify which module caused the fault.
  - Reboot the switch, and verify after the reboot whether the issue has been resolved. This method has a high probability of clearing the I2C fault but will significantly impact network traffic. It also does not help identify which module caused the fault.
  - Reboot the switch and verify whether the issue is resolved after the reboot. This method offers a very high chance of clearing the I2C fault but also causes significant impact on traffic. It does not provide a way to identify which module caused the fault.
  - If the I2C fault recurs shortly after rebooting, apply a binary elimination strategy by removing half of the modules at a time combined with a reboot.
  - If the I2C error still persists after removing all modules and powering off the switch, the next step is to remove each power supply and fan one by one after power-off to determine if any of these devices are blocking the I2C bus.
  - If you have removed all modules and each power supply, and tested the fans individually, but the I2C issue still persists, the final step is to replace the switch.

If the switch can operate again using one of the above methods but you haven't yet identified the module causing the issue, try the following steps:

- If the "syslog" file contains historical error records for a specific module before the failure occurred, start by removing or replacing that module.
- Replace any modules that have caused problems in the past.

- Replace all modules in the switch.

Ⅳ、MLNX-OS (Infiniband) System Troubleshooting Guide

NVIDIA InfiniBand technology, as a world-leading high-performance interconnect solution, has become the cornerstone for building AI factories, supercomputing centers, and cloud-native infrastructure thanks to its ultra-low latency, ultra-high throughput (supporting NDR/HDR rates up to 400 Gbps and beyond), and excellent scalability. By acquiring Mellanox and deeply integrating its Quantum-2 platform, Spectrum-X series switches, and BlueField DPU, NVIDIA has built an end-to-end accelerated computing ecosystem from chip to network, empowering real-time data processing needs in scenarios such as GPT large model training, scientific simulation, and autonomous driving.

Although InfiniBand technology offers outstanding performance, its complex architecture and demanding operational environment still pose multiple challenges. When using Quantum-2 platforms like the MSN2700-CS2RC IB switches paired with 100G modules, many issues remain.

From a daily customer perspective, we have gathered frequently reported problems. You can identify causes based on problem codes and follow the corresponding steps for troubleshooting and resolution.

1. Introduction to Core Diagnostic Commands of MLNX-OS

To quickly locate and resolve issues, it is essential to be familiar with MLNX-OS system commands. Using core diagnostic commands, you can swiftly obtain the operational status of transceivers and switches, facilitating analysis and troubleshooting. The following commands are commonly used diagnostics that help SR quickly gather product usage information and improve the efficiency of after-sales problem resolution.

1.1 Show Interfaces Ib Link-diagnostics

show interfaces ib link-diagnostics//show diagnostic information for a specific InfiniBand port or all InfiniBand ports.

- You can see interface information, fault codes, and interface status. Based on the fault codes, you can analyze the current port status, and refer to Chapter 4, Section 2 **Fault Code Display and Troubleshooting Approach** for detailed explanations and troubleshooting of the codes.
- Example：

```
switch (config) # show interfaces ib link-diagnostics
Interface Code Status
IB1/1 0 The port is Active.
IB1/2 0 The port is Active.
IB1/3 1024 Cable unplugged
IB1/4 1024 Cable unplugged
IB1/5 1024 Cable unplugged
IB1/6 1024 Cable unplugged
IB1/7 1024 Cable unplugged
IB1/8 1024 Cable unplugged
IB1/9 1024 Cable unplugged
IB1/10 1024 Cable unplugged
IB1/11 1024 Cable unplugged
IB1/12 1024 Cable unplugged
IB1/13 1024 Cable unplugged
IB1/14 1024 Cable unplugged
IB1/15 1024 Cable unplugged
IB1/16 1024 Cable unplugged
IB1/17 1024 Cable unplugged
IB1/18 1024 Cable unplugged
IB1/19 1024 Cable unplugged
IB1/20 1024 Cable unplugged
IB1/21 1024 Cable unplugged
IB1/22 1024 Cable unplugged
IB1/23 1024 Cable unplugged
IB1/24 1024 Cable unplugged
```

```
IB1/25 1024 Cable unplugged
IB1/26 1024 Cable unplugged
IB1/27 1024 Cable unplugged
IB1/28 1024 Cable unplugged
IB1/29 1024 Cable unplugged
IB1/30 1024 Cable unplugged
IB1/31 1024 Cable unplugged
IB1/32 1024 Cable unplugged
IB1/33 1024 Cable unplugged
IB1/34 1024 Cable unplugged
IB1/35 1 The port is closed by command.
IB1/36 2 Auto-Negotiation failure..
```

1.2 Show Interfaces Ib Internal Leaf Link-diagnostics

show interfaces ib internal leaf link-diagnostics//Show diagnostic information for a specific InfiniBand internal spine module/port link.

- You can see interface information, fault codes, and interface status. Based on the fault codes, you can analyze the current port status, and refer to Chapter 4, Section 2 **Fault Code Display and Troubleshooting Approach** for detailed explanations and troubleshooting of the codes.
- Example：

```
switch (config) # show interfaces ib internal leaf 1 link-diagnostics
----------------------------------------------------------------------
Interface Code Status
----------------------------------------------------------------------
IB1/1/19 0 No issue was observed
IB1/1/20 0 No issue was observed
IB1/1/21 0 No issue was observed
IB1/1/22 0 No issue was observed
IB1/1/23 0 No issue was observed
IB1/1/24 0 No issue was observed
IB1/1/25 0 No issue was observed
IB1/1/26 0 No issue was observed
IB1/1/27 0 No issue was observed
IB1/1/28 0 No issue was observed
IB1/1/29 0 No issue was observed
IB1/1/30 0 No issue was observed
```

1.3 Show Interfaces Ib Internal Spine Link-diagnostics

show interfaces ib internal spine link-diagnostics//show diagnostic information for a specific InfiniBand internal spine module/port link.

- You can see interface information, fault codes, and interface status. Based on the fault codes, you can analyze the current port status, and refer to Chapter 4, Section 2 **Fault Code Display and Troubleshooting Approach** for detailed explanations and troubleshooting of the codes.
- Example：

```
switch (config) # show interfaces ib internal spine 3/1/1 link-diagnostics
----------------------------------------------------------------------
Interface Code Status
----------------------------------------------------------------------
IB3/1/1 0 No issue was observed
```

2. Fault Code Display and Troubleshooting Approach

Monitoring Code	Detailed Description	Detailed Countermeasures
0—No issue observed		After this prompt appears, if the device port does not show a normal link, please wait 5 minutes and then check again. If the code persists and the port status hasn't improved, please check the opposite end.
1—Port is close by command	The port has been disabled via command, usually by using shutdown/no shutdown to turn the port off or on.	Check if a shutdown command was issued; please use the command to reopen the port.
2—AN failure	Adaptive failure of speed/FEC or DME on both sides	Troubleshooting steps: Check TX and RX power on both ends: If TX power is too low: check the transceiver module by swapping to see if the inserted module is faulty. If RX power is too low: check the TX power on the other side, test the line attenuation if possible, and clean the connector endfaces of both devices and patch cords after inspection. Check configurations on both ends: Ensure matching speed settings, FEC configuration, or adaptive mode.
9—Logical mismatch between link partners	Link partner logical mismatch, no lock between lane blocks	Check if the configurations on both ends of the device are correct The same speed settings, FEC configuration, or auto-negotiation mode. If the problem persists, more data should be collected and then proceed with an upgrade.
10—Logical mismatch between link partners	Link partner logic mismatch, AM region blocks not locked (NO FEC)	
11—Logical mismatch between link partners	Link partner logic mismatch, align_status not received; auto-negotiation configured but signal remains unlocked	
12—Logical mismatch between link partners	Link partner logic mismatch, FC FEC not specified	
13—Logical mismatch between link partners	Link partner logic mismatch, RS FEC not specified	
14—Remote fault received	Received remote error indication	After this message appears, please wait 5 minutes and check again. If the code persists and the port status doesn’t improve, please check the remote side.

15/52—Bad signal integrity	Low raw bit error rate (BER). Please note, when checking the raw BER, do so after the device has been running for some time.	Phenomenon: The port connection is up, but the raw bit error rate (BER) is relatively high. Troubleshooting steps: Wait for a period of time and observe if the issue improves. Clean the end faces of the modules and patch cords on both sides. Check the TX and RX power on both ends of the device: If the transmit power is too low: check the transceiver module and use a replacement method to identify if the module inserted in the port is faulty. If the receive power is too low: check the transmit power on both ends, and if possible, test the attenuation of the fiber link. After checking, be sure to clean the end faces of the modules and patch cords on both sides. Collect the signal-to-noise ratio (SNR) for both optical and electrical sides on both ends. You can use mlxlink -m or other supported tools to collect this data. In such cases, use the module's internal PRBS (Pseudo-Random Bit Sequence) function to test. For PRBS testing, the following adjustments are needed: If PRBS test is fine, firmware debugging of the module is needed. If there is a very low BER, SerDes interface debugging is needed. If a specific channel is not locked, it may be caused by the module, NIC, firmware, or interface issues.
16—Cable compliance code mismatch (protocol mismatch between cable and port)	Cable type error, or the cable type is not supported by the device	Check if the port speed is configured according to the cable's specifications. Verify whether the cable type is supported by the device.
20—Stamping of non-NVIDIA Cables/Modules	Using third-party modules on NVIDIA devices causes errors	Replace with NVIDIA original cables or NVIDIA-certified cables. For third-party cables, the product's EEPROM must contain encryption information certified by NVIDIA devices.
21—Down by PortInfo MAD	Port shutdown caused by changes from PortInfo MAD	Check who issued the shutdown command for the port and re-enable the port using the appropriate command.
23—Internal error	Information calibration failed	Collect product information from both ends of the device, preferably including the EEPROM encoding data. Reboot the device to check if the issue persists. If it does, compare the calibration bits in the EEPROM to verify if they are normal.
24—EDR speed is not allowed due to cable stamping: EDR stamping		

25—FDR10 speed is not allowed due to cable stamping: FDR10 stamping	The cable is invalid or non-NVIDIA branded, not certified, and does not support the corresponding InfiniBand technology.	Replace with an original NVIDIA cable or an NVIDIA-certified cable. If using a third-party cable, ensure that the product's EEPROM contains the encrypted information required for certification by NVIDIA devices.
26—Port is closed due to cable stamping: Ethernet_compliace_code_zero		
27—Port is closed due to cable stamping: 56GE stamping		
28—Port is closed due to cable stamping: non-NVIDIA QSFP28		
29—Port is closed due to cable stamping: non-NVIDIA SFP28		
30—Port is closed, no backplane enabled speed over backplane channel	The port is shut down, and the backplane speed is not enabled on the backplane channel.	Check whether the port configuration is correct: ensure that the port matches the cable's speed, bandwidth, FEC type, and whether Auto-Negotiation (AN) is fully enabled.
31—Port is closed, no passive protocol enabled over passive copper channel	Port is down because passive protocol is not enabled on the passive copper channel.	
32—Port is closed, no active protocol enabled over active channel	Port is down because no active protocol is enabled on the active channel.	
33—Port width is does not match the port speed enabled	Port bandwidth does not match the enabled port speed	

		Phenomenon: The device port is linked up, but the rate is lower than expected. Troubleshooting steps: Wait and observe if the issue improves over time. Clean the end faces of modules and jumpers on both sides. Check TX and RX power on both ends: If TX power is too low: check the transceiver module, use swap method to identify faulty module. If RX power is too low: check TX power on both ends, test line attenuation if possible, then clean modules and jumpers. Collect optical and electrical SNR on both ends: Use mlxlink -m or other supported tools. If needed, use the module's internal PRBS function for testing; perform adjustments as required: If PRBS is fine, tune the module firmware. For very low error rate, adjust SerDes interface. If a specific channel fails to lock, it could be module, NIC, firmware, or interface related.
34—Local Speed degradation	Local rate decrease	
35—Remote Speed degradation	Remote Rate Degradation	Check the remote device's port status and troubleshoot following the procedures outlined for error code 34 as a reference.
36—No Partner detected during force mode. 37—Partial link indication during force mode.	When rate and FEC mode are forced, the link partner is not detected (due to mismatched configuration) or only partial link indicators are observed.	Troubleshooting steps: Check TX and RX optical power on both ends: If TX power is too low: Investigate potential issues with the transceiver module. Try replacing it to identify if the issue lies with the module inserted into the port. If RX power is too low: Examine TX power on both sides. If possible, test for link attenuation. After checking, thoroughly clean the ends of both modules and patch cables. Verify port configuration on both ends: Ensure both ends have matching speed settings, FEC type, and auto-negotiation (AN) is fully enabled if required.
38—AN failure	FEC type mismatch under override configuration mode.	
39—AN failure	No HCD detected	Check if the configurations on both ends of the port are correct. Verify that the rate settings, FEC type, and auto-negotiation are the same and fully enabled.
42—Bad SI, cable is configured to non optimal rate	Poor cable SI parameters; the cable is configured for a non-optimal rate.	
51—HST speed mismatch	HST rate mismatch	
53—Link failure due to MCB at link up	When the link comes up, the link failure is caused by MCB.	

54—PLR didn't get Rx good non sync cell	PLR did not receive Rx good non-synchronous cells.	Wait for 10 seconds; if the error persists, then exchange shared information between both sides and switch the link.
55—PSI fatal error	PSI fatal error	
57—signal not detected 59—Did not get module conf done	Power detection not detected in SerDes	Wait and observe if the issue improves over time. Clean the end faces of modules and jumpers on both sides. Check the TX and RX power on both ends of the device: If the transmit power is too low: check the transceiver module and use a replacement method to identify if the module inserted in the port is faulty. If the receive power is too low: check the transmit power on both ends, and if possible, test the attenuation of the fiber link. After checking, be sure to clean the end faces of the modules and patch cords on both sides. Collect the signal-to-noise ratio (SNR) for both optical and electrical sides on both ends. You can use mlxlink -m or other supported tools to collect this data. In such cases, use the module's internal PRBS (Pseudo-Random Bit Sequence) function to test. For PRBS testing, the following adjustments are needed: If PRBS test is fine, firmware debugging of the module is needed. If there is a very low BER, SerDes interface debugging is needed. If a specific channel is not locked, it may be caused by the module, NIC, firmware, or interface issues.
128—Troubleshooting in process	Troubleshooting is in progress	Wait for 3 seconds, then rerun the port diagnostic command to check if the fault indication has cleared.
1023—Info not available	Information unavailable	Wait for 10 seconds. If the problem persists, restart the device and collect more information.
1024—Cable is unplugged	The switch port did not detect the transceiver.	Insert the transceiver (optical module or high-speed cable).
1025—Long Range for non Mellanox cable/module .	Non-NVIDIA long-distance cables/modules are not supported	Replace with genuine NVIDIA original parts.
1026—Bus stuck (I2C Data or clock shorted)	A fault was detected on the I2C EEPROM communication line — the bus is stuck (I2C data or clock line is shorted).	Reset the transceiver (disable/enable). If the issue persists, reboot the device and collect more diagnostic information.
1027—Bad/unsupported EEPROM	The EEPROM is corrupted or unsupported. The system cannot read the EEPROM from the transceiver or cannot recognize the transceiver ID.	This is usually due to a compatibility issue or transceiver failure. Replace it with a known good module for comparison and validation.
1028—Part number list	The transceiver is not listed in the approved vendor list.	This is usually identified as a compatibility issue. You should replace the cable with one from the supported list, or modify the EEPROM information in the cable to include NVIDIA-certified data.

1029—Unsupported cable.	The transceiver is not supported.	
1030—Module temperature shutdown	The module temperature has exceeded the allowed threshold range.	Check and confirm the transceiver temperature and ambient temperature. If the temperature remains high, take the following actions to mitigate the issue: Clean the transceiver and patch cord end faces. Increase the device's fan speed or cooling level to enhance heat dissipation. Lower the ambient temperature, e.g., by enabling external cooling systems. Reduce the transceiver density per device to decrease the load on the ports. Use lower-power-consumption modules.
1031—Shorted cable	The operating current on the transceiver is too high.	The transceiver is damaged and needs to be replaced with a new one.
1032—Power budget exceeded	The power budget of the circuit board has been exceeded.	Check the transceiver and device information to confirm the total supported power load of the device.
1033—Management forced down the port	The unit/module was shut down via server command.	Check the server commands and use them to enable the unit/module.
1034—Module is disabled by command	The transceiver's management status is disabled.	Enter the port mode and enable the port using the command, usually by executing no shutdown.
1036—Module's PMD type is not enabled (see PMTPS).	The transceiver's Type is not supported.	Check whether the transceiver's Type is supported by the device. Replace with a transceiver type supported by the device. Modify the EEPROM information to a Type supported by the device.
1044—Module's stamping speed degeneration	HDR rate is not supported.	Replace with a product matching the corresponding IB technology. If it's already the correct IB rate product, check the transceiver's EEPROM bytes declaring the IB technology.
1045—Module's stamping speed degeneration	EDR rate is not supported.	
1046—Module's stamping speed degeneration	FDR 10 rate is not supported.	
1047/1048—Modules DataPath FSM fault	Transceiver configured speed (application) failure.	Wait 10 seconds and observe if the alarm improves; if not, restart the switch and collect startup logs. Output the transceiver's EEPROM information on the device and check whether the application inside the transceiver is compliant.
2048—MPR Violation (Under 64 bytes between two starts).	MPR vulnerability (interval between two startups less than 64 bytes)	Wait 10 seconds and observe if the alarm improves; if not, restart the switch and collect startup logs.

V、Typical FS Case Collection

1. Transmission Issues

1.1 Flapping

Symptom description:

The customer reported CRC errors on the QSFP-LR4-100G module. Replacing and cleaning the patch cable did not resolve the issue.

Involved devices:

S5860-48SC

Related products:

QSFP-IR4-40G

QSFP-LR4-100G

Analysis approach:

1. Physical layer causes flapping due to line issues, resulting in unstable optical or electrical signals; it may also be caused by improper port speed settings, leading to negotiation failures between devices and intermittent transmission.

2. Network/data layer causes flapping due to improper routing design, resulting in frequent equal-cost path changes or route update oscillations, causing route instability during transmission. It can also be caused by MAC address flapping, leading to interface status fluctuations.

3. Module-related flapping occurs when internal components of the module have issues, causing transmission faults and intermittent link connectivity.

4. Port-related flapping happens when the device port itself has faults, causing intermittent transmission and unstable link status.

Problem solution:

Flapping is a common transmission issue caused by multiple factors:

1. Flapping caused by physical layer issues: Usually check by swapping cables to see if the problem is cable-related. If yes, troubleshoot or replace the cable. Also, verify both ends have matching port speeds; if it’s a config issue, resetting should fix it.

2. Flapping caused by network/data layer: Remove the faulty device from the system and test it standalone with loopback or connected to one device. If no flapping occurs, the problem lies in network/data layer, so check routing and network configs.

3. Flapping caused by the module itself: Check the module’s DDM for abnormal parameters. Swap the module or high-speed cable to test if replacing them solves the issue.

4. Flapping caused by the port itself: Use the module/cable on another port; if no flapping, the original port is faulty.

Solution：

1. Flapping caused by the physical layer: Replace the connecting cable for comparative testing to check if the issue is cable-related. If it is, troubleshoot or replace the cable. Check whether the port speeds on both ends match; if it's a configuration issue, resetting the speed should fix it.

2. Flapping caused by network/data layer: Remove the faulty device from the system, run a loopback test or connect it to a single device to see if flapping persists. If the issue disappears, it indicates a network/data layer problem, requiring investigation of routing and network configurations.

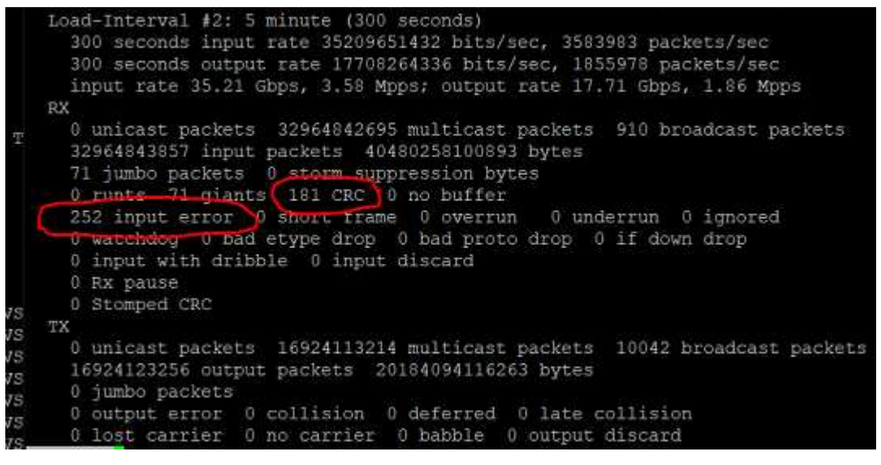
3. Flapping caused by the module itself: Read the module’s DDM to check for abnormal parameters. You can also swap modules or high-speed cables in the same environment to verify if the problem is resolved.

4. Flapping caused by the port itself: Use the module/cable on another port; if flapping no longer occurs, the issue lies with the original device port.

1.2 CRC Errors

Problem manifestations on the switch:

After the module is inserted into the switch, the switch will pop up an alarm message for DATA_CRC_ERROR (or it can be viewed by reading the switch's working day log information via show logging), or CRC error statistics faults on the port can be viewed via show interfaces Port Number (the command to view port status), etc.



Principle of CRC generation:

CRC is a common technique used to detect data transmission errors. The message to be sent is divided into fixed-length blocks, which are divided by a fixed divisor. The remainder from this division is appended to the message and sent together. Upon receiving, the computer recalculates the remainder and compares it with the sent remainder. If the numbers don't match, an error is detected.

For example, when transmitting a video segment defined to have length 10, it is divided by 3, resulting in a remainder of 1. This remainder is attached to the video data before transmission. The receiver also defines the video length as 10, divides by 3, and calculates its remainder. If the remainders match, there's no error; if they differ, it means the signal was corrupted during transmission.

Causes of the issue and solutions:

1. CRC errors caused by cable issues occur because the cable itself has problems, leading to transmission errors related to CRC checking. Typically, customers experience errors on this link, but after replacing the cable, the errors disappear. In such cases, CRC errors are usually caused by the cable. It is recommended that customers either fix the cable or use a different link.

2. CRC errors caused by hardware issues occur when the module or cable itself is faulty, causing transmission errors related to CRC checking. This usually shows up as CRC error alarms on one or some devices on the same link, while other modules work normally. This kind of problem is generally due to hardware faults in the product. It's important to try to reproduce the issue in the lab—if reproducible, it can likely be resolved (depending on the switch). If not, the customer should replace the product for testing.

3. CRC errors caused by encoding issues happen when the product's encoding is incompatible with the customer's device, resulting in CRC alarms during use. Typically, this manifests as alarms on the customer's equipment:

%GBIC_SECURITY_CRYPT-4-VN_DATA_CRC_ERROR: GBIC in port 54 has bad crc *Mar 21 13:21:31.656: Warning: SFP gbic-security check fail

This belongs to encoding recognition CRC errors. Such issues are generally caused by encoding problems leading to CRC errors. Depending on the specific brand of equipment, provide the customer with a suitable encoding replacement.

2. Encoding Issues

2.1 PID Not Supported

Symptom description:

When the customer uses the **QSFP-100G-ER4** module to connect to the device, the switch or NIC cannot recognize the module. The logs show errors such as **"Unsupported PID"** or **"Invalid Transceiver"**. The device port status shows **DOWN**, rate negotiation fails, and the module cannot be activated.

Involved devices:

Cisco 9901 ASR router

NVIDIA SN4700

Related products:

QSFP-100G-ER4

Logs:

```
[root@Switch]# show interface ethernet 1/1/1 transceiver
Status: Invalid Transceiver (PID: 0x0000)
Status: Invalid Transceiver (PID: 0x0000)
```

```
Vendor: FS
Part Number: QSFP-DD-100G-ER4
Error: PID not supported by current firmware.
[root@Server]# lspci -vv | grep Mellanox
01:00.0 Ethernet controller: Mellanox Technologies MT28908 [ConnectX-6 HDR] Subsystem: Mellanox Technologies Device 0020
Physical Port: 0 Link: Down (Unsupported transceiver detected)
```

Analysis approach:

1. Check firmware version compatibility

- Verify whether the device firmware version supports the module's PID (refer to the vendor compatibility list).

2. Module PID definition error

- The PID field in the module's **EEPROM** is not defined according to the standard protocol (e.g., CMIS/SFF-8024), or the value written is incorrect.
- Use i2cdump or vendor tools to read the module EEPROM and check whether **Bytes 148-163 (the PID field)** comply with the specification.

3. Hardware compatibility limitations

- Hardware design limitations (e.g., ASIC chip version) prevent support for newly released PIDs.
- Refer to the vendor's published **Hardware Compatibility List (HCL)** to verify that the module is compatible with the device model.

4. Vendor-specific customization restrictions

- Some vendor devices lock out non-certified modules (e.g., "original module only" policy) by blocking third-party PIDs through firmware restrictions.

Solution：

1. Firmware upgrade or patch installation

- Switch firmware upgrade

```
[Switch]# download firmware from tftp://10.1.1.1/sn4700_fw_v5.0.0.img
[Switch]# install firmware sn4700_fw_v5.0.0.img
[Switch]# reboot
```

2. Module EEPROM PID repair

- Use programming tools (e.g., ethtool or vendor EEPROM utilities) to modify the module's PID field to match the device-supported PID codes.
- Example (correcting CMIS PID field):

```
[root@Server]# ethtool -m eth0 | grep "0x0094"
0x0094: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
# 写入标准PID（如QSFP-DD-400G-LR4 PID: 0x1E 0x1F） [root@Server]# ethtool -E eth0 magic 0xABCDEF offset 0x94 value 0x1E
[root@Server]# ethtool -E eth0 magic 0xABCDEF offset 0x95 value 0x1F
```

3. Disable vendor authentication

Enable **third-party module support** in the device configuration (requires vendor authorization):

```
[Switch]# configure terminal
```

```
[Switch(config)#]interface ethernet 1/1/1
[Switch(config-if)#]transceiver third-party enable
```

2.2 Encoding Display Issues

Symptom description:

The customer recently placed an order for 10 pcs of #168310 QSFP-ZR4-40G. The modules function properly, but the device displays them as "ER4".

Involved devices:

JunOS

Related products:

QSFP-ZR4-40G

Logs:

```
eknell@br1.che.ecolink.coop-re0> show chassis hardware
Hardware inventory:
Item Version Part number Serial number Description
Chassis V3121 JNP204 [MX204]
Routing Engine 0 BUILTIN BUILTIN RE-S-1600x8
CB 0 REV 34 750-069579 CAMG3374 JNP204 [MX204]
FPC 0 BUILTIN BUILTIN MPC
 CPU REV 02 750-066879 CAGC8782 MPC
 Xcvr 1 REV 01 740-059185 G2006590566 QSFP+-40G-ER4
 Xcvr 2 REV 01 740-060379 S2120726422 QSFP+40GE-AOC-15M
 Xcvr 3 REV 01 740-059185 C2312707295 QSFP+-40G-ER4
Fan Tray 0 Fan Tray, Front to Back Airflow - AFO
Fan Tray 1 Fan Tray, Front to Back Airflow - AFO
Fan Tray 2 Fan Tray, Front to Back Airflow - AFO
```

Analysis approach:

1. **EEPROM information error**

- The model field in the module's **EEPROM** (e.g., **Byte 148–163**) may be incorrectly written or failed verification, causing abnormal display on the device.
- Check whether the "Part Number" field in the EEPROM complies with the SFF-8636/CMIS protocol format (ASCII encoding, fixed length of 16 bytes).

2. Alternative Coding Display

- The primary purpose of module coding is to allow switch devices to properly recognize the module. It does not affect the execution of the module's hardware functions.

Solution：

1. **Correct the module EEPROM model field**

- Use programming tools (such as "ethtool" or vendor-specific EEPROM utilities) to rewrite the correct Part Number.
- Enable third-party module model display in the switch configuration (may require admin privileges).
- **Check the port mode settings** and revert the port to auto-detection mode.

3. Quality Issues

3.1 Abnormal Heat Dissipation

Symptom description:

The #48722 received by the customer operates normally on the C9500-48Y4C. However, a warning appears every hour. Technical assessment suggests that the underlying coding needs to be modified, adjusting the temperature threshold to 0–70°C.

Involved devices:

Cisco C9500-48Y4C

Related products:

QSFP-BD-40G

DWDM2-Q28100G-80

QSFP28-LR4-100G

Logs:

```
 High Alarm High Warn Low Warn Low Alarm
 Temperature Threshold Threshold Threshold Threshold
Port (Celsius) (Celsius) (Celsius) (Celsius) (Celsius)
--------- ----------------- ---------- --------- --------- ---------
Hu1/0/51 0.0 75.0 70.0 10.0 5.0

 High Alarm High Warn Low Warn Low Alarm
 Voltage Threshold Threshold Threshold Threshold
Port (Volts) (Volts) (Volts) (Volts) (Volts)
--------- ----------------- ---------- --------- --------- ---------
Hu1/0/51 0.00 3.63 3.46 3.10 2.97

 High Alarm High Warn Low Warn Low Alarm
 Current Threshold Threshold Threshold Threshold
Port Lane (milliamperes) (mA) (mA) (mA) (mA)
--------- ---- --------------- ---------- --------- --------- ---------
Hu1/0/51 1 0.0 10.0 9.5 1.0 0.5
Hu1/0/51 2 0.0 10.0 9.5 1.0 0.5
Hu1/0/51 3 0.0 10.0 9.5 1.0 0.5
Hu1/0/51 4 0.0 10.0 9.5 1.0 0.5

 Optical High Alarm High Warn Low Warn Low Alarm
 Transmit Power Threshold Threshold Threshold Threshold
Port Lane (dBm) (dBm) (dBm) (dBm) (dBm)
--------- ---- --------------- ---------- --------- --------- ---------
Hu1/0/51 1 -40.0 5.0 4.0 -2.0 -3.0
Hu1/0/51 2 -40.0 5.0 4.0 -2.0 -3.0
Hu1/0/51 3 -40.0 5.0 4.0 -2.0 -3.0
Hu1/0/51 4 -40.0 5.0 4.0 -2.0 -3.0

 Optical High Alarm High Warn Low Warn Low Alarm
 Receive Power Threshold Threshold Threshold Threshold
Port Lane (dBm) (dBm) (dBm) (dBm) (dBm)
--------- ---- --------------- ---------- --------- --------- ---------
Hu1/0/51 1 -40.0 5.0 4.0 -8.0 -9.0
Hu1/0/51 2 -40.0 5.0 4.0 -8.0 -9.0
Hu1/0/51 3 -40.0 5.0 4.0 -8.0 -9.0
Hu1/0/51 4 -40.0 5.0 4.0 -8.0 -9.0
```

Analysis approach:

1. Inadequate environmental cooling

- Poor ventilation in the equipment rack or fan failure may cause high ambient temperatures around the module.
- Check rack temperature, fan speed, and airflow design for any issues.

2. Module thermal design flaws

- Poor contact between the heatsink and module casing or insufficient thermal material performance may result in ineffective heat dissipation.
- Compare temperature performance with similar modules (e.g., test another brand with the same model).

3. Excessive port density

- High-density port configurations (e.g., 32x400G) can concentrate heat, exceeding the system's cooling capacity.
- Verify if the device's thermal specs (e.g., SN4600C's maximum port power support) match the module's requirements.

4. Abnormal module power consumption

- Internal circuit issues (e.g., excessive laser driver current) may cause a power surge.
- Use DDM to check if **TX Bias Current** and **TX Power** are beyond the normal range.

5. The module's threshold range is incorrect

- The module's threshold range must be consistent with the temperature range defined in the module's coding thresholds.

Solution：

1. Optimize cooling environment

- Clear obstructions in the rack airflow path, ensure at least 30cm of front-to-back clearance for heat dissipation.
- Replace faulty fans or upgrade to high-speed fan modules (e.g., NVIDIA SN4600C high-speed fan kit).

2. Replace with higher thermal performance modules

- Choose modules with a wider operating temperature range (e.g., industrial-grade modules).

3. Reduce port density or power consumption

- Distribute high-power modules across different line cards or switches to avoid localized overheating.
- Enforce port rate limits to lower power consumption (if business allows).

4. Modify module temperature thresholds

- Use DDM to check abnormal parameters, then edit **Byte** 220 to 00 in the EEPROM coding to disable DDM reporting.

3.2 TX Abnormality

Symptom description:

When the customer uses **QSFP-100G-LR4** or **QSFP28-100G-SR4** modules, the link frequently drops or experiences a high bit error rate. The device logs show alarms such as "**Tx Power Low**" or "**Laser Fault**." DDM parameters indicate the **Tx Power** is significantly below specification (for example, LR4 module Tx Power < -10 dBm, while the normal range is -4.3 to 4.5 dBm).

Involved devices:

NVIDIA SN5600

Related products:

QSFP-100G-LR4

QSFP28-100G-SR4

Logs:

Port	Lane	Optical Transmit Power (dBm)	High Alarm Threshold (dBm)	High Warn Threshold (dBm)	Low Warn Threshold (dBm)	Low Alarm Threshold (dBm)
Fo1/1/1	1	1.1	4.3	3.3	-7.0	-8.2
Fo1/1/1	2	-0.1	4.3	3.3	-7.0	-8.2
Fo1/1/1	3	1.3	4.3	3.3	-7.0	-8.2
Fo1/1/1	4	0.5	4.3	3.3	-7.0	-8.2

Port	Lane	Optical Receive Power (dBm)	High Alarm Threshold (dBm)	High Warn Threshold (dBm)	Low Warn Threshold (dBm)	Low Alarm Threshold (dBm)
Fo1/1/1	1	-17.6	4.5	2.5	-14.1	-16.4
Fo1/1/1	2	-1.1	4.5	2.5	-14.1	-16.4
Fo1/1/1	3	-2.5	4.5	2.5	-14.1	-16.4
Fo1/1/1	4	-0.7	4.5	2.5	-14.1	-16.4

Analysis approach:

1. TOSA hardware damage

- Laser diode (LD) aging, driver circuit faults, or optical path contamination causing abnormal transmit power.
- Check DDM parameters: **Tx Bias Current** exceeding limits (e.g., >150mA) or **Tx Power** approaching zero.

2. Excessive fiber link loss

- Fiber bend radius too small, connector contamination, or excessively long patch cables causing signal attenuation beyond module tolerance.
- Measure end-to-end link loss with an optical power meter (LR4 modules tolerate up to 4.5dB loss).

3. Module power supply abnormality

- Insufficient or unstable power at device port causing laser driver failure.
- Compare module **Voltage** value (normal range: 3.13–3.47V).

4. Firmware compatibility issues

- Outdated device firmware misinterprets power thresholds (e.g., interpreting normal -5dBm as low power alarm).

Solution：

1. Replace faulty optical module

- If DDM shows **Tx Power consistently below -10dBm** and cleaning the fiber doesn’t help, replace the module directly.

2. Clean or replace fiber link

- Use a fiber cleaning pen to clean the module endface and patch cable connectors.
- Test with low-loss patch cables (e.g., UPC to APC connectors) or shorten the fiber length.

Shenzhen（China）

Address：Room 1903-1904, Block C, China Resources Tower, Dachong Community, Yuehai Subdistrict, Nanshan District
Email：sales@feisu.com
Tel：+86(400)865 2852

Shanghai（China）

Address：Unit 1201, Lee Gardens Shanghai Office Tower, No. 668 Xinzha Road, Jing'an District
Email：sales@feisu.com
Tel：+86(400)865 2852

Wuhan（China）

Address：Building A1-A4, Chuangxin Tiandi, No. 88 Guanggu Sixth Road, Hongshan District
Email：sales@feisu.com
Tel：+86(400)865 2852