DESIGNCON® 2024
WHERE THE CHIP MEETS THE BOARD

800G Linear Direct Drive Network System Design & Implementation

Feng Luo	ByteDance
Luoyi Fang	ByteDance
Chuansheng Cheng	ByteDance
Lei Guo	ByteDance
Dayong Shen	ByteDance
Haohao Kang	ByteDance
Anbing Sun	Ruijie Networks
Zhan Su	Ruijie Networks
Xiuguo Jiang	Keysight

Abstract

In the data center network system, the application of LPO (Linear-direct-drive Pluggable Optics) modules has certain advantages in terms of cost, power consumption, and latency compared to traditional DSP modules. However, it also poses significant challenges to the end-to-end system design, including signal integrity design between devices and modules, the impact of optical-electrical link design consistency on the system, and the controversy over testability. These challenges currently hinder the practical implementation of LPO. This article will showcase high-speed design and testing schemes for LPO systems from a system perspective. It will present the testing results, data, and product architecture of a linear direct-drive switch in conjunction with the LPO module system.

Authors Biography

Dayong Shen is a engineer at ByteDance. He is focusing on data center network solutions. He has 16 years of high-speed signal integrity design experience.

Haohao Kang is a staff engineer at ByteDance. He is familiar with data center optical module solutions and has a good understanding of system end-to-end link characteristics. He has 5 years of experience in the selection and certification of data center optical module.

Anbing Sun is a senior signal and power integrity engineer at Ruijie. He currently leads the PCB/SI team in system solutions and has 15 years of experience in the development of data center network products.

Zhan Su， an optical architect at Ruijie Networks, has years of experience in the advanced research of novel optical technologies.

Xiuguo Jiang is Great China PSS SE&CSM Manager at Keysight Technologies, where he focuses on Signal Integrity, Power Integrity, and EMC. He has more than 13 years of experience in Hardware and Signal integrity. He is the author of over 5 books on SI, PI, and EE.

1 Introduction

With the rapid development of AIGC/cloud computing technology, higher demands are placed on the bandwidth capacity and forwarding latency of HPC/DCN networks. During the upgrade process of network equipment switch capacity from 25.6T to 51.2T, and port rates from 400G to 800G, the increase in system density and complexity poses significant technical challenges in terms of heat dissipation and power consumption design. Faced with this challenge, the industry has proposed a solution to the system's heat dissipation design challenge by removing the oDSP chip from the optical module and adopting a linear direct-drive approach to reduce the power consumption of the optical module, while also reducing data transfer latency.

Currently, the mainstream 800G optical modules in the industry typically adopt traditional re-timer solutions, implementing signal regeneration and utilizing oDSP for digital signal compensation techniques such as dispersion compensation, non-linear compensation, noise removal, etc. This design can achieve better system performance, lower electrical signal error rates, and provide reliable support for the transmission of network signals.

In comparison to traditional re-timer module solutions, as shown in Figure 1-1, LPO modules retain the advantages of pluggable modules. By subtracting on the module side, they reduce the power consumption and heat dissipation challenges of switch systems. CDR/oDSP and other re-timer components are removed, and high-performance DRV/TIA chips with higher bandwidth and stronger SI compensation capabilities are used. The signal regeneration and digital signal compensation functions originally implemented by oDSP are now handled by network equipment ASIC chips.

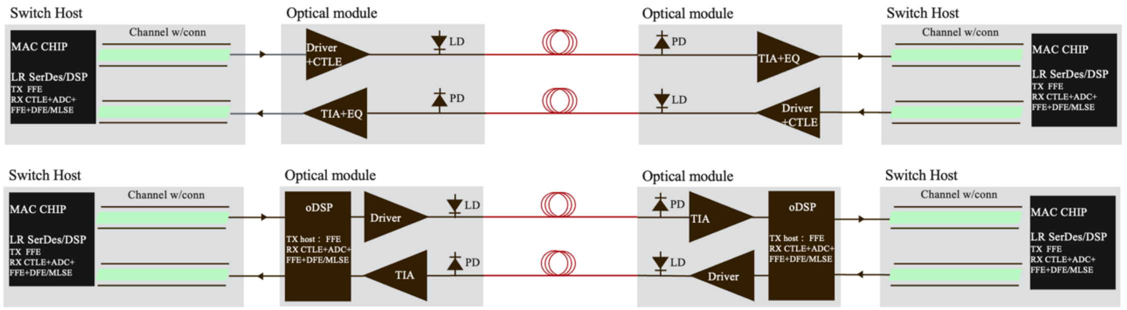


Figure 1-1: LPO module system vs DSP module system

In comparison to the oDSP solution modules, LPO modules have the following advantages:

1. Low Cost: From the perspective of the Bill of Materials (BOM) cost, it can save approximately 30%.

2. Low Latency: After removing oDSP, the data transmission latency for each direction is reduced by approximately 60-90 ns.

3. Low Power Consumption: Compared to oDSP solution modules, DR8 power consumption is reduced by approximately 50%, and SR8 reduces power consumption by about 70%.

4. Mature Ecosystem: Compared to solutions such as CPO, NPO, OBO, etc., the LPO solution maintains a pluggable module form, preserving the current mature industry ecosystem.

To address the challenges posed by the removal of oDSP in LPO modules to the system's signal integrity. This article will focus on describing how to optimize high-speed channels in system design to achieve better SI performance and enhance link performance through system tuning. The second section analyzes the theory and application challenges of LPO systems. The third section analyzes the design and implementation of LPO systems from a high-speed design perspective. The fourth section introduces the practical test results of LPO systems. The fifth section reflects on the industrial implementation of LPO systems. The sixth section provides a summary of this article on LPO systems and outlines future work plans.

2 Theoretical Foundations and Application Challenges of LPO System

2.1 Theoretical Analysis of Linear System in LPO

The ideal communication channel is regarded as an LTI (Linear and Time-Invariant system). It has linear and time-invariant characteristics.

A linear system satisfies both additivity and homogeneity:

$$\text{Additivity: if } x(t) = x_1(t) + x_2(t),\ y(t) = y_1(t) + y_2(t).$$

$$\text{Homogeneity: If } x(t) = ax_i(t),\ y(t) = ay_i(t).$$

Combining additivity and homogeneity, linearity can be expressed as:

$$\text{input } x(t) = \textstyle\sum_{k=1}^{N} a_k x_k(t),\ \text{output} y(t) = \textstyle\sum_{k=1}^{N} a_k y_k(t)$$

A time-invariant system means that if the input signal is delayed by τ, then the output exists with the same delay τ. If $y(t) = tx(t)$, satisfying $y_1(t) = tx(t - \tau)$, the system is considered time-invariant.

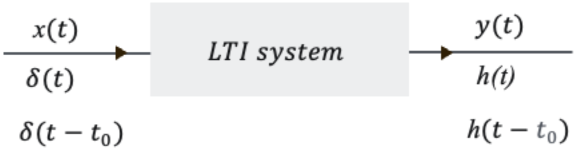


Figure 2 -1: LTI system impulse response

The impulse response of the LTI system as show in Figure 2-1 of the model. Based on its linearity and time-invariant characteristics, it can linearly decompose any input function $x(t)$,Assuming $\Delta(t-t_0)$ is a function that is nonzero only at t_0 and equals 1, then $x(t)$ is the summation of all functions $x(t_0)\Delta(t-t_0)$ for all real numbers t_0, expressed as $x(t) = \sum_{k=1}^{N} a_k x_k(t-t_k)$. If the response function of $\Delta(t)$ is denoted as $Y(t)$,then the response function $y(t)$ of $x(t)$ is the summation of all functions $x(t_0)Y(t-t_0)$,given by $y(t) = \sum_{k=1}^{N} a_k y_k(t-t_k)$.

Any input signal $x(t)$ can be represented as the infinite sum of shifted and weighted unit impulse functions $\delta(t)$: $x(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)d\tau$. Here, $\delta(t)$ is the unit impulse function satisfying $\delta(t) = \begin{cases} \infty, (t=0) \\ 0, (t \neq 0) \end{cases}$, $\int_{-\infty}^{\infty} \delta(t)\, dt = 1$. The output response of the LTI system is given by $y(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau$, where $h(t)$ is the unit impulse response.

For the frequency characterization of the LTI system, if the input is $X(s) = \sum_{-\infty}^{\infty} F_n\, e^{st}$, and the system's frequency response is $H(s)$,then the corresponding system's frequency response is $Y(s) = \sum_{-\infty}^{\infty} Y_n\, H(s)e^{st}$,where F_n is the Fourier coefficient of the input signal, and Y_n is the Fourier coefficient of the output response. Hence, $Y(s) = X(s)H(s)$ is the Fourier transform of $y(t) = h(t) * x(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)dt$ .These expressions represent the output in the frequency and time domains, respectively. The frequency domain multiplication corresponds to time domain convolution. Therefore, during channel compensation, simplifying analysis can be performed in the frequency domain, and compensation calculations can be done in the time domain.

Taking the example of a 112Gbps telecommunication channel. The channel model is shown in Figure 2-2. Here, $X(s)$ is the original transmitted signal, $H_1(s)$ is the pre-emphasis transfer function, $H_2(s)$ is the channel transfer function,$H_3(s)$ is the equalizer transfer function, and $Y(s)$ is the final received signal. If $H_1(s) * H_2(s) * H_3(s) = 1$,then $Y\ (s) = X(s)$,and the signal is perfectly restored.

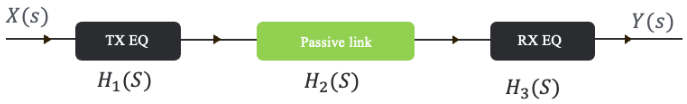
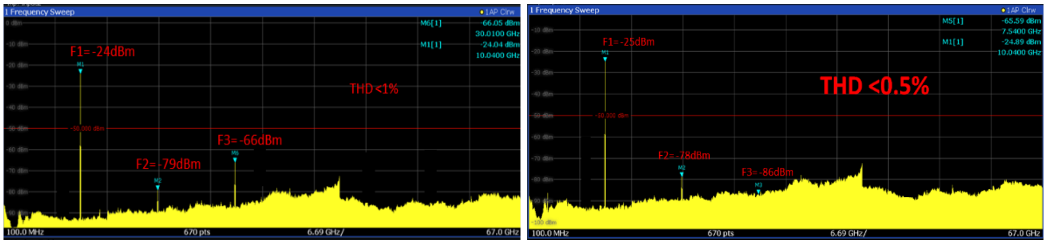


Figure 2-2: channel system model

Due to its optoelectronic conversion characteristics, the channel model of the LPO system is different significantly from the ideal telecommunication channel model. This system is mainly composed of the network equipment's main chip, passive channel, and LPO optical modules. The linear system consists of the SerDes with ADC+DSP architecture in the ASIC chip, which directly drives the optical engine to emit light through a channel with some insertion loss. At the receiver, the light signal is detected by a photodiode (PD), through TIA’s equalization and PCB transmission line reaching the ASIC SerDes Rx EQ. The signal is further reshaped and regenerated.

Under specific conditions, LPO system can be considered an LTI system. The passive channel between the SerDes and the optical module, including but not limited to PCB traces, PCB vias, connectors, footprint, solder joints, etc. The passive channel is linear (analyzed according to linear superposition theory, the introduction of reflection points doesn't change the system's linear characteristics). As for the LPO optical module after removing the oDSP, the internal key functional components mainly include analog devices such as optical and electrical chips. These devices typically operate in a linear range and have clear linearity requirements. Under suitable operating conditions, they can be considered linear. The linearity of analog chips is described by the THD (Total Harmonic Distortion). The DRV and TIA used in LPO have excellent THD characteristics, ensuring signal distortion-free amplification. The typical test results for THD of the analog chips used in LPO, including DRV and TIA, are shown in Figure 2-3, demonstrating that the linearity meet requirements for PAM4 application.

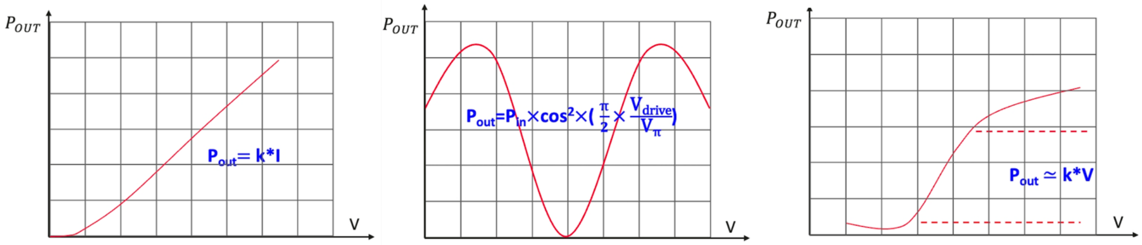


(a) Driver THD test results

(b) TIA THD test results

Figure 2-3: Typical THD Test Results for the Electrical Chips

The main optical chips widely used in LPO modules include VCSEL, MZ, and EML. Under the conditions of selecting suitable operating points, the electro-optical response curves of these chips could meet the requirements of linear modulation. As shown in Figure 2-4.



(a) VCSEL

(b) MZ

(c) EML

Figure 2-4: Electro-optical response curve of mainstream optical chips

Based on the theoretical analysis mentioned above, the signal channel of the LPO system is simplified into the model shown in Figure 2-5. Here, $X(s)$ is the original transmit signal, $H_1(s)$ is the pre-emphasis transfer function at the Transmitter, $H_2(s)$ is the transmission function of the sending channel, $H_3(s)$ is the transmission function of the receiving channel, $H_4(s)$is the pre- emphasis transfer function at the Receiver, $Y(s)$ is the final received signal, thus $Y\ (s) = \phi X(s)$, and due to the nonlinear factors in the optical link, $\phi = H_1(s) * H_2(s) * H_3(s) * H_3(s) < 1$.

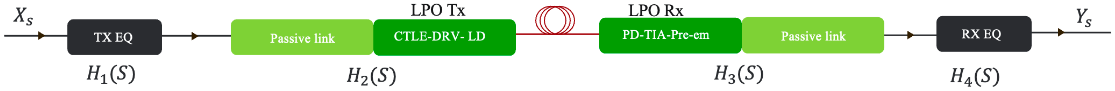


Figure 2 -5: LPO System Channel

2.2 Challenges in LPO System Application

The 112G LPO system demonstrates theoretical feasibility. And there is already a relatively enough test data to further validate it. However, to achieve commercialization and application, there are still some challenges to be overcome.

1. SI Design Challenges in LPO System

In the LPO system, with the removal of oDSP from the module, the signal system essentially becomes an LR channel. The SerDes at both transmitter and receiver need to cover the electrical and optical loss of the entire channel. Even with Driver/TIA providing equalization, the system still faces challenges such as jitter, crosstalk, and reflections that all superimpose onto the LR channel within the optoelectronic channel.

2. Eliminate Optical Nonlinear in LPO Modules

Although the design principle of the LPO system is aimed at linear operation, some nonlinear effects are inevitably generated during the electro-optical modulation process in practical applications. Additionally, nonlinear effects can occur due to optical dispersion and distortion during the transmission of optical signals. The elimination or compensation of optical nonlinearity will affect the performance of the LPO system.

3. Standardization of LPO System Interfaces

LPO optical modules, as a pluggable short-distance solution, are mainly used for intra-data center interconnection. Therefore, "interconnectivity between devices" is a crucial feature that the LPO system must support. To achieve "interconnectivity," it is necessary to standardize the optoelectronic interfaces.

4. Control of Production Metrics in LPO System

Due to the removal of oDSP in LPO optical modules, traditional testing methods for DSP modules can’t be directly applied. Therefore, how to implement quality control in the production of LPO modules will be a critical factor influencing the mass production of LPO modules.

3 Design and Implementation of the LPO System

In the HPC/DCN network system based on the CLOS architecture, LPO serves as the interconnection medium, ensuring high-speed data communication between switches and server NICs, as well as between switches. The diagram below illustrates a typical 800G HPC network. The scale of each pod's network can be adjusted based on the convergence ratio and network hierarchy.

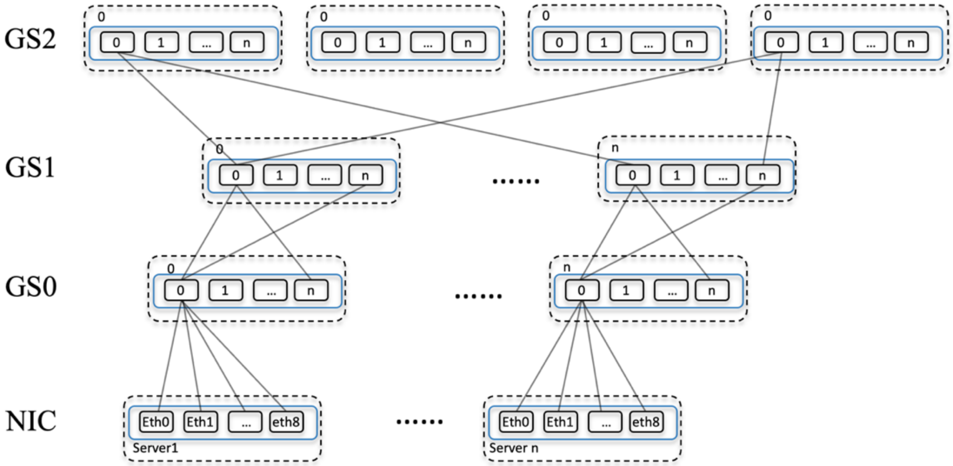


Figure 3-1: CLOS Architecture in HPC Networks

Considering the data center fiber cabling lengths' plan, for the 800G-HPC network, each level of switching equipment serves as a Core Switch, where GS0 in the DCN network is typically a TOR switch. LPO-SR modules can be used for GS0-NIC and GS0-GS1 connections, while LPO-DR modules can be applied to various levels of the network. In conjunction with the application requirements of the 800G LPO system, the following analysis focuses on the high-speed design of network equipment and modules within the system.

3.1 Signal Integrity Analysis of Network Hardware Devices

In the Leaf-Spine architecture of DCN/HPC networks, the Core Switch, as the most powerful network switching equipment with a capacity of 51.2T, a single-chip has 512 pairs 112G SerDes. The front panel provides 64 ports of 800G. It could deliver the core switching capability to the network.

Hardware architecture of the Core Switch: For pluggable module devices, various solutions can be employed, such as the card with Retimer, cable solutions, and PCB + backplane multilayer architecture. However, considering cost, reliability, and SI performance comprehensively, the most competitive solution is to use a single PCB combined with a dual-layer IO connector bell-belly scheme.

Our approach adopts a novel wiring design, utilizing an N+N stacking design, allowing the upper-layer network on the chip to traverse from the chip's lower side to the lower ports on the front panel. We control the total length of all PCB traces to be less than 9 inches. And through the reasonable PCB layout design, we could control the total PCB layer is not more than 28 layers. The PCB dielectric (Core/PP) thickness is 6 mil/6 mil. The host loss is within 7 dB for Df=0.0015@10GHz and within 9dB for Df =0.0021@10GHz.

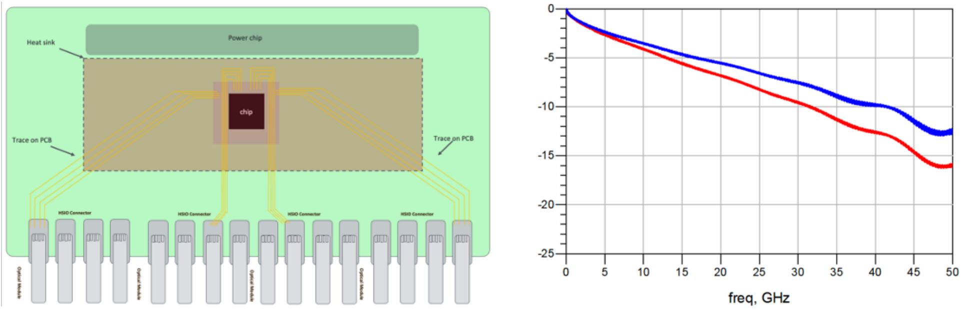


Figure 3-2: PCB Design and C2M Loss of CORE Switch Equipment

The TOR Switch is a 16T device designed to interface with 400G network cards. It has 8 upstream ports with 800G and 24 downstream ports with 400G. The PCB thickness is 3.6mm. It has 18 layers. All PCB traces length are controlled to be less than 9 inches. The PCB material’s Df is 0.0015 @ 10GHz. And the dielectric (Core/PP) thickness is 6 mils/6 mils.

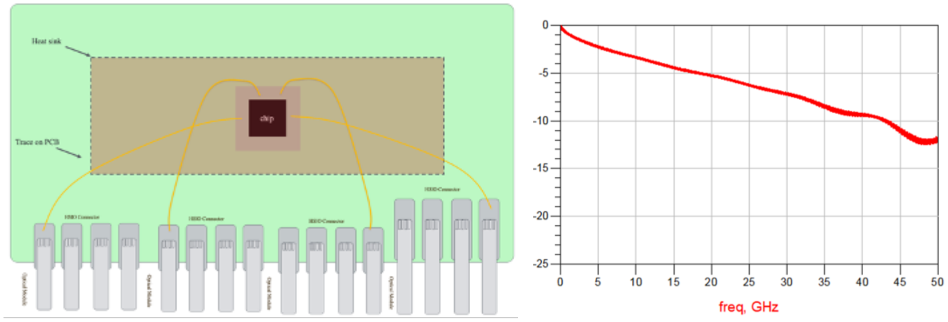


Figure 3-3:PCB Design and C2M Loss of TOR Device

For the NIC, the transmission line length is controlled to be 2.5 inches, and the PCB layer is 14 layers. For 112G data rate’s transmission lines, the dielectric (Core/PP) thickness is 3.5 mil/3 mil. The PCB material’s Df is 0.0021 @ 10GHz.

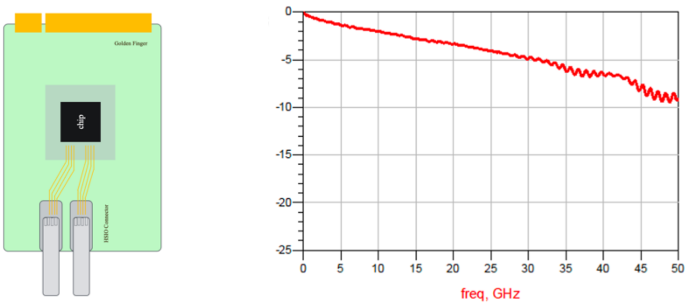


Figure 3-4: PCB Design and C2M Loss of T NIC

Chart 3-1: The simulation results of the 112 G C2M channel

Item	Material Df @10GHz	Host PCB Trace length (inch)	Channel insertion loss (dB) @26.56GHz	Channel insertion loss w/pkg (dB)@26.56GHz	ICN (mV)	VEC (dB)	ERL (dB)	EH (mV)	COM (dB)
Core Switch	0.0021	8.8	11.24	15.66	1.00	7.85	15.10	21.29	4.51
Core Switch	0.0015	8.8	9.65	13.79	1.15	8.13	14.38	23.68	4.32
TOR Switch	0.0015	8.5	9.45	13.51	1.17	7.61	14.23	25.87	4.67
NIC	0.0021	2.5	7.00	9.90	3.05	7.85	11.03	35.99	4.51

Channel insertion loss: include BGA solder ball-via -PCB trace- HSIO connector-HCB

Analysis of COM simulation results for C2M, the board card's SI margin is sufficient.

3.2 Prototype Testing Validation.

In the early stages of the project, we developed a prototype supporting 112G SerDes, as shown in Figure 3-8. The device includes various types of system channels, and this chapter discusses C2M links with different loss. The feasibility of the LPO system is tested and validated in conjunction with the 400G-DR4 LPO module. The switch chip utilizes SerDes DSP architecture, supporting a 40dB die-die loss capability.

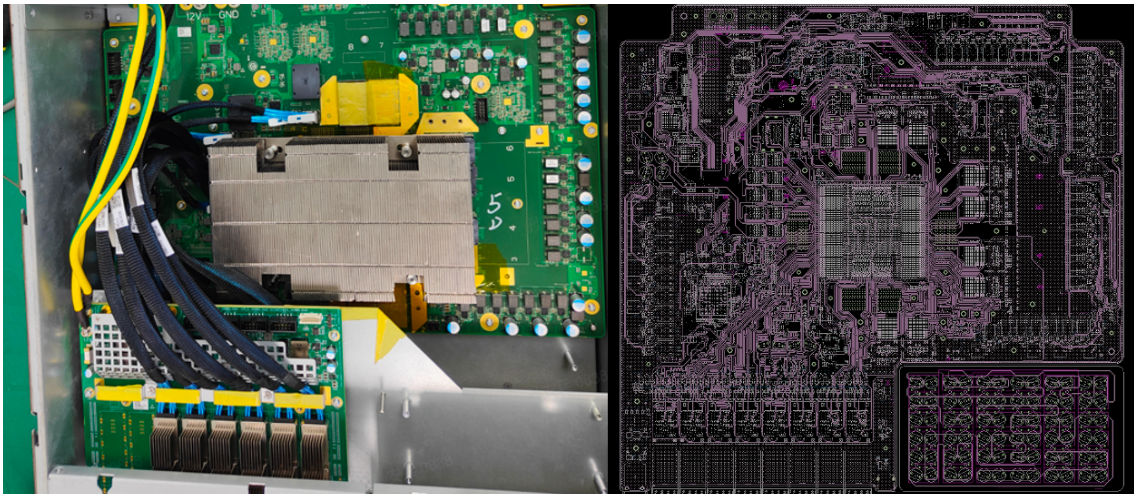


Figure 3-5:112G SerDes System prototype verification machine

Based on this prototype, we will optimize the TXRQ, analyze the channel BER levels under different loss, and compare the performance of the LPO system between two architectures: a pure PCB solution and an in-board cable solution.

The LPO system is an optoelectronic hybrid channel that includes an electrical signal channel, driver, TIA, and optical device. In comparison to passive electrical signal channels, the LPO channel exhibits the characteristic of active gain. Moreover, the parameter configuration of the driver and TIA in the optical module has a significant impact on the link performance.

For the LPO module, due to the lack of system-level equipment in the early stages of module development, manufacturers typically use a Bit Error Rate Tester (BERT) combined with an evaluation board to set up a testing environment. The performance is then improved by adjusting the parameters of the driver and TIA. Parameters adjusted based on the evaluation board may not match the actual system link. Therefore, it is common in subsequent stages to integrate with the switch system for system-level optimization. This involves jointly adjusting the TXEQ parameters at the transmit end of the switch system and the parameters of the LPO module's driver and TIA to achieve optimal link performance.

It is important to note that all tests in this article are conducted based on parameters that have been optimized for the LPO module during the testing phase.

1. TX FIR Parameter Optimization Test: Combining the typical loss chain of the actual channel switch host's TX/RX, adjusting the TXEQ parameters, and validating the debugging results on the system. Based on the actual design scheme, the Host loss can be controlled within 7dB. In this section, the test is conducted with Host TX + Package/Host RX + Package Insertion loss (IL)both set at 12dB. TX EQ is adjusted from 5dB to 11dB, and the test results are as figure 3-6.

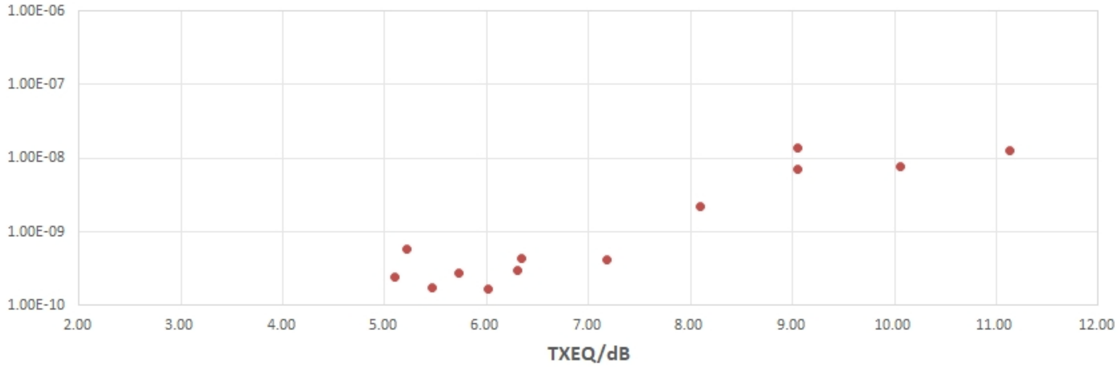


Figure 3-6: TX Host + Package /RX Host + Package IL 12dB EQ debugging test

Based on the test results, under different link loss conditions, TXEQ significantly affects the Bit Error Rate (BER) of the LPO system according to different system links. The following tests in this chapter will use the optimized EQ parameters for relevant testing.

2. CASE2: With the RX Host + Package IL fixed at 12dB, the TX Host + Package IL is increased from 8dB to 17dB. For each link, the TXEQ parameters are optimized based on the BER results. The system's BER changes as the TX loss varies are as figure 3-7.

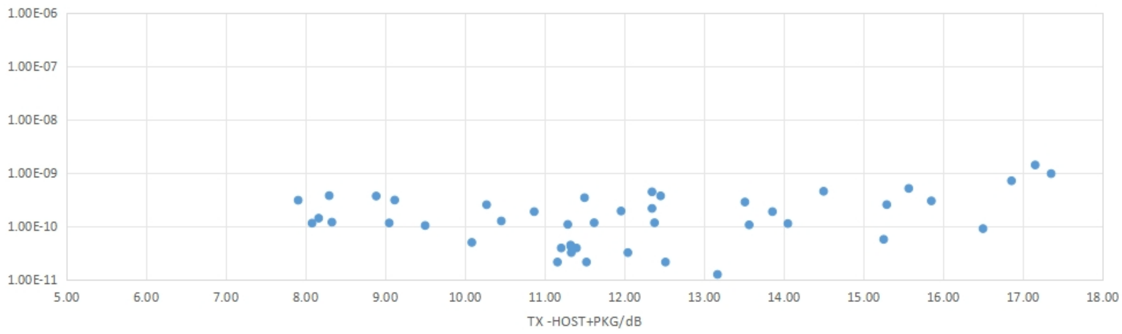


Figure 3-7: TX host + pkg IL 8-17dB RX host + PKG IL 12dB   BER Testing

From the test data, it is evident that the change in BER is not highly correlated with the variation in TX loss ranging from 8 to 17dB. Overall, even when the channel loss reaches 17dB, the system's BER can be controlled at the level of E-9, indicating sufficient margin in the system.

3. With TX loss fixed, RX loss varies, and the TXEQ parameters are optimized based on BER results for each link. For this scenario, TX loss is set for a short link at 10dB and a long link at 17dB. RX loss varies in the range of 7.5dB to 14.5dB, and as RX loss changes, the system's BER varies as follows.

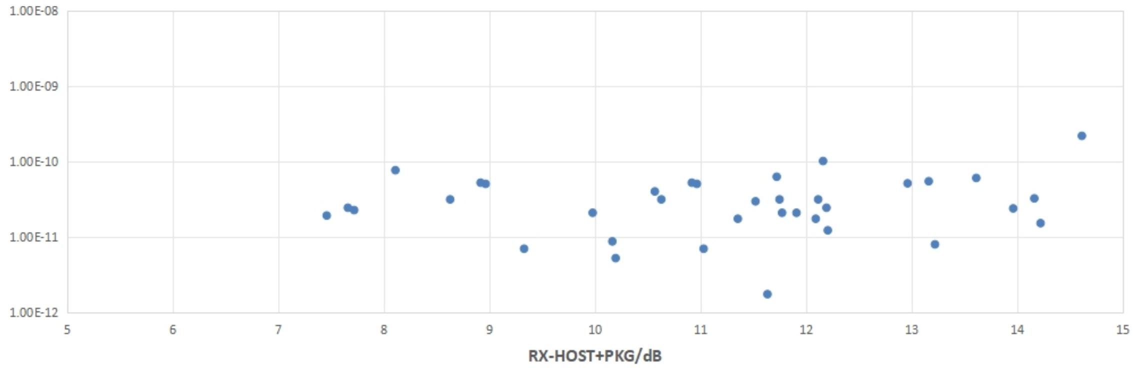


Figure 3-8: TX host + pkg IL 10dB RX host + PKG IL 7.5-14.5dB   BER Testing

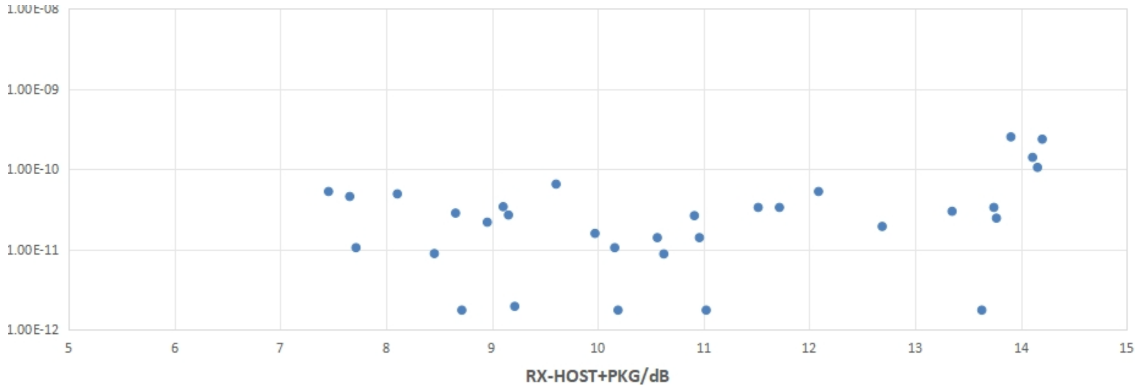


Figure 3-9: TX host + pkg IL 17dB RX host + PKG IL 7.5-14.5dB   BER Testing

From the test data, it is evident that when RX loss varies from 7.5dB to 14.5dB, the distribution of BER for both the TX short link and the TX long link remains stable. Overall, the BER for the entire system can be controlled at levels below E-9, indicating sufficient margin in the system.

4. In the hardware system architecture design of the switch, the cable solution is often mentioned for its low-loss characteristics. We conducted comparative LPO performance tests on this prototype under two different architectures, and in both cases, the TX/RX host loss was measured at 7dB.

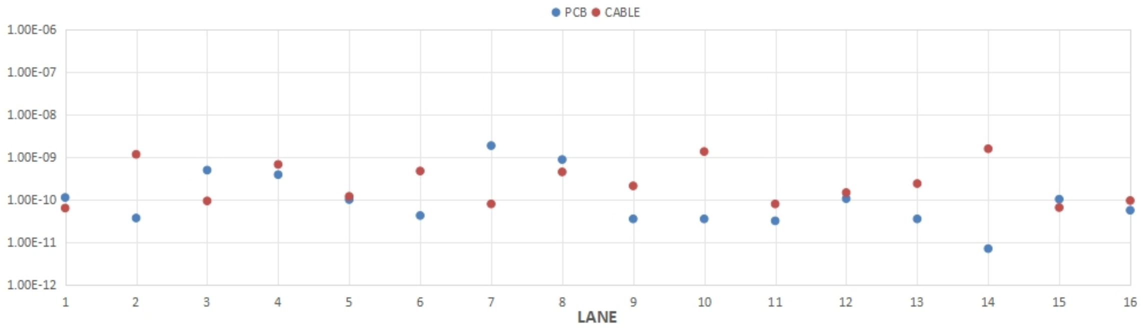


Figure 3-10: PCB/Cable Solution LPO System BER Testing

Through the test data, it can be observed that the overall BER (Bit Error Rate) of the LPO system for both architectures varies between E-8 and E-11, showing no significant difference. The error rates of the two methods are essentially consistent.

Analyzing the test results from the prototype mentioned in this article.

1. With the optimization of TXEQ (transmitter equalization) and the equalizer parameters of the module Driver/TIA (Transimpedance Amplifier), based on the capabilities of the SerDes demo machine, BER (Bit Error Rate) tests were conducted on channels with different loss lengths. Under the conditions of a VSR channel, the LPO system demonstrates good operational performance.

2. Within the range of satisfying the OIF Linear Optical specifications for loss, there is no significant difference between the on-board cable scheme and the pure PCB scheme. Both can effectively support the LPO system.

3. Considering system margin and the variation in SerDes capabilities among different manufacturers, we strongly recommend adhering to more stringent metrics when designing the Host Signal Integrity performance of the equipment. Reference can be made to Section 3.1, while also meeting the requirements outlined in IEEE 802.3 CK Annex 162A for CR (Continuous Rate) links, CEI-112G-LINEAR-PAM4 specifications, and OIF-CEI-5.0-VSR-PAM4 specifications.

4 LPO System Testing

4.1 Test Environment and Experimental Setup

To thoroughly assess whether the core switch has the capability to fully support LPO optical modules on all ports, this paper establishes a test environment as shown in Figure 4-1. Using ByteDance 51.2T core switch. Each port's host PCB loss less than 7dB.

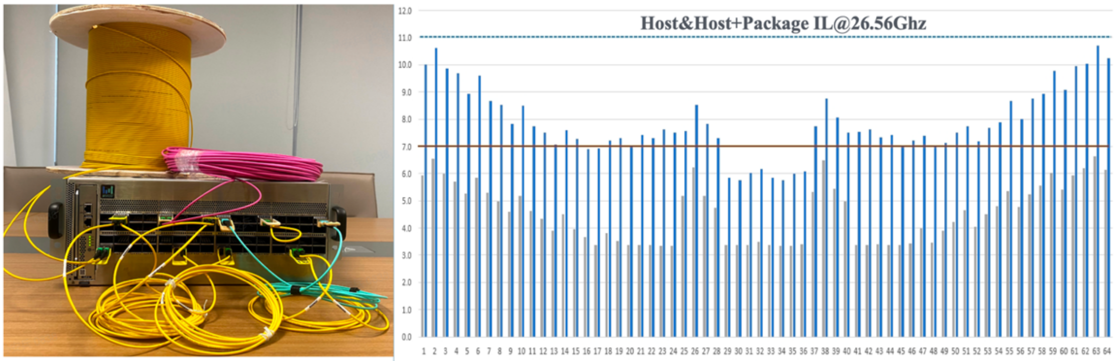


Figure 4-1: LPO System BER Testing Environment

Key materials:

Switch: One 51.2T 64-800G port core switch

Fiber: 2 pieces of 50m OM4 MPO12 MM, 4 pieces of 3m OM4 MPO12 MM, 2 pieces of 500m MPO12 SM fiber, 4 pieces of 5m MPO12 SM, and 4 pieces of 5m MPO16 SM

LPO modules: 8 pieces of 800G 2xVR4 (Vendor A), 8 pieces of 800G 2xDR4 (Vendor A & B)

Evaluation Metrics

BER: The Bit Error Ratio is the ratio of the number of erroneous bits to the total transmitted bits during a specific time interval. In this context, it specifically refers to the pre-FEC error rate. In communication systems, the BER at the receiver may be influenced by channel noise, interference, distortion, bit synchronization issues, attenuation, wireless multipath fading, etc.

Symbol Error Distribution: It indicates the number of consecutive errors in a codeword and is used to evaluate whether the error distribution in the link is within the correction capability of FEC (Forward Error Correction). For PAM4 links, the KP4 RS (544,514) FEC is commonly used, which can correct up to 15 consecutive erroneous bits. Therefore, Symbol Error Distribution is used to assess the quality of the link.

Experimental Setup

1. Experiment 1a: Use 8 sets of 3m short fibers to connect 4 pairs of 800G 2xDR4 LPO modules from 2 vendors. Fixed the LPO module's Driver/TIA parameters. And set the

switch under the PRBS test mode. Optimize the TXEQ parameters on each port, and test the BER and symbol error for all 64 ports.

2. Experiment 1b: Connect 2 sets of 800G 2xDR4 LPO modules from vendor A using 2 pieces of 500m MPO12 fibers. Maintain the port TXEQ parameters and LPO module parameters from Experiment 1a and test the BER and Symbol error for all 64 ports.

3. Experiment 2: Adjust the single-mode LPO Driver/TIA parameter settings from Experiment 1 and traverse all 64 ports to evaluate the impact of different module configurations on link performance.

4. Experiment 3: Keep the port TXEQ parameters of Experiment 1a unchanged. Connect 4 pairs of 800G 2xVR4 LPO modules with 3m/50m OM4 fibers and test the BER and Symbol error for all 64 ports.

5. Experiment 4: Based on Experiment 3 with 50m OM4 fibers, optimize the port TXEQ parameters and 800G 2xVR4 LPO Driver/TIA parameter settings separately for all 64 ports and test the BER and Symbol error.

4.2 Single-Mode LPO Module Testing

Following the settings of experiments 1a and 1b, the data obtained by traversing all ports of two different vendors' 800G 2xDR4 LPO optical modules are shown in Figures 4-2 and 4-3.

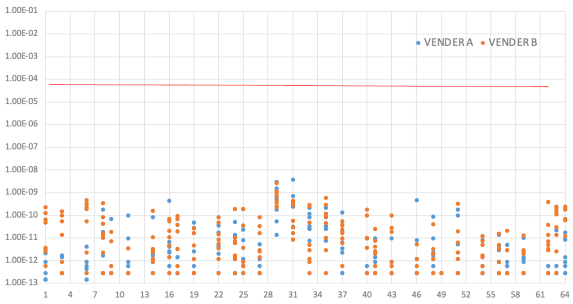


Figure 4-2: Testing Data of 2xDR4 Vendor A/B

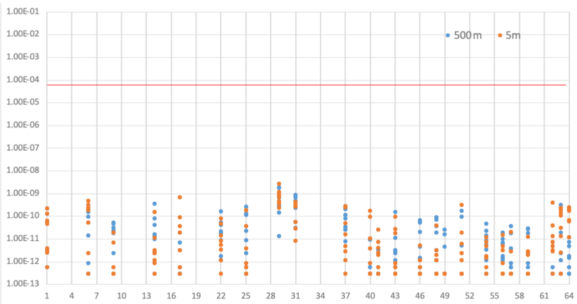


Figure 4-3: Testing Data of 2xDR4 500m vs. 5m

Figure 4-2 shows the test data for the 800G 2xDR4 modules from Vendor A and B, respectively, across 64 ports of the switch with a 5m short fiber interconnect. Under relatively optimal TXEQ parameter configurations for the switch, the BER for both Vendor A and Vendor B falls between 1E-8 and 1E-13, significantly lower than the IEEE 802.3 standard requirement of 2E-4. This link margin is sufficient. Comparing ports with shorter traces, such as 29-31, under the same settings, the BER is 1 to 2 orders of magnitude lower than ports with relatively longer traces. Therefore, the optimal performance cannot be achieved on all switch ports with the same set of LPO parameters.

In Figure 4-3, the testing data for Vendor A modules are presented under scenarios with 500m and 5m fiber interconnects across 64 ports. A comparison shows no significant difference in BER between the 5m and 500m fiber scenarios, both ranging from 1E-9 to

1E-13. Thus, it can be inferred that fiber distance has almost no impact on the single-mode LPO link.

As shown in Table 4-1, 5-10 ports of different insertion loss physical ports are selected to test the FEC margin of the link under an 800G 2xDR4 configuration with a 500m interconnect. Looking at the Symbol Error distribution test data for lane 0 of typical ports, it can be observed that errors occurring on the link fall within one random error, indicating sufficient FEC RS (544,514) margin.

Table 4-1 Testing Results for Symbol Error Distribution on Typical Ports of 800G 2xDR4

PORT	P KG + host IL	BER(lane0)	Symbol error
1	10dB	9.17E-11	s1/s16
6	9.5dB	1.02E-11	s1/s16
9	7.8dB	2.05E-12	s1/s16
14	7.5dB	2.92E-13	s1/s16
17	6.9dB	8.77E-13	s1/s16
22	7.2dB	1.17E-12	s1/s16
25	7.5dB	2.31E-11	s1/s16
30	5.7dB	3.01E-11	s1/s16
32	5.8dB	2.60E-11	s1/s16

From the data in Figures 4-1 and 4-2, under the same set of Driver/TIA parameters, the performance of ports with shorter traces is worse than those with longer traces. To optimize the performance of ports with shorter traces, Experiment 3 adjusted the Driver/TIA parameters, and the test data is presented in Figure 4-3. Examining the test data for ports 29-31, it can be observed that the BER with the optimized Driver/TIA parameters is two orders of magnitude lower than the original parameters, all below 5E-10. Therefore, it can be concluded that tuning the LPO module parameters based on trace loss can achieve better link performance. It could provide configuration guidance for LPO applications.

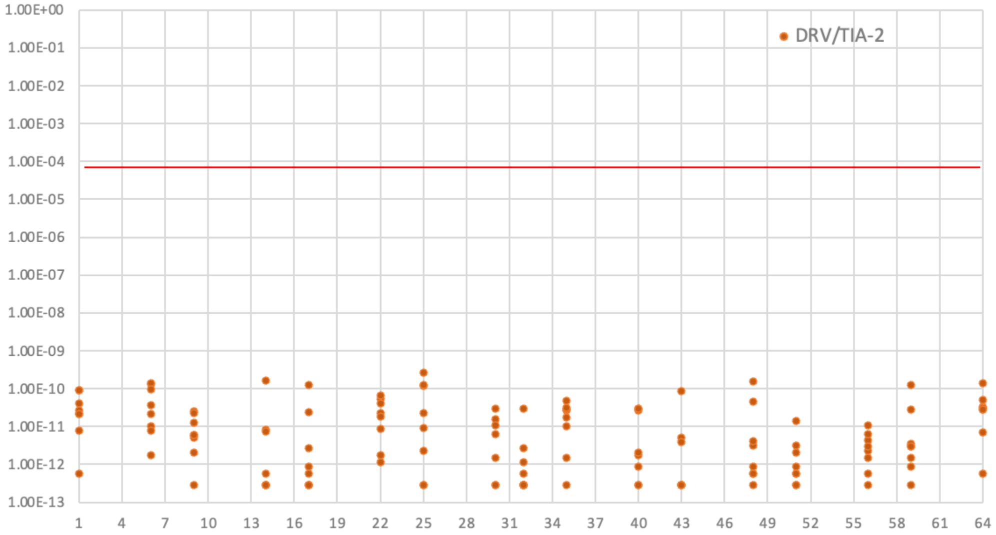
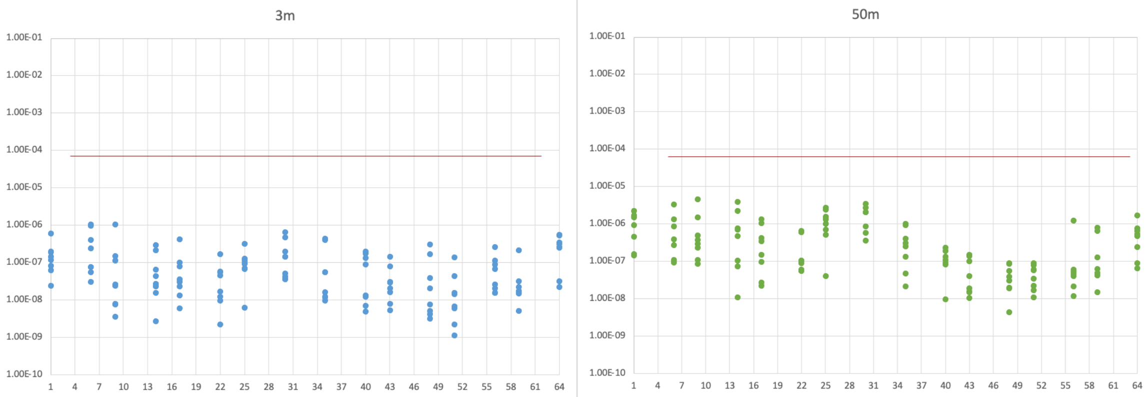


Figure 4-4 Testing Data for Different Configurations of Driver/TIA Parameters

4.3 Multimode LPO Optical Module Testing

Following the settings of experiments 3 and 4, tests were conducted on Vendor A's 800G 2xVR4 LPO module under normal temperature conditions. The testing involved traversing through all ports with 3m length and 50m length fiber interconnections.



(a) 3m length fiber testing data

(b) 50m length fiber testing data

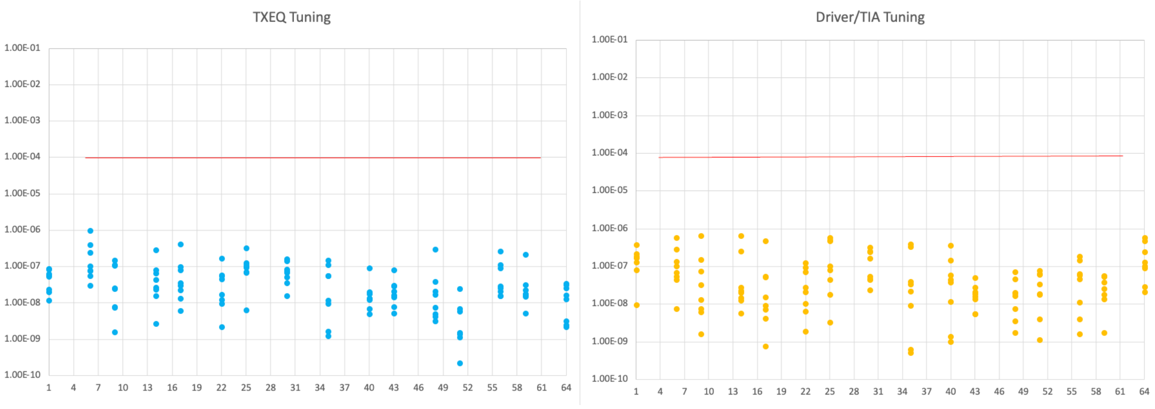
Figure 4-5: Comparison of Testing Results for 800G 2xVR4 with Different Fiber Lengths

Experiment 3 maintained the port TXEQ parameters unchanged from experiments 1 and 2 to evaluate the link performance of the 800G 2xVR4 module under equivalent parameters. As shown in Figure 4-5 (a), in the scenario of a 3m length short fiber interconnection, the BER level of the multimode LPO module ranged from 1E-6 to 1E-8, although lower than the IEEE 802.3 standard requirement of 2E-4, the BER margin of the link was relatively small. Figure 4-5 (b) shows that in the scenario of a 50m length long fiber interconnection, the BER level of the multimode LPO module ranged from 5E-6 to 1E-8, decreasing by 0.5 to 1 order compared to the short fiber interconnection data. At the same time, considering the Symbol Error Distribution of lane0 for typical ports in Table 4-2, the continuous errors on the link ranged from 2 to 5.

Table 4-2 Symbol Error Distribution Test Results for Typical Ports of 800G 2xVR4

PORT	P KG + host IL	BER(lane0)	Symbol error
1	10dB	6.47E-8	s4/s16
6	9.5dB	1.82E-7	s3/s16
9	7.8dB	2.45E-7	s3/s16
14	7.5dB	2.92E-8	s3/s16
17	6.9dB	8.97E-8	s2/s16
22	7.2dB	3.17E-7	s3/s16
25	7.5dB	4.31E-7	s3/s16
30	5.7dB	4.01E-6	s4/s16
33	5.8dB	1.80E-6	s5/s16

In Experiment 4, the TXEQ parameters of the ports and the Driver/TIA parameters of the 800G 2xVR4 LPO module were optimized separately. As shown in Figure 4-6, the optimized BER averaged between 1E-6 and 1E-9. Compared to Figure 4-5, there is no significant change observed. For the optimization of parameters in the 800G 2xVR4 LPO module, the improvement in link performance is limited.



(a) Optimization TXEQ of switch Ports (b) Optimization the parameters of Driver/TIA

Figure 4-6: Test Data After Optimization of Parameters for 800G 2xVR4

5 Industrialization Analysis of LPO

In considering and researching port standardization, we found that the EECQ indicator of Tp1a is more advantageous in establishing a connection with TDECQ of the Tp2 eye diagram to guide LPO link debugging.

We studied and tested the correlation between EECQ and TDECQ, as shown in Figure 5-1. The horizontal axis represents four different TXEQ configurations. From the comparative data, good EECQ results in better TDECQ. Under the second set of TXEQ parameter configurations, the link BER performance of measured is better. However, in this TXEQ parameter setting, the link BER does not show a strong correlation with EECQ. Further exploration is needed for the optical-electrical interface standards of LPO systems.

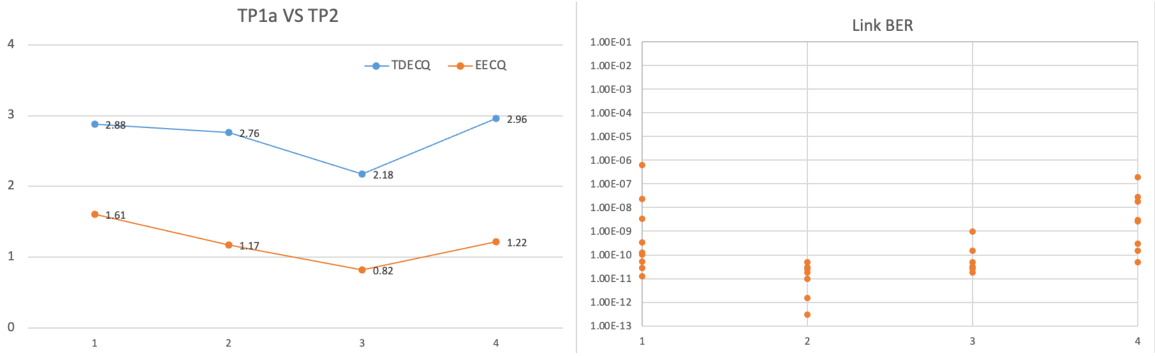


Figure 5-1: Comparison of Correlation between TP1a and TP2 Metrics

Regarding the current production criteria for the LPO system, the considerations are as follows:

1. On the module production line, using the switch directly as a production testing instrument, measure directly the module performance using link BER and symbol error distribution, and guide module production by calibrating a set of reasonable thresholds.

2. For the equipment, it is also recommended to use link BER and symbol error distribution to calibrate the switch port's TXEQ.

3. To ensure optimal link performance, during the R&D phase, tuning the Driver/TIA parameters based on the channel loss range and store them in different modes (such as Long Reach Mode, Short Reach Mode, and Normal Reach Mode) in the module's App Code register area. After the module is inserted into the switch system, network equipment selects the optimal module parameters based on port cable's loss and actively configures the module.

6 Conclusion

This article describes the advantages and design challenges of the LPO system from the perspectives of theoretical foundations, system simulation, and signal integrity. It also presents the system test results of the ByteDance 51.2T core switch combined with the LPO module.

From the viewpoint of comprehensive system design and implementation, this 64*800G network switch with an innovative SI scheme, ensuring sufficient system margin in the switch's SI design, and all ports support the application of LPO modules.

Considering the comprehensive test data, the overall performance of the single-mode LPO module is much higher than that of the multi-mode LPO module. The single-mode LPO module has the potential for mass production and can be applied to production networks first. The multi-mode LPO module, due to its optical non-linear characteristics, has a smaller overall system margin.

In the comprehensive testing process, optimizing the TXEQ parameters of the switch ports and the Driver/TIA parameters of the LPO module has revealed the following insights.

1. Adjusting the Driver/TIA parameters of the LPO module in a targeted manner, considering the equipment-side loss range, leads to improved link performance.

2. The optimized TXEQ parameters for single-mode LPO modules are not entirely applicable to multi-mode LPO modules. Moreover, adjusting multi-mode parameters requires consideration of optical characteristics, posing higher debugging difficulties.

Considering the advantages of the LPO module in power consumption, cost, delay, and maintainability, combine with the practical test results, it can be concluded that the 800G linear direct drive system is achievable.

Prospects for the 224G SerDes High-Speed System： oDSP optical modules, LPO, and CPO are all optional solutions, but each of them face their own challenges. For the oDSP module, it will bring greater power consumption and cost challenges. Regarding CPO, there is still no effective solution to its reliability and maintainability issues. For the LPO module, the main challenges lie in device performance and design. Silicon photonic device bandwidth needs further improvement, and thin-film lithium niobate devices need to achieve low-cost mass production. The capability of 224G SerDes is still to be verified. However, with the optimization of switch link design and advancements in PCB materials, connectors, cables, chip technology, and other aspects, overall channel insertion loss and reflection will be further optimized. In terms of technology, LPO may potentially extend to 224G.

References

[1] Dayong Shen et al. “High-speed system architecture design of DCN core switch design”, Design Conference 2023.
[2] OIF CEI-112G-Linear-PAM4-Draft
[3] IEEE Std 802.3ck™-2022
[4] BP Lathi, Signal Processing and Linear Systems