System-Level Design & Characterization of a 5 Tbps Router

Austin Rothschild and Tejus Rao

Stanford University, Department of Electrical Engineering

June 8, 2025

Contents

1 Introduction

In this project, we describe the design of a signaling system that implements a 5 Tbps router fabric composed of eight line-cards and four crossbar cards. Each crossbar card supports a full-duplex bandwidth of 160 Gb/s from each line card. The line-cards each support a total of 640 Gb/s duplex bandwidth to the four crossbar boards. Both the line-cards and crossbar cards contain an identical switch chip transceiver, which has 128 duplex differential pairs in the SERDES interface. Additionally, each line-card has dedicated IO from the switch chip to interface with 16 different QSFP optical modules, each offering 10 Gb/s x 4 duplex optical fiber communications to the line-card front panel. Each crossbar also communicates with two external CPU modules with a 1 Gb/s link used for control and status information.

This report outlines the design process for the router. First, we discuss the system-level design including the major cost and performance tradeoffs, system architecture diagrams, placement and routing study, and connector and PCB material choices. Next, we describe the signaling scheme used including the driver, equalization techniques, and clock and timing methodology. Then the PCB stackup and crosstalk are discussed before showing thorough design characterization via SPICE simulations of crosstalk, equalization, and link pulse response. Eye diagrams are measured and noise and jitter budgets are shown along with calculations of the allowable noise to meet a BER of 10^{-14}

2 System Design Description

Given the power, price, performance, and area constraints imposed by the design rules, we arrived at our final router architecture after considering the various tradeoffs involved with each option. As such, we elected to focus our efforts on *cost reduction* while operating with moderate power, robust noise and timing budgets, and minimal area overhead.

Opting for the higher signaling speed option, 10.7 Gb/s 64/66 coded, enables 10 Gb/s throughput per SERDES differential pair. Therefore, the required link bandwidth from each board was achieved with only one switch chip per board (line-card and crossbar). The cost saved by only using one switch chip per board far outweighs the cost incurred by using a higher power signaling rate. Furthermore, our system does not require any cabling. All connections are made with Amphenol XCede+ orthogonal midplane connectors.

In addition to the eight line cards and four crossbar cards, a fully passive and structural midplane board offers orthogonal connectivity between the line and crossbar cards and signal fanout for low-speed CPU and control routing.

2.1 Cost Summary

Cost Summary (WIP)						
PCB Costs	***# Layers***	***Tan δ***	***# Cards***	***Panel Cost***	***Cards/Panel***	***Total Cost****
Line Card	14	0.006	8	$249	1	$1,992.00
XBar Card	14	0.006	4	$249	3	$249.00
Midplane	8	0.006	1	$193	1	$193.00
Package Costs	**Type**	**# Balls**	**Pkg Cost**	**# Pkgs/card**	**Card Cost**	**Total Cost**
Line Card	BGA	2116	$300	1	$300.00	$2,400.00
XBar Card	BGA	2116	$300	1	$300.00	$1,200.00
Connector Costs	**Type**	**# Signals**	**Piece Cost**	**# / Card**	**Card Cost**	**Total Cost**
Line Card	4x8 Orthogonal Midplane & Receptacle	32	$11.20	4	$44.80	$358.40
XBar Card	4x8 Orthogonal Midplane & Receptacle	32	$11.20	8	$89.60	$358.40
XBar Card <--> CPU (low speed)	3x6 Orthogonal Midplane & Receptacle	18	$6.30	4	$25.20	$100.80
Link Power Costs	**I (mA)**	**Vdd (V)**	**Power/Link (W)**	**# Links**	**Pwr Cost**	**Total Cost**
1.3 V Rail (Link pwer)	234.375	1.3	0.305	1536	$20.0/W	$9,360.00
Other Power Costs	**I (A)**	**Vdd (V)**	**Power (W)**	**# Chips**	**Pwr Cost**	**Total Cost**
1.0 V Rail	60	1	60	12	$20.0/W	$14,400.00
3.3 V Rail	0.3	3.3	0.99	12	$20.0/W	$237.60
Total System Cost						$30,849.20

Figure 1: System cost summary. Note that one of the PCB panels shares one line card and one crossbar card, for a total of ten manufacturing panels.

2.2 Block Diagrams

Here we provide a high level overview of the fabric architecture. The final design uses a single switch chip on each crossbar board and line card, making full use of the transceiver's IO capabilities in order to meet the required interface bandwidths. To enable this, the chip is operated in the 10.7 Gb/s 64/66 coded throughput mode. Note that due to the forward error correction overhead, the effective throughput is 10 Gb/s. This signaling architecture fully occupies each switch chips IO capability, minimizing the need for extra chips and hence saving cost, power, and area.

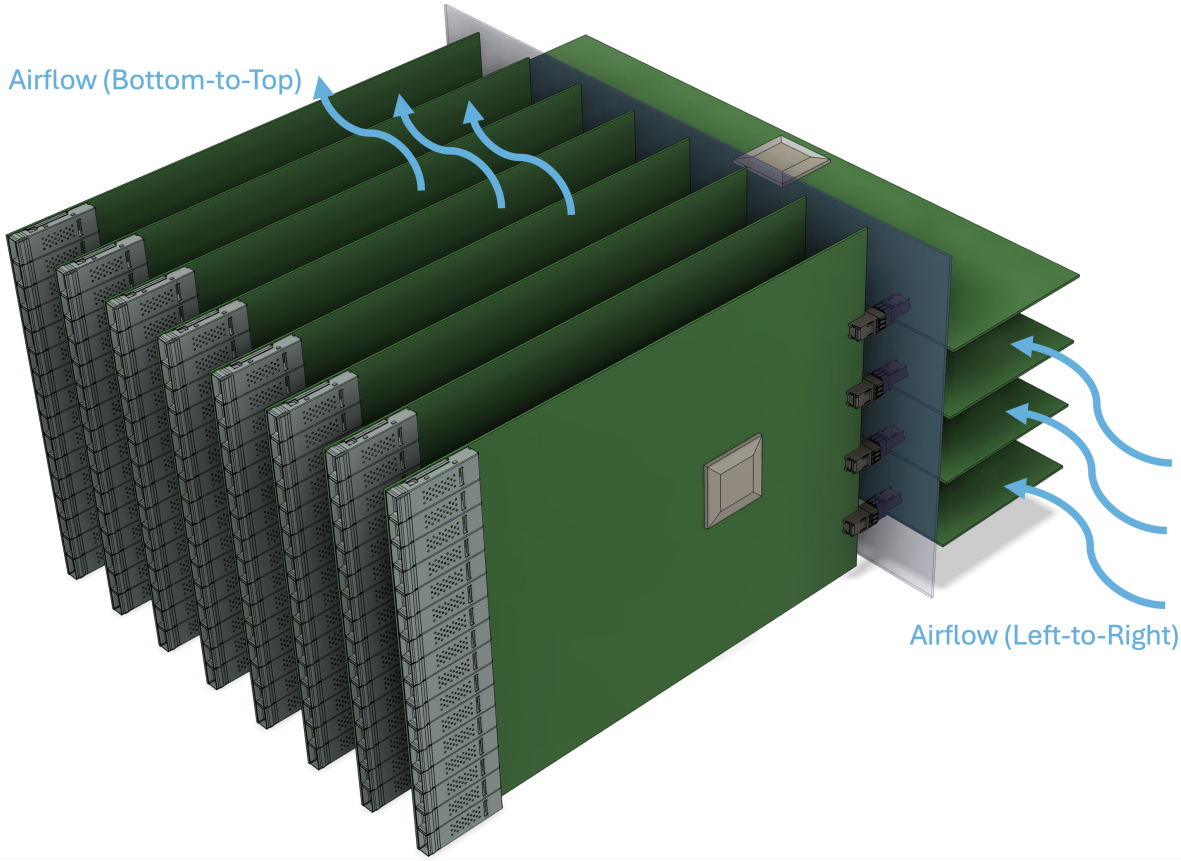


Figure 2: A 3D model of the system. The blue board is the midplane, made translucent for display purposes connectors shown. Airflow is indicated by the blue arrows. The eight line cards are vertical, and the four crossbar cards are horizontal. (Midplane connectors are not to scale.)

On the line card, a 640 Gb/s throughput requirement means that 64 10 Gb/s duplex links are allocated from the line card chip to each crossbar interface. The remaining 64 links are used to interface with the QSFP optical modules, with 4 TX/RX pairs allocated to each module in order to support 40 Gb/s throughput per QSFP transceiver slot. The linecard board architecture is shown in Fig. 3. Each bus drawn on the diagram represents two overlapping stripline layers of routing. The bus width is calculated by a 4-4-4-28 mil spacing convention for the differential pairs. A total of 64 duplex differential pairs are used for the links to the optical modules and another 64 duplex differential pairs are used for the links to the crossbar cards. The longest trace length for the line-card switch chip to midplane link is approximately 5.25 inches and the shortest link is approximately 3.25 inches.

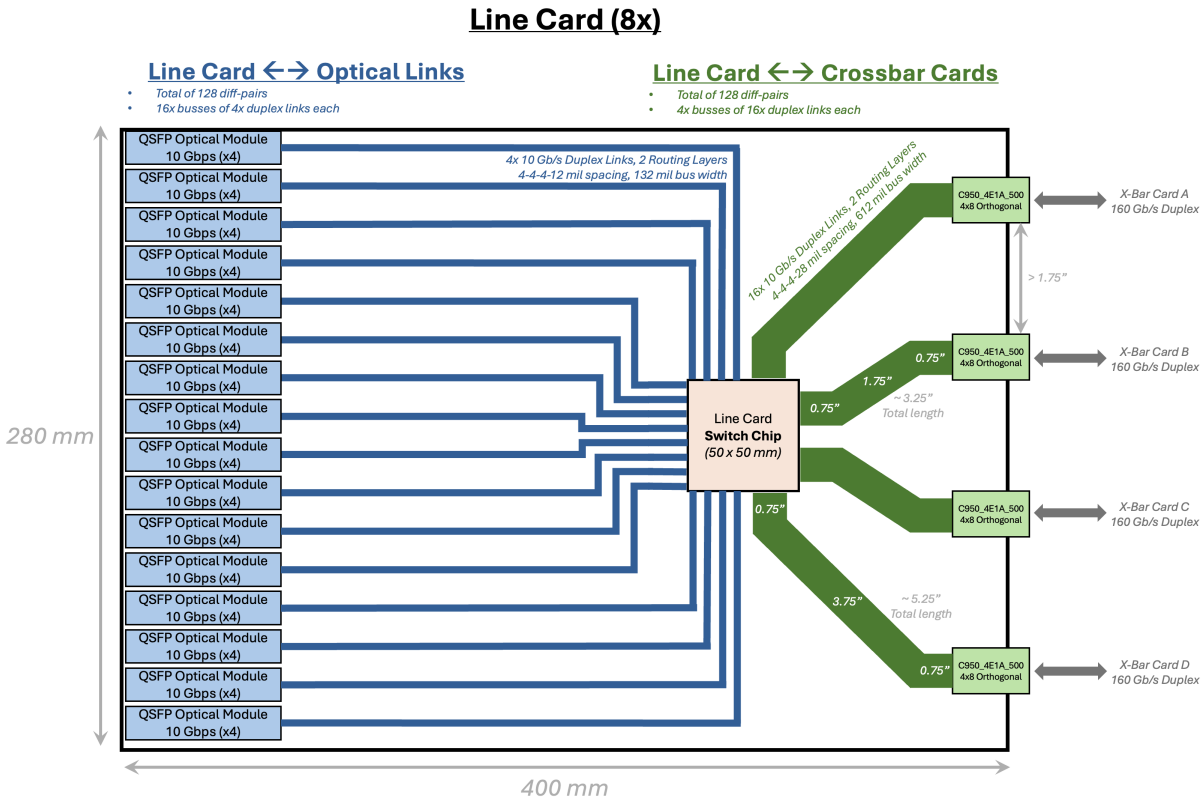


Figure 3: Linecard block diagram. Longest link: 5.25". Shortest link: 3.25".

Each crossbar switch chip contains 128 TX/RX duplex pairs. Thus, 16 pairs are allocated per line-card interface in order to meet the 160 Gb/s bandwidth requirement. This allows for the use of a single switch chip to meet the crossbar IO needs. The crossbar board architecture is shown in Fig. 4. Again, each bus drawn on the diagram represents two overlapping stripline layers of routing. The longest link between the midplane and the crossbar card switch chip is approximately 10.6" and the shortest link is approximately 1.5 inches.

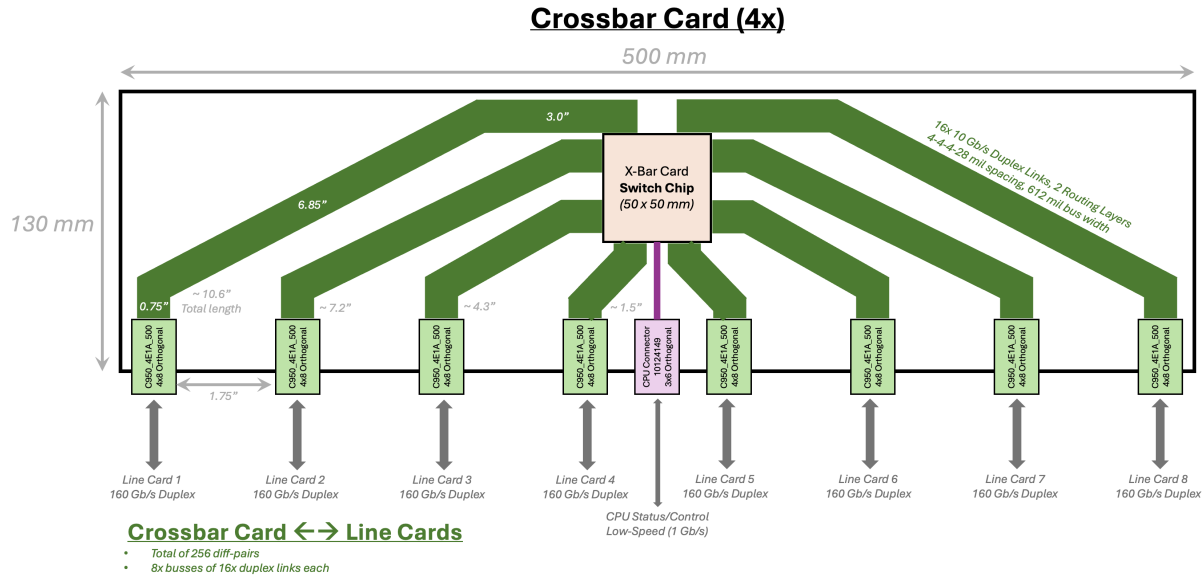


Figure 4: Crossbar card block diagram. Longest link: 10.6". Shortest link: 1.5".

The two cards are mated via an orthogonal midplane board which has press-fit connectors, discussed further in the next section. The longest total path from switch chip to switch chip is approximately 16.5", and the shortest total path is approximately 5.25".

2.3 Placement Drawing

The crossbar and line cards are interfaced using an orthogonal mid-plane pass-through system. This architecture was selected based on a variety of factors after analyzing the relative cost, signal integrity, compactness, and thermal performance of all the different solutions. With the use of an orthogonal mid-plane, the line cards and crossbars can directly interface to pass-through connectors on the mid-lane, eliminating any extra routing or stubs that would be introduced by a planar back-plane solution. Additionally, this architecture removes the need for extra cables/harnesses to interface the line-card and cross-bar boards. This improves the signaling channel by removing the loss incurred by the cables' frequency-dependent attenuation, which scales with length. Additionally, cost, area, and complexity are optimized when avoiding cable interfaces and opting for the plug-and-play based connectivity approach.

The Amphenol XCede+ orthogonal mid-plane connector is the implemented solution on the final design. This connector has 32 differential pairs, with a differential characteristic impedance of 97 Ω. Each line-card to crossbar interface has a total of 16 duplex links (16 TX, 16 RX) can be supported per connector. This allows for one connector to support the entire interface throughput requirement between one line card and one crossbar card. The crossbar and line-card connectors are both XCede+ 950-4E1A-B3 connectors, having a female adapter pinout as observed in Fig. 6b. The midplane contains the XCede+ 951-419-0/1-90D male sockets for the crossbar and line-card connectors, with the pinout displayed in Fig. 7a. These connectors press-fit into the midplane from both sides of the board, sharing through-holes to avoid any stubs or trace routing on the midplane board. The line-card connector is configured using a 90° degree rotation with respect to the cross-bar connector. Additionally, the pass-through footprint with shadow vias is used in order to maintain interface impedance control and minimize adjacent link crosstalk.

2.4 Signal/Connector Mapping

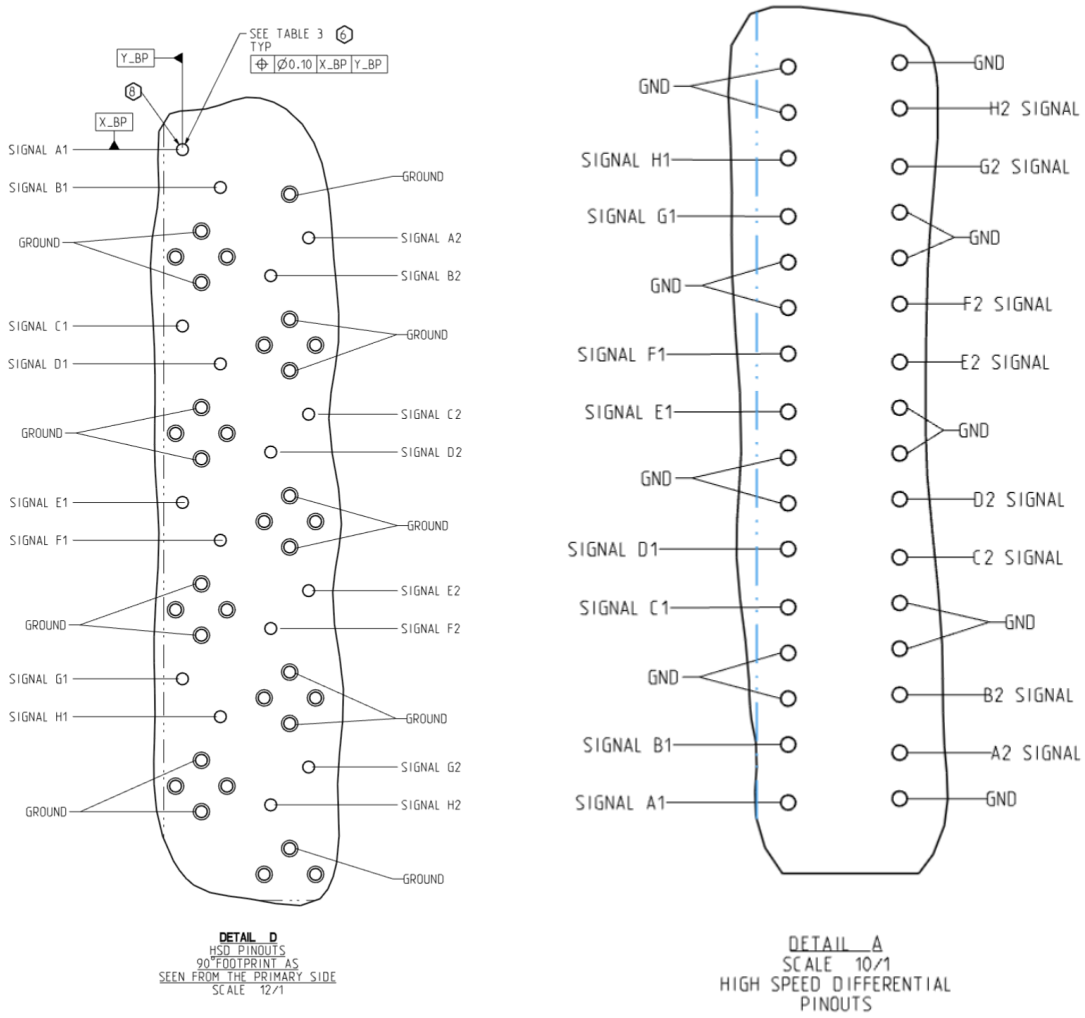
The following signal net-name assignments are adopted in the connector pinout mapping:

$$\mathbf{LC}[0:15]\_\mathbf{TX}[0:15]\_\mathbf{P}, \quad \mathbf{LC}[0:15]\_\mathbf{TX}[0:15]\_\mathbf{N}$$
$$\mathbf{LC}[0:15]\_\mathbf{RX}[0:15]\_\mathbf{P}, \quad \mathbf{LC}[0:15]\_\mathbf{RX}[0:15]\_\mathbf{N}$$

The line-card (LC) is considered as the "main", while the crossbar is considered as the "peripheral". All signals are prefixed by "LC" whether they originate from the line-card or not. The subsequent "TX" and "RX" identifiers indicate whether the signal is an input or output from the line-card. An input to the line-card is an output of the crossbar card, and vice versa. The pair number is indicated by the numbering identifier from $[0:15]$. Lastly, the polarity is denoted by "P" and "N".

PASS THROUGH CONFIGURATION 90°								
PIN MAPPING SIDE 1-SIDE 2	SIDE1-SIDE2							
	A1-A1	A2-B2	A3-C1	A4-D2	A5-E1	A6-F2	A7-G1	A8-H2
	B1-B1	B2-A2	B3-D1	B4-C2	B5-F1	B6-E2	B7-H1	B8-G2
	C1-A3	C2-B4	C3-C3	C4-D4	C5-E3	C6-F4	C7-G3	C8-H4
	D1-B3	D2-A4	D3-D3	D4-C4	D5-F3	D6-E4	D7-H3	D8-G4
	E1-A5	E2-B6	E3-C5	E4-D6	E5-E5	E6-F6	E7-G5	E8-H6
	F1-B5	F2-A6	F3-D5	F4-C6	F5-F5	F6-E6	F7-H5	F8-G6
	G1-A7	G2-B8	G3-C7	G4-D8	G5-E7	G6-F8	G7-G7	G8-H8
	H1-B7	H2-A8	H3-D7	H4-C8	H5-F7	H6-E8	H7-H7	H8-G8

Figure 5: Amphenol XCede+ orthogonal mid-plane 90° passthrough pinout mapping.



(a) Midplane male connector.

(b) Linecard/Crossbar female connector.

Figure 6: First two columns in the orthogonal midplane connector interface for the line-card and crossbar signals.

Line Card								
	1	2	3	4	5	6	7	8
A	LC_TX_00_P	LC_TX_01_P	LC_TX_02_P	LC_TX_03_P	LC_TX_04_P	LC_TX_05_P	LC_TX_06_P	LC_TX_07_P
B	LC_TX_00_N	LC_TX_01_N	LC_TX_02_N	LC_TX_03_N	LC_TX_04_N	LC_TX_05_N	LC_TX_06_N	LC_TX_07_N
C	LC_TX_08_P	LC_TX_09_P	LC_TX_10_P	LC_TX_11_P	LC_TX_12_P	LC_TX_13_P	LC_TX_14_P	LC_TX_15_P
D	LC_TX_08_N	LC_TX_09_N	LC_TX_10_N	LC_TX_11_N	LC_TX_12_N	LC_TX_13_N	LC_TX_14_N	LC_TX_15_N
E	LC_RX_00_P	LC_RX_01_P	LC_RX_02_P	LC_RX_03_P	LC_RX_04_P	LC_RX_05_P	LC_RX_06_P	LC_RX_07_P
F	LC_RX_00_N	LC_RX_01_N	LC_RX_02_N	LC_RX_03_N	LC_RX_04_N	LC_RX_05_N	LC_RX_06_N	LC_RX_07_N
G	LC_RX_08_P	LC_RX_09_P	LC_RX_10_P	LC_RX_11_P	LC_RX_12_P	LC_RX_13_P	LC_RX_14_P	LC_RX_15_P
H	LC_RX_08_N	LC_RX_09_N	LC_RX_10_N	LC_RX_11_N	LC_RX_12_N	LC_RX_13_N	LC_RX_14_N	LC_RX_15_N

(a) Line-card connector, duplex signal mapping.

Crossbar Card								
	1	2	3	4	5	6	7	8
A	LC_TX_00_P	LC_TX_01_N	LC_TX_08_P	LC_TX_09_N	LC_RX_00_P	LC_RX_01_N	LC_RX_08_P	LC_RX_09_N
B	LC_TX_00_N	LC_TX_01_P	LC_TX_08_N	LC_TX_09_P	LC_RX_00_N	LC_RX_01_P	LC_RX_08_N	LC_RX_09_P
C	LC_TX_02_P	LC_TX_03_N	LC_TX_10_P	LC_TX_11_N	LC_RX_02_P	LC_RX_03_N	LC_RX_10_P	LC_RX_11_N
D	LC_TX_02_N	LC_TX_03_P	LC_TX_10_N	LC_TX_11_P	LC_RX_02_N	LC_RX_03_P	LC_RX_10_N	LC_RX_11_P
E	LC_TX_04_P	LC_TX_05_N	LC_TX_12_P	LC_TX_13_N	LC_RX_04_P	LC_RX_05_N	LC_RX_12_P	LC_RX_13_N
F	LC_TX_04_N	LC_TX_05_P	LC_TX_12_N	LC_TX_13_P	LC_RX_04_N	LC_RX_05_P	LC_RX_12_N	LC_RX_13_P
G	LC_TX_06_P	LC_TX_07_N	LC_TX_14_P	LC_TX_15_N	LC_RX_06_P	LC_RX_07_N	LC_RX_14_P	LC_RX_15_N
H	LC_TX_06_N	LC_TX_07_P	LC_TX_14_N	LC_TX_15_P	LC_RX_06_N	LC_RX_07_P	LC_RX_14_N	LC_RX_15_P

(b) Crossbar connector, duplex signal mapping.

Figure 7: Connector pin mappings based on the 90° orthogonal midplane interface design.

2.5 Routing Study

The switch chips on each line-card and crossbar board have the following ball-out landing pattern:

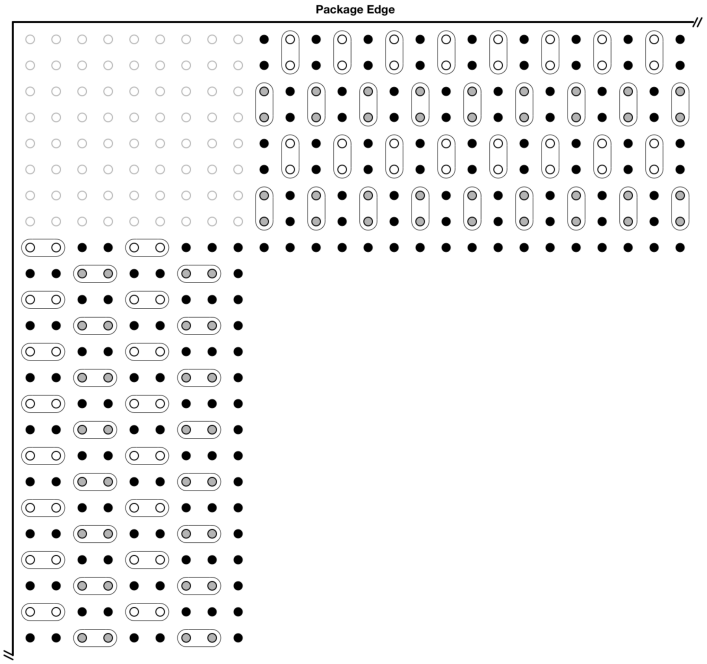


Figure 8: Switch chip BGA land-pattern.

Additionally, based on the desired duplex bus links on both the linecard and cross bar card, a switch chip escape pattern is designed such that TX and RX signals arrive/leave from BGA pins on different stripline layers.

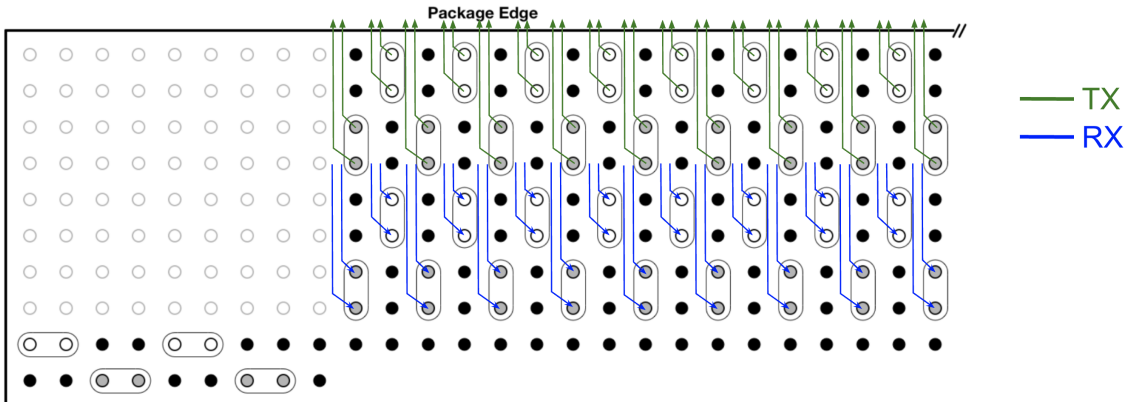


Figure 9: TX and RX duplex escape pattern.

The escape pattern for the duplex signal fanout pattern is shown below in Fig. 4.2. Via drill diameters are specified as 10 mil both for signal and GND vias. [1]. A standard via-to-via pitch of 40 mils results in 30 mils of separation between between the edges of each drilled vias. The drill-to-metal clearance is 12 mils, thus the allowable routing space is 30 mil - (2*9 mil) = 12 mils. Using a 4-4-4 trace separation, the fanout allows for two traces to fanout from each via pair. Adopting this approach, we fit all TX signals on one stripline layer, and all RX on another as observed in Fig. 9.

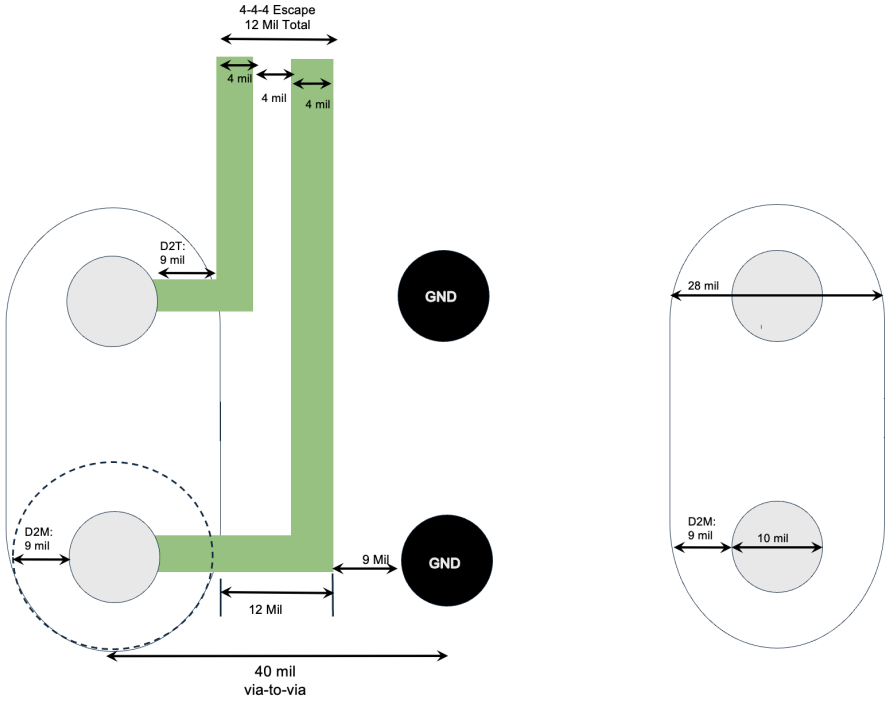


Figure 10: Switch chip TX/RX BGA escape pattern.

Additionally, the worst-case pair-to-pair spacing at the BGA escape can be observed to be 28 mils given the drill-to-trace (D2T) and via diameter between adjacent links. This spacing is the number used for stripline crosstalk analysis in Sect. 6.1.

[1] A single via size is used throughout the board to minimize PCB production cost

The XCede+ Orthogonal 4x8 connector's escape routing plan is shown in Fig. 11. This escape plan confirms that only two stripline routing layers will be necessary on both the crossbar and line card boards. This routing diagram is not perfectly to scale, but it is enough to verify that no additional routing layers are needed to fan out its signals.

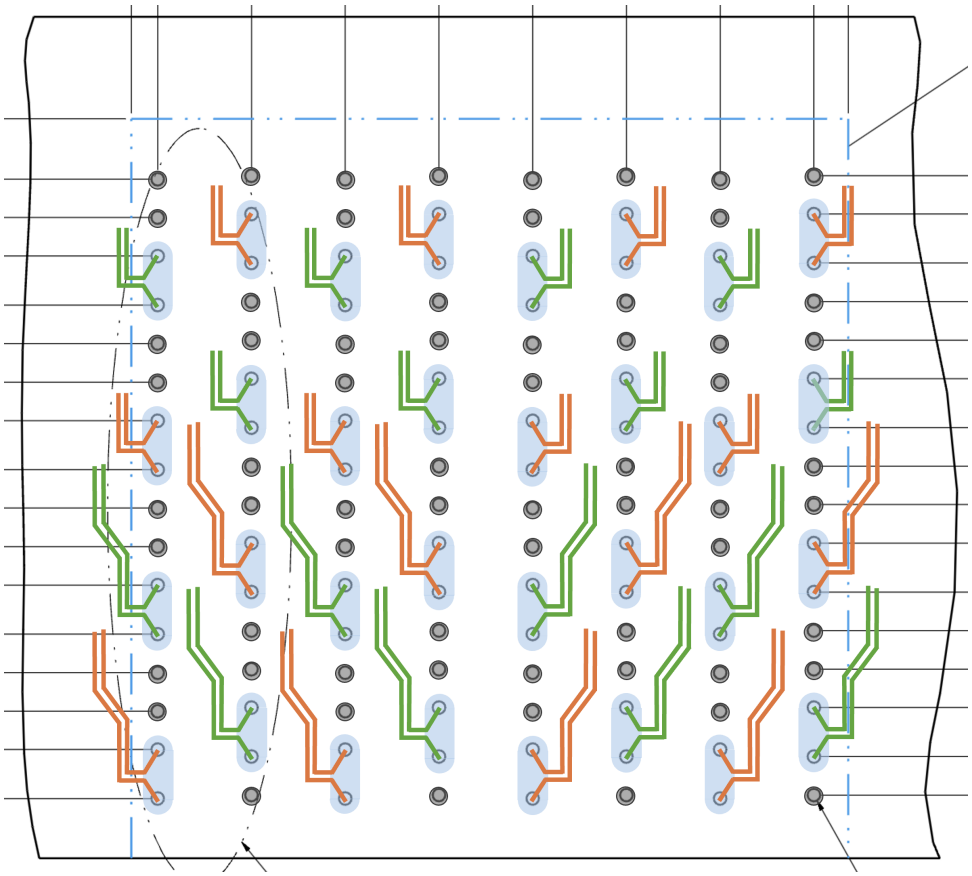


Figure 11: Amphenol XCede orthogonal connector escape (applicable to line and crossbar cards).

All differential links are spaced as follows: 4-mil trace, 4-mil gap, and 28-mil spacing between differential pairs. This was initially chosen to reduce bus width, but we understand that loss could be improved with larger traces. However, a good eye diagram was achieved with straightforward equalization techniques, so we did not optimize trace geometry beyond this.

2.6 Design Discussion

Sections 2.1-2.5 provide a high level overview of the router's architecture, while section 3 discusses link specification/operation. Sections 4 and 5 qualify the signaling performance and display the achieved equalized channel responses and allowable random voltage/jitter in the system.

3 Signaling Description

3.1 Link Block Diagram

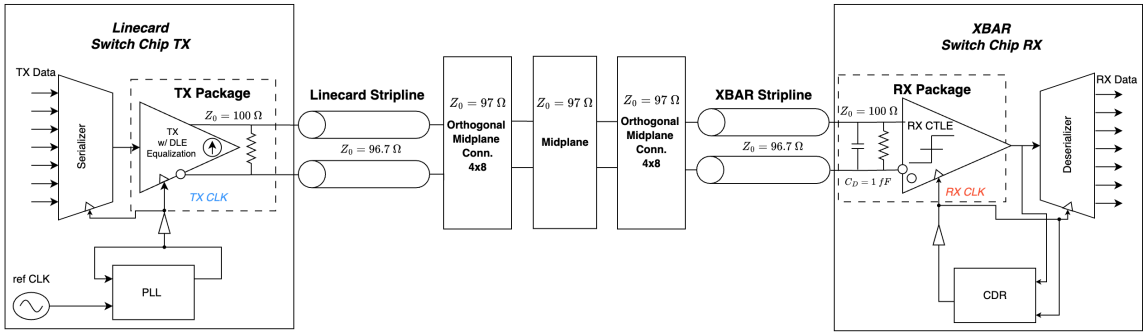


Figure 12: Signaling system: TX/RX link block diagram.

3.2 Signaling Convention

3.2.1 Current Mode Logic Driver

As displayed in Fig. 12, the transmitter utilizes current-mode logic (CML) based signaling approach. The CML driver utilizes a Norton-equivalent parallel termination of $Z_T = 2 * Z_{oo} = 100\ \Omega$. CML drivers have high output impedance, allowing for more precise impedance control set by Z_T.

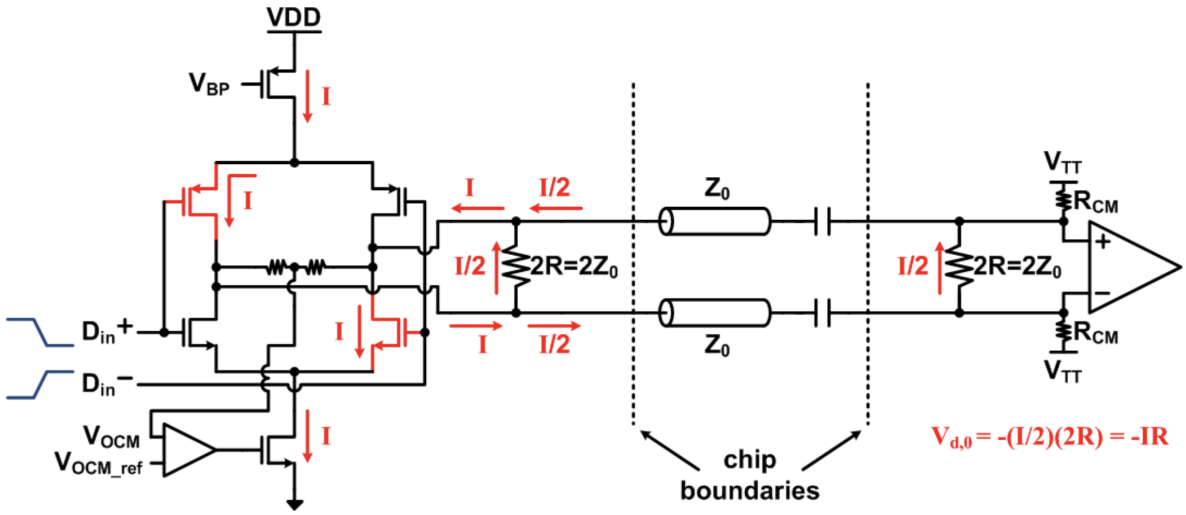


Figure 13: Push-pull CML TX Driver [1].

In particular, a push-pull CML driver is utilized due to their near constant tail current, resulting in low dI/dt noise. Additionally, differential signaling is used in order to mitigate common-mode noise that may be induced by other noise sources on the chip coming from separate digital logic, power supply noise, or other EMI coupling to signal lines. Differential signaling also allows for twice the noise margin as a single-ended signal, allowing for potentially lower voltage swings to be used, enabling power savings.

3.2.2 TX: Discrete Linear Equalizer (DLE)

At the transmitter side a 4-tap discrete linear equalizer (DLE) is used to pre-shape the TX waveforms in order to combat channel attenuation and intersymbol interference (ISI). In particular, each transmitted symbol is of the form:

$$y[n] = c_{-1} \cdot x[n+1] + c_0 \cdot x[n] + c_1 \cdot x[n-1] + c_2 \cdot x[n-2] \tag{1}$$

whereby c_{-1} is the pre-cursor, c_0 is the cursor, and c_1/c_2 are the post-cursors. The 4-tap DLE circuit model is shown in Fig. 14 in which a half-rate, interleaving architecture is used. The DLE taps are sized such that the maximum relative weights of each are 0.25, 1.0, 0.5, and 0.25.

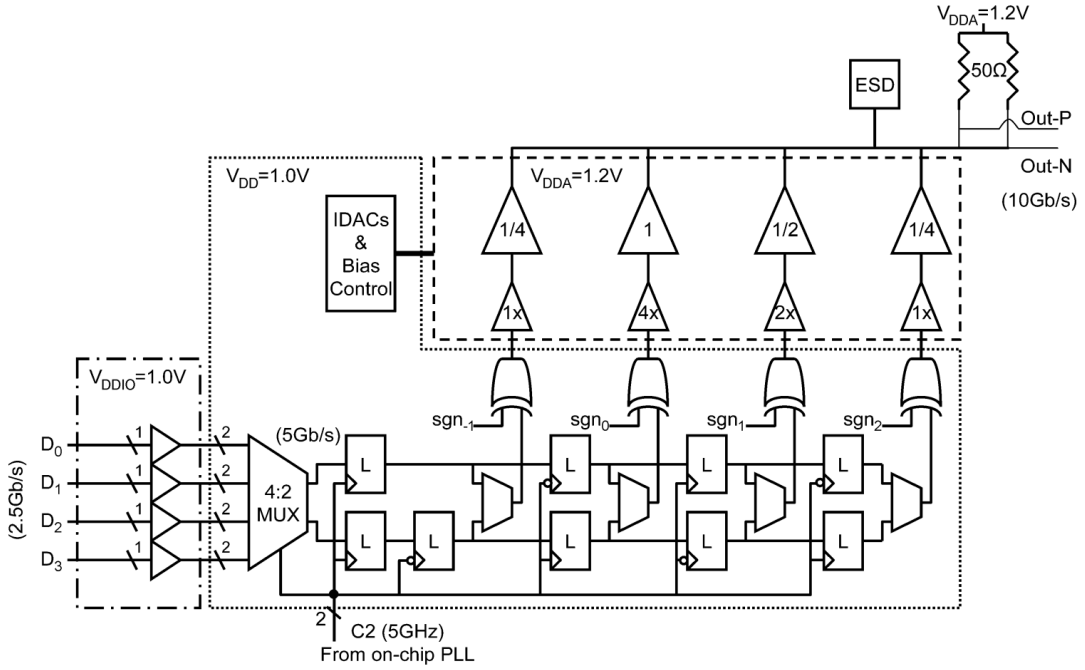


Figure 14: Transmitter with 4-tap DLE [2]

.

To implement the TX DLE, the zero-forcing (ZF) technique is used whereby a TX FIR filter is constructed such that the coefficients result in removal of ISI [3]. This scheme is simple and linear, however, it also results in noise amplification and signal attenuation which can be problematic under certain channel conditions. The ZF system for a 4-tap DLE equalizer can be expressed as:

$$\mathbf{y}_{\text{target}} = X\mathbf{c} \tag{2}$$

For the 4-tap DLE model, this has the following matrix formulation.

$$\begin{bmatrix} y_{-1} \\ y_0 \\ y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_0 & x_{-1} & 0 & 0 & 0 \\ x_1 & x_0 & x_{-1} & 0 & 0 \\ x_2 & x_1 & x_0 & x_{-1} & 0 \\ 0 & x_2 & x_1 & x_0 & x_{-1} \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \\ c_2 \end{bmatrix}$$

The vector $\mathbf{y}_{\text{target}}$ sets the location of the target response, and for this specific setup is defined as $\mathbf{y}_{\text{target}} = [0, 1, 0, 0]^T$ in order to retain the primary cursor and set the pre/post-cursors to zero. Thus, the DLE becomes a simple matrix inversion in order to extract the tap coefficients:

$$\mathbf{c} = \mathbf{x}^{-1}\mathbf{y}_{\text{target}} \tag{3}$$

Normalizing these coefficients to satisfy a fixed transmit power constraint leads to:

$$\bar{c}_i = \frac{c_i}{\sum_n |c_i|} \tag{4}$$

Eq. (2)-(4) are used in order to compute the DLE coefficients during the TX equalization procedure described in Sect. 6.3.

3.2.3 RX: Continuous Time Linear Equalizer (CTLE)

In addition to the driver equalization discussed in the previous section, this system uses a continuous time linear equalizer (CTLE) at the receiver. In essence, the CTLE provides an inverse frequency response to the channel, equalizing loss and distortion caused by the low-pass nature of the transmission lines and connectors on the PCB. The CTLE used in our system is characterized by single zero and dual pole transfer function as follows.

$$F(s) \;=\; \frac{1 + s\,\tau_z}{\left(1 + s\,\tau_{p_1}\right)\left(1 + s\,\tau_{p_2}\right)} \tag{5}$$

The peaking gain can be analyzed in the frequency domain via the ratio of the first pole and the zero.In our simulations, this transfer function is implemented by setting various resistances in a behavioral amplifier modeled by a dependent voltage source in HSPICE. The receiver model also captures termination impedance in resistance and capacitance due to ESD protection diodes.

3.3 Timing Conventions

There are two broad approaches to clocking between transmitters and receivers in high-speed links. An embedded clock system uses clock-and-data recovery (CDR) units within the receiver to extract a clock signal from the data links and synchronize the receiver's samples with this recovered clock. A forwarded clock system employs an additional link between each transmitter and receiver that provides a dedicated clock signal. At the receiver end, a deskew block is required to align the clock timing with the data. In order to save on system cost by reducing the number of links (and therefore connectors), we choose to use the former methodology, with CDR units in the receiver. This has been proven to work at frequencies above 10 Gb/s [4].

Each switch chip in our system is locally clocked, meaning that an on-board reference crystal and on-chip PLL provide global clock signals throughout a chip. As such, our overall system is plesiochronous — there are small frequency offsets between each chip's clock which cause slowly drifting phase offsets between a given transmitter and receiver. The CDR units for our system must not only correct for phase offsets, but also for frequency mismatch. Many types of architectures have been designed to fulfill these needs, including dual loop architectures that use both a PLL and DLL, multi-phase CDRs, injection-locked CDRs, and various phase interpolation methods.

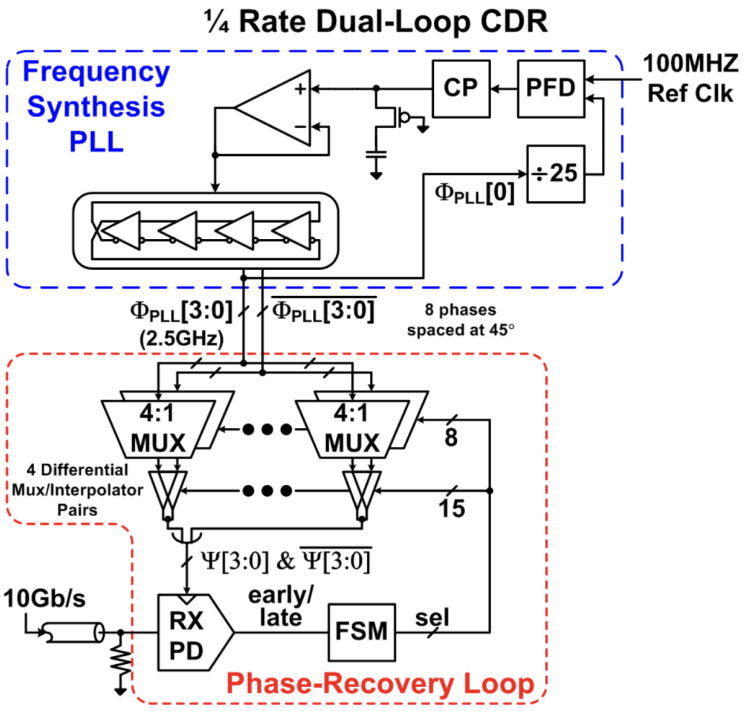


Figure 15: Simplified block diagram of a dual-loop phase interpolator CDR.

We will briefly discuss a phase interpolator (PI) based CDR here, shown in in Fig. 15, which is one candidate CDR that would work in our system [5],[4]. A main PLL generates the clock signal based on the reference crystal or another global clock reference. A ring-oscillator-based VCO is tapped at multiple nodes to produce multiple clock phases. Then, a phase interpolator loop selects between and mixes the clock phases to give fine control over the clock phase. An early-late detector and finite state machine form a feedback loop with the phase interpolation multiplexers to precisely control the phase of the recovered sampling clock. This architecture can accommodate some frequency mismatch between transmitter and receivers.

4 Interconnect Description

4.1 PCB Stack-ups

The linecard and crossbar PCBs are both manufactured using 14 layers with mid-loss grade DS-7409D (X) dielectric core and pre-preg layers. A mid-loss grade dielectric was chosen in order to minimize cost-overheard after a satisfactory channel attenuation was verified in simulation, allowing for the design to reach the target BER levels of 10^{-14}. The final stackup is shown in Fig. 16. A secondary stackup from an alternative dielectric vendor is shown as well, allowing for a multi-sourced design. Note that all design calculations and simulations are carried out using the primary stackup parameters. Both stackups are designed such that the total thickness is relatively constant between the two, allowing for a common mechanical design between the multi-sourced PCB design. Additionally, the impedance of the stripline signals is approximately the same between the two cases when using the same trace geometries and pair-to-pair spacing, allowing for a common PCB design file.

Primary Stackup DS-7409D (X)		
1.8 mil	**Top**	1.5 oz
2.3 mil	**Pre-preg**	1x1035
0.6 mil	**GND**	0.5 oz
3.9 mil	**Core**	2x1035
1.2 mil	**Power (3V3)**	1 oz
2.3 mil	**Pre-preg**	1x1035
0.6 mil	**GND**	0.5 oz
5.7 mil	**Core**	2x1078
0.6 mil	**Signal**	0.5 oz
5.1 mil	**Pre-preg**	2x1035
1.2 mil	**GND**	1 oz
2.0 mil	**Core**	1x1035
2.4 mil	**Power (1V0)**	2 oz
5.6 mil	**Pre-preg**	2x1035
2.4 mil	**Power (1V3)**	2 oz
2.0 mil	**Core**	1x1035
1.2 mil	**GND**	1 oz
5.1 mil	**Pre-preg**	2x1035
0.6 mil	**Signal**	0.5 oz
5.7 mil	**Core**	2x1078
0.6 mil	**GND**	0.5 oz
2.3 mil	**Pre-preg**	1x1035
1.2 mil	**Power (3V3)**	1 oz
3.9 mil	**Core**	2x1035
0.6 mil	**GND**	0.5 oz
2.3 mil	**Pre-preg**	1x1035
1.8 mil	**Bottom**	1.5 oz

Secondary Stackup Megatron 4		
1.8 mil	**Top**	1.5 oz
2.3 mil	**Pre-preg**	1x1067
0.6 mil	**GND**	0.5 oz
2.4 mil	**Core**	1x1080
1.2 mil	**Power (3V3)**	1 oz
2.3 mil	**Pre-preg**	1x1067
0.6 mil	**GND**	0.5 oz
5.9 mil	**Core**	2x1080
0.6 mil	**Signal**	0.5 oz
6.4 mil	**Pre-preg**	2x1080
1.2 mil	**GND**	1 oz
2.0 mil	**Core**	1x1067
2.4 mil	**Power (1V0)**	2 oz
5.9 mil	**Pre-preg**	2x1080
2.4 mil	**Power (1V3)**	2 oz
2.0 mil	**Core**	1x1067
1.2 mil	**GND**	1 oz
6.4 mil	**Pre-preg**	2x1080
0.6 mil	**Signal**	0.5 oz
5.9 mil	**Core**	2x1080
0.6 mil	**GND**	0.5 oz
2.3 mil	**Pre-preg**	1x1067
1.2 mil	**Power (3V3)**	1 oz
2.4 mi.	**Core**	1x1080
0.6 mil	**GND**	0.5 oz
2.3 mil	**Pre-preg**	1x1067
1.8 mil	**Bottom**	1.5 oz

Figure 16: Primary stackup (left): $t = 65.0$ mil, $Z_{0,o} = 97.89\ \Omega$. Secondary stackup (right): $t = 65.3$ mil, $Z_{0,o} = 98.01\ \Omega$.

Each switch chip consumes a total of 100 W of power when operating at 10.7 Gb/s and has three voltage rails:

- 1.0 volt at 60 amperes
- 1.3 volts at 30 amperes
- 3.3 volts at 0.3 amperes

The low current carrying 3V3 rails were assigned two different 1oz power planes symmetrically about the half stack close to the top and bottom layers. The high current planes are built-up using 2oz copper in order to minimize the supply rail impedance and potential IR drops during high-current switching events. These planes are allocated to the center of the board in order to comply with the standard adajcent side copper thickness differences. Placing these planes at the center of the board increases the loop inductance for these power plane return paths, posing a potential issue for the robustness of the power-delivery network (PDN). However, during standard link operation, the current consumption is not expected to vary drastically. Thus, extreme dI/dt events are not expected, and the switch chip voltage regulators can thus handle inductance penalty posed by the increased loop area can be tolerated on the PDN. Additionally, each power planee has an associated 1oz ground plane to further reduce the supply return loop area and optimize the PDN impedance.

4.2 Signal Trace Descriptions

The final router architecture avoids the use of cables to interconnect between switch chips on different boards, thus saving on cost and mechanical complexity. Thus, the only controlled impedance signal paths from the TX to RX consist of the linecard transmission lines, and orthogonal midplane connector interface. Both the linecard and crossbar PCBs utilize differential striplines for duplex signaling in each link. Striplines are selected in the design in order to provide robustness to the signaling link. By using striplines, far-end crosstalk (FEXT) is greatly reduced, since stripline signals in a homogeneous medium generate the same even and odd-mode propagation velocities (i.e., $k_C = k_M$). This results in no FEXT, since the FEXT crosstalk coefficient $K_F = \frac{k_C - k_M}{2}$. Additionally, striplines allow for tight impedance control over distance and provide EMI robustness from on-board or external aggressors to the information carrying high speed links.

In the primary stackup shown in Fig. 16, the pre-preg and core layers surrounding the signaling layer are inhomogeneous, however, their geometry and dielectric constant are designed to yield a near zero K_F coefficient. In particular, the pre-preg layer is comprised of a 2.3 mil + 2.6 mil stack (5.1 mil total) of 1035 G/F glass yielding an effective dielectric constant of $D_K = 3.205$ and an effective loss-tangent of tan $\delta = .006$ at 5 GHz. The core is made of a 2.6 mil + 3.1 mil (5.7 mil total) stack of 1035 glass, yielding an effective dielectric constant of $D_K = 3.445$ and a loss-tangent of tan $\delta = .005$ at 5 GHz. Analytical equations from [3] were used in order to select the optimal dielectric thicknesses and dielectric constants in order to achieve an odd-mode impedance close to the nominal connector impedance, while also maintaining a relatively constant total stackup height between the primary and secondary stackups. As previously displayed in Fig. 3 and 4.2, the striplines utilize a 4-4-4 spacing, whereby each polarity signal in the pair has a trace width of 4 mil and a 4 mil gap between them. Lastly, the PCB process makes use of via back-drilling to remove stubs.

To confirm the differential impedance of the signaling trace, a differential pair using the primary stackup and trace parameters was modeled in HSPICE using a 2-D electromagnetic field-solver. The self/mutual inductances and capacitances were extraced to be:

$$\begin{aligned} L_1 &= 339.5\ n\text{H/m} \\ L_m &= 46.35\ \text{nH/m} \\ C_1 &= 0.11\ \text{nF/m} \\ C_m &= 15.05\ \text{pF/m} \end{aligned} \tag{6}$$

The resulting coupling coefficients for the magnetic and electric fields are thus:

$$\begin{aligned} k_m &= \frac{L_m}{L_1} = 0.1365 \\ k_c &= \frac{C_m}{C_1} = 0.1363 \end{aligned} \tag{7}$$

The backward crosstalk coefficient can be found as:

$$K_{B,1} = \frac{1}{4} \cdot (k_C + k_M) = 0.0682 \tag{8}$$

Lastly, we can see that the differential mode impedance closely matches the design intent:

$$\begin{aligned} Z_{oo} &= \sqrt{\frac{L_1}{C_1}} \cdot \sqrt{\frac{1 - k_m}{1 + k_c}} = 48.33\ \Omega \\ Z_{o,\text{diff}} &= 2 \cdot Z_{oo} = 96.67\ \Omega \end{aligned} \tag{9}$$

Note that the switch chip contains parallel termination resistors having a nominal impedance of $Z_T = 100\ \Omega$. Thus, there exists a small impedance discontinuity between the stripline and the BGA TX/RX terminations given as: $\Gamma_T = \frac{96.67-100}{100+96.67} = -0.01609$. This mismatch is acceptable and still allows for 99.97% power transfer at these discontinuities.

4.3 Connectors and Cables

To model the connector crosstalk, a 24-port S-parameter model of 4x8 orthogonal midplane connector is simulated in HSPICE. A differential signal on the female interface connector - representing a TX signal from either the linecard or crossbar board - is excited with a 1 Vpp signal. All ports are terminated in the nominal connector impedance. The two nearest adjacent pairs are probed in order to view the near-end and far-end cross-talk that is generated at the connector interface.

Figure 17: Orthogonal midplane connector crosstalk analysis.

From the probed signals, it is observed that the connector has a peak FEXT coefficient of $K_F = \frac{3.4\ \text{mV}}{1\ \text{V}_{\text{pp}}} = 0.0034$, while the peak NEXT coefficient is $K_B = \frac{3.53\ \text{mV}}{1\ \text{V}_{\text{pp}}} = 0.00353$. As observed from Fig. 7, there exists a row where TX pairs immedietly adjacent to a row of RX pairs, thus the worst case total crosstalk noise is the summation of the peak FEXT and NEXT crosstalk. Doing this results in a total peak connector crosstalk of:

$$K_{X,\text{tot}} = 0.0034 + 0.00353 = 0.0069 \tag{10}$$

Note that by electromagnetic reciprocity, the coupling from the differentially excited signal to the adjacent signal lines is identical to the case where the adjacent signal lines are the aggressors instead of the victims. This symmetry is used when computing the noise margins required for estimating the link BER performance in Sect. 6.4.

4.4 Integrated Circuit Packages

The total BGA package size has a dimension of 5cm x 5cm, with a via-to-via pitch spacing of 1mm in the X and Y directions. Figure 8 shows one corner of the chip ball-out, with the others being identical. The total possible ball-count for this size of package would be 2601 assuming a fill factor of 100%. In practice, not all balls are flooded on the backside of the BGA. There may be a ground paddle in the

center, or a guard ring in between the SERDES logic and the core IO and power ring balls. Thus, we assume a nominal ball count of 2116, or a fill factor of 81.35%.

5 Design Analysis

Below we discuss the simulation approach of the signaling system, including extraction of crosstalk coefficients, noise, channel responses for the longest/shortest links, equalization, noise, and bit-error-rate (BER) performance. For all simulations, a differential output level of 1000 mV is used. Both the TX and RX termination resistors assume their nominal values, since on average this will be the case. For a more robust approach, Monte Carlo methods should be used to capture statistical process variation, which will have different spreads depending on whether on-chip or off-chip terminations are used. Additionally, voltage drive tolerance is not accounted for in simulation. That being said, the signaling gross margin should ideally take all significant non-idealities into account in order to conservatively estimate the BER performance, so the variation/offsets are modeled into the computation of the effective margin V_{nm} described in Sect. 5.4.

5.1 SPICE Analysis

In order to evaluate the link performance and extract the signaling margins used for BER calculations, a link-level model representing the behavior of Fig. 12 is developed in HSPICE. This model is used to extract channel S-parameters as well as identify the proper transmit and receive equalization settings. The transmitter uses a behavioral model emplying current-mode logic and $Z_T = 100\ \Omega$ parallel termination resistors, allowing for a max differential output swing of 1250 mV. In the HSPICE simulations, the differential output swing is set to 1000 mV. Additionally, a 4-Tap DLE with 1 pre-cursor and 2 post-cursors is used to combat channel intersymbol interference. The receiver employs a CTLE equalizer with 2 tunable poles and 1 tunable zero. The CTLE has 0 dB DC gain and a maximum peaking gain of 15 dB to combat post-cursor ISI.

5.2 Crosstalk

As was mentioned in Sect. 3.5, the worst case pair-to-pair spacing is 28 mils, which is the number used in all crosstalk simulations in order to conservatively model the worst-case across the end-to-end link. To model the coupling between neighboring stripline pairs on the linecard and crossbar PCBs, the longest link is simulated on HSPICE having a transmitter/receiver parallel terminations of 100 Ω, nominal stripline differential impedance of 96.67 Ω, as shown in Fig. 18. The aggressor diff pair comprised of lines 1 and 2 is driven with a 2 V differential step, and the resulting crosstalk voltage is probed victim differential pair comprised of lines 3 and 4.

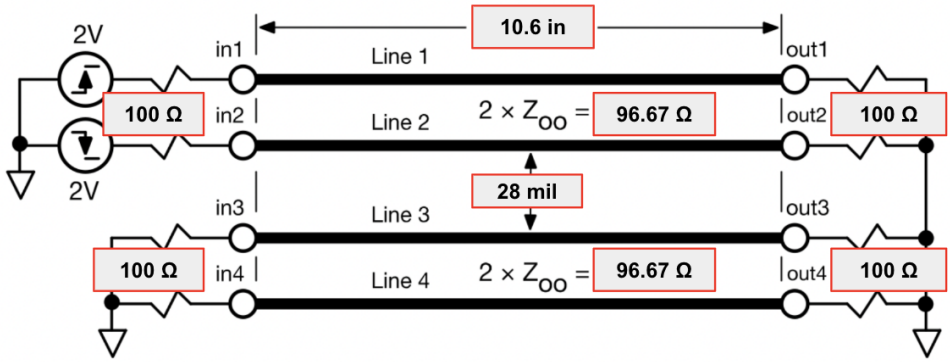


Figure 18: Stripline pair-to-pair crosstalk simulation setup.

The observed crosstalk from the HSPICE simulation is shown Fig. 19:

Figure 19: Stripline pair-to-pair crosstalk simulation (longest link).

The peak NEXT observed at V(in3) is only -68 μV, with the majority of this coming from the nearest aggressor line2. This has the same polarity as the aggressor signal and occurs right after the differential step on line2 is applied, lasting two line propagation delays before the line returns to a quiet state. We also see that FEXT in V(out3) has a peak of 82 μV, occuring one propagation delay after the aggressor signal is launched. Again, this is due to the nearest aggressor line2, and has a polarity opposite to that of the differential step on line2 since $k_M > k_C$. Note that in the BER analysis, FEXT is ignored since TX/RX signals are routed on different layers in the PCB. Additionally, we see only a small amount of crosstalk induced on line4 with the same shape profile as line3, just with much lower amplitudes and thus only the adjacent signal lines are considering when budgeting margin degradation from crosstalk signals. Based on the above model, we extract the NEXT crosstalk coefficient between

the differential traces as:

$$K_{B,2} = \frac{\text{peak}(|V(in3) - V(in4)|)}{V(in1) - V(in2)} = 3.09 \cdot 10^{-5} \tag{11}$$

5.3 Equalization

5.3.1 Equalization: Longest Link

In the end-to-end signaling system, the longest link between a TX driver and RX pair is 16.1 inches. Modeling this channel in HSPICE, the insertion loss S_{21} is observed in Fig. 20. At the signaling frequency of 10.7 Gb/s, the link sees 26.3 dB of attenuation.

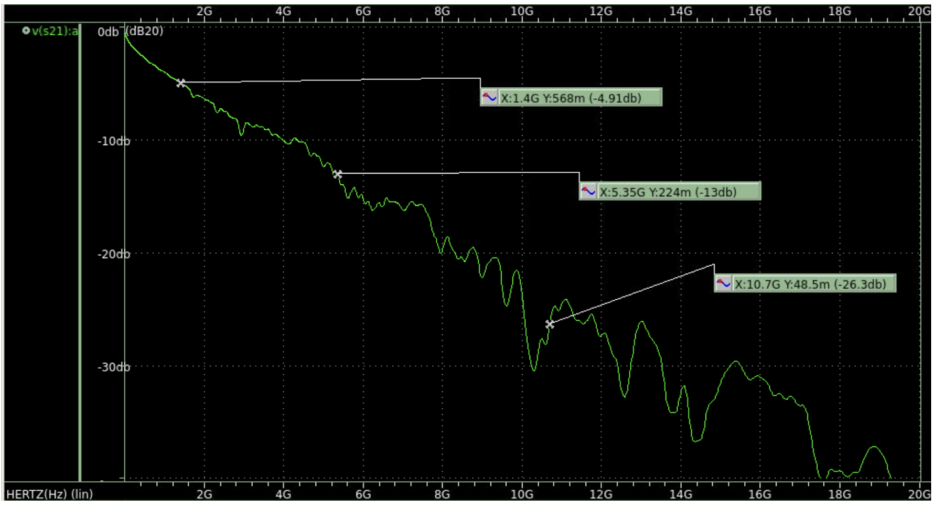


Figure 20: Longest link channel attenuation.

In order to equalize the link, the pulse response of a single bit is sent in order to determine the appropriate pre-emphasis and post-emphasis settings as shown in Fig. 21.

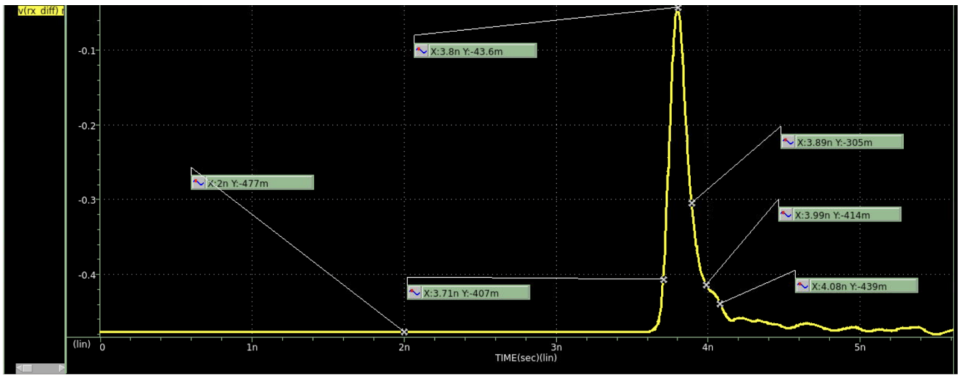


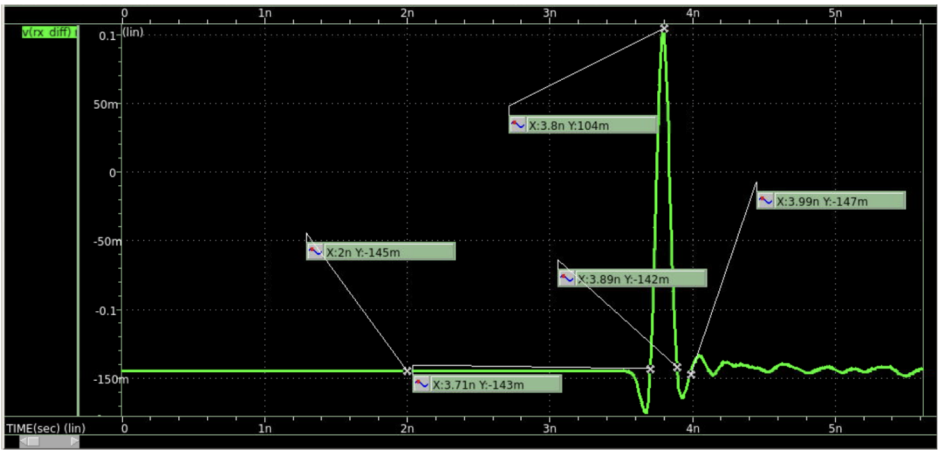
Figure 21: Longest link unequalized pulse response.

Given a coded bit-rate of $R_b = 10.7$ Gb/s, the bit period - also known as the unit interval (UI) - is $1/R_b = 93.458$ ps. The pre-cursor, cursor, and post-cursor signal amplitudes are extracted at fixed UI time offsets before and after the cursor. It is evident from the pulse response that it is mostly post-cursors that have contributed to the ISI; this is introduced by the band-limiting of the channel. Equalization is

done using equations (2)-(4) described in Sect. 4.2.2., from which we extract the normalized DLE tap coefficients $\bar{c}_i$.

	Received Voltage [mV]	Reference Centered [mV]	Normalized Tap Coefficient $\bar{c}$
Pre-cursor	-407.0	70.00	-0.105
Cursor	-43.6	433.40	0.650
Post-cursor 1	-305.00	172.00	-0.243
Post-cursor 2	-414.00	63.00	0.0019
Reference	-477.00	0.00	-

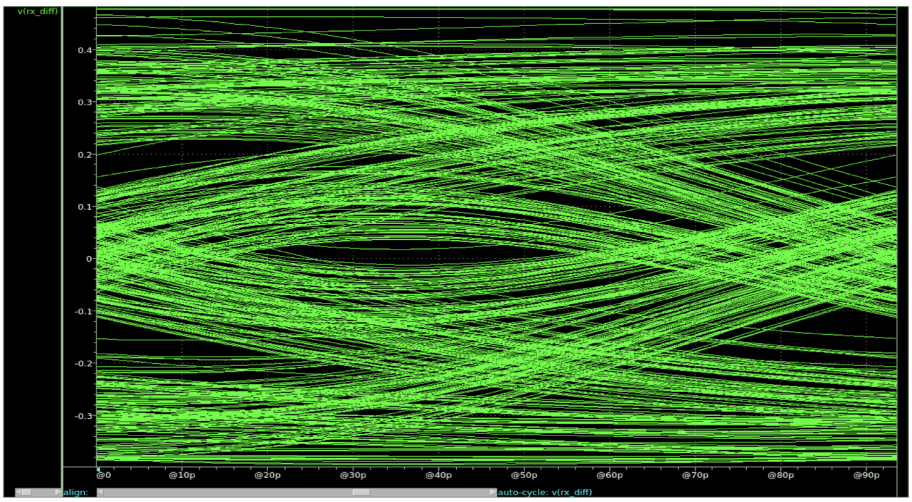
(a) Longest link DLE taps.



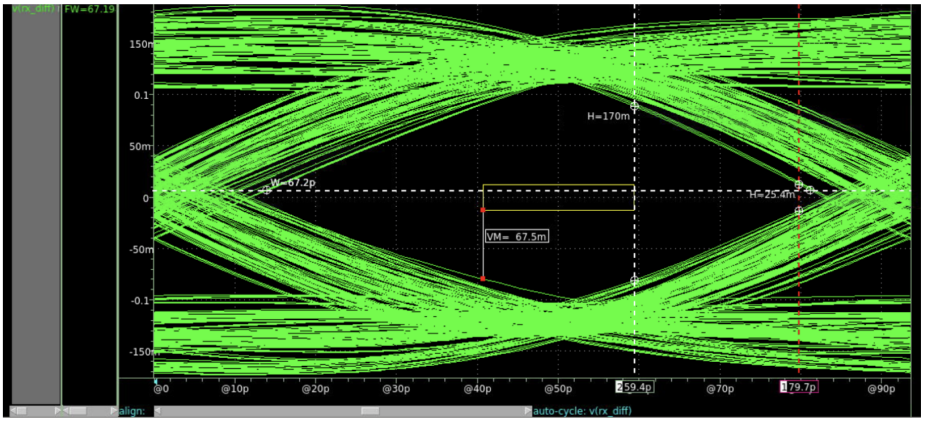
(b) Longest link equalized pulse response; DLE only.

Figure 22: Equalization settings and resulting pulse response for the longest link.

As seen in Fig. 22b, the pre and post cursors have been set to zero by the TX DLE. The resulting improvement in the channel's eye diagram is seen in Fig. 32b, demonstrating successful channel equalization. Note that the DLE has reduced the amplitude swing at the RX input in order to equalize the channel for a fixed symbol power budget.



(a) Longest link unequalized eye.



(b) Longest link equalized eye; DLE only.

Figure 23: DLE only equalization for the longest link.

The link was also independently equalized via the receiver CTLE. The equalizer's zero was swept to zero-force the first post-cursor. Figure 24 shows the pulse response of the longest link using the CTLE with no equalization from the DLE. The zero was set to 1.4 GHz, with the poles at 5.35 GHz and 10 GHz. Figure 25 is the eye diagram for the same link.

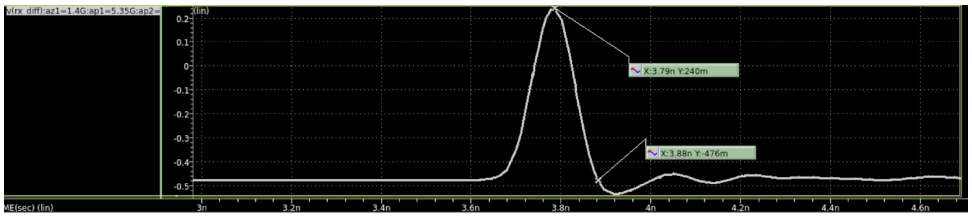


Figure 24: Longest link equalized with CTLE.

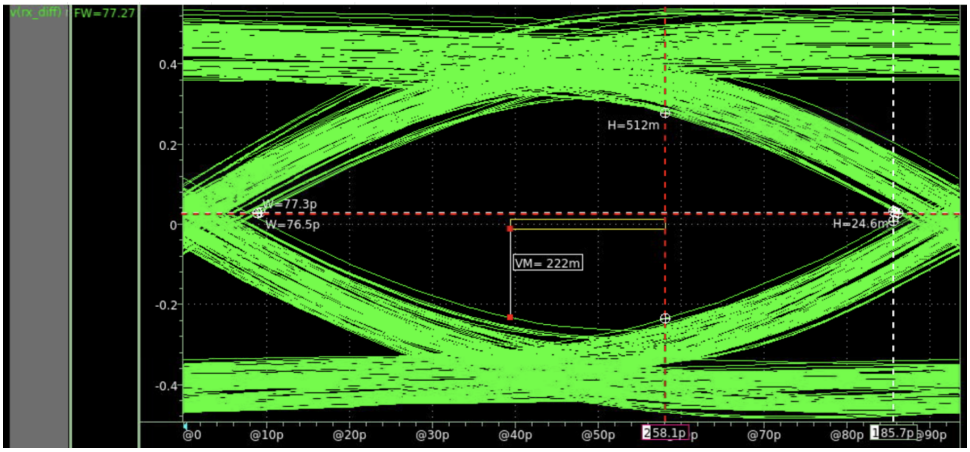


Figure 25: Longest link equalized with CTLE.

5.3.2 Equalization: Shortest Link

We perform the same exact procedure as was previously highlighted in 5.3.1 for the shortest link which has a total length of 5.0 inches from the TX driver to RX receiver.

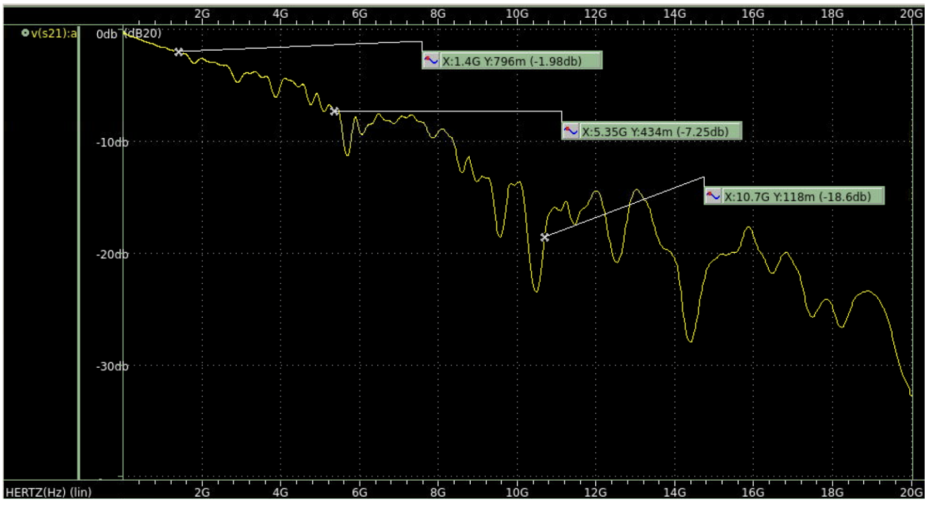


Figure 26: Shortest link channel attenuation.

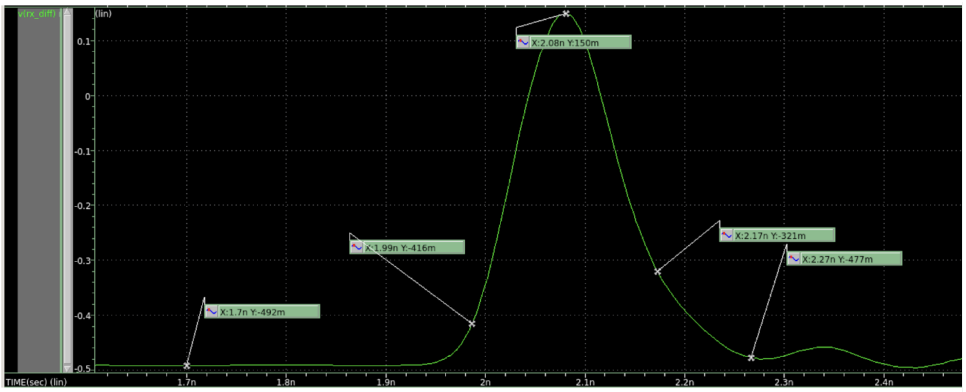
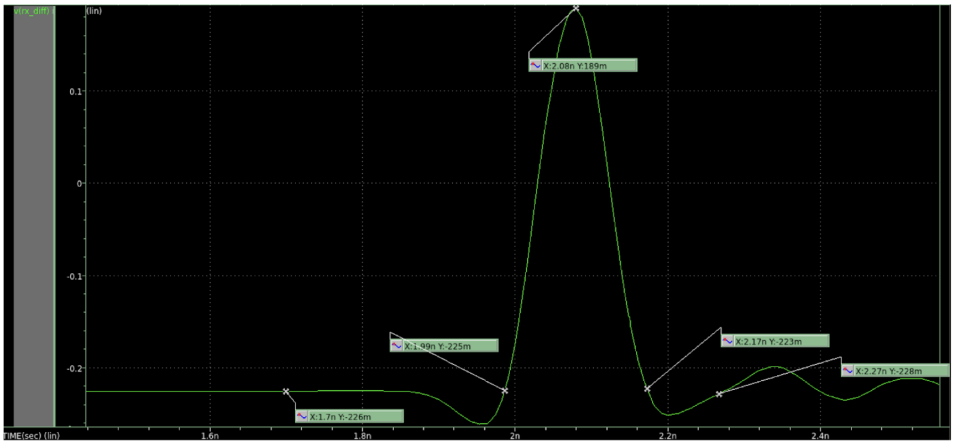


Figure 27: Shortest link unequalized pulse response.

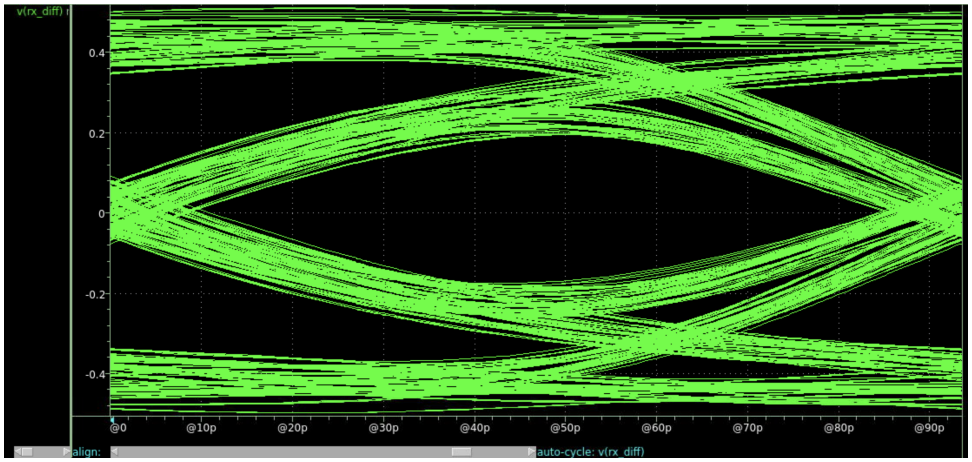
	Received Voltage [mV]	Reference Centered [mV]	Normalized Tap Coefficient $\bar{c}$
Pre-cursor	-416.0	75.00	-0.082
Cursor	150.0	641.00	0.695
Post-cursor 1	-321.00	170.00	-0.188
Post-cursor 2	-477.00	14.00	0.0340
Reference	-491.00	0.00	-

(a) Shortest link DLE taps.

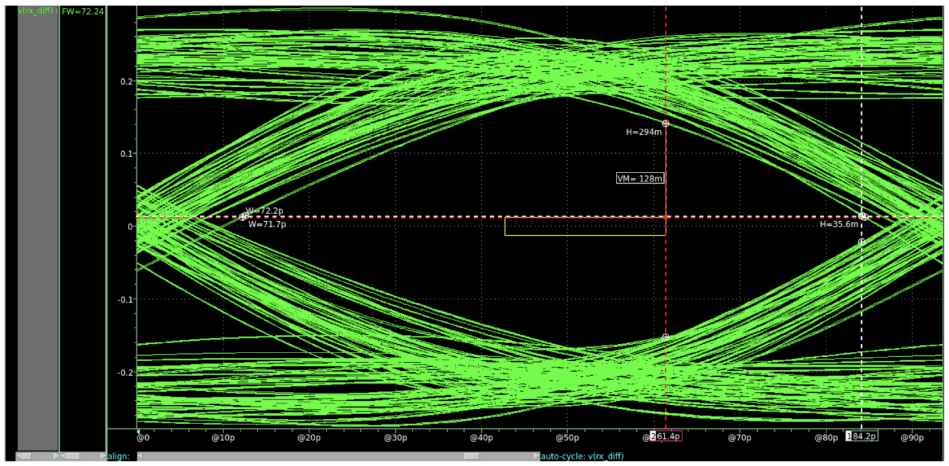


(b) Shortest link equalized pulse response; DLE only.

Figure 28: Equalization settings and resulting pulse response for the shortest link.



(a) Shortest link unequalized eye.



(b) Shortest link equalized eye; DLE only.

Figure 29: DLE only equalization for the shortest link.

The link was also independently equalized via the receiver CTLE, with the same procedure used for the longest link. The equalizer's zero was swept to zero-force the first post-cursor. Figure 30 shows the pulse response of the longest link using the CTLE with no equalization from the DLE. The zero was set to 2.25 GHz, with the poles at 5.35 GHz and 10 GHz. Figure 31 is the eye diagram for the same link.

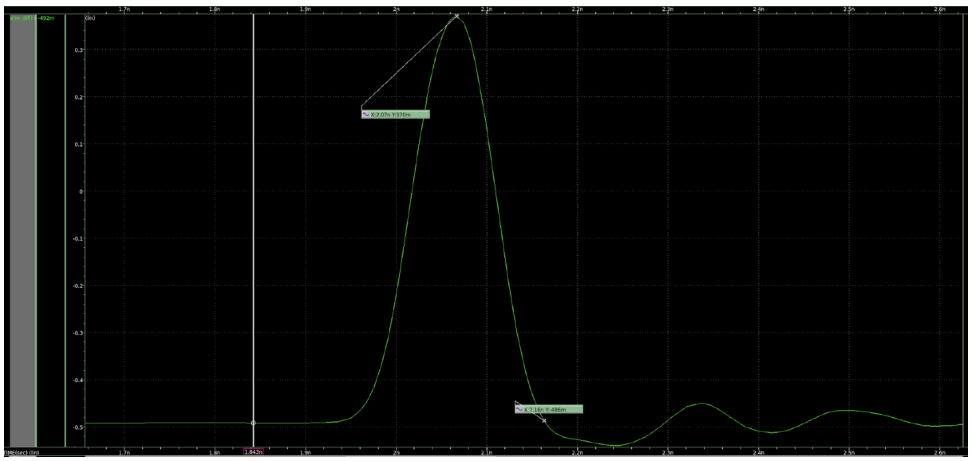


Figure 30: Shortest link equalized with CTLE.

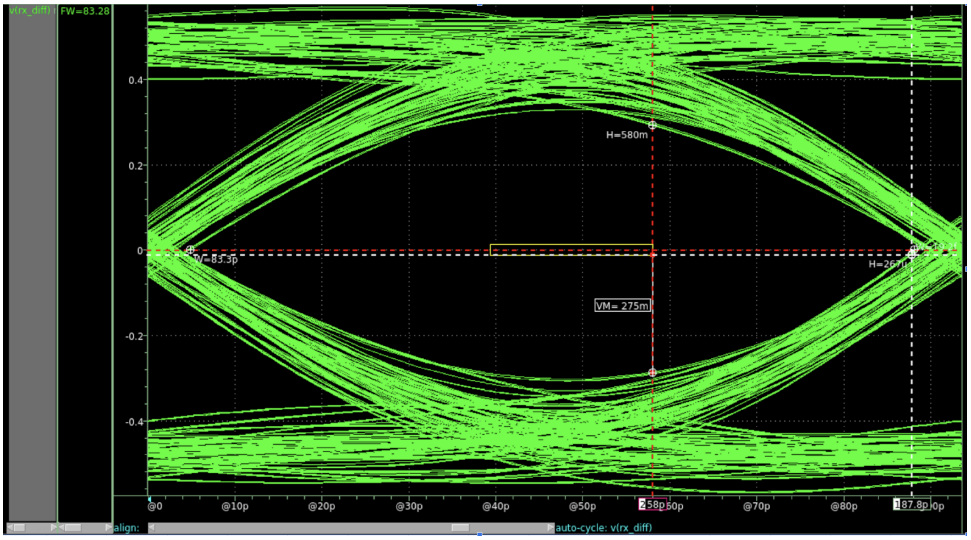
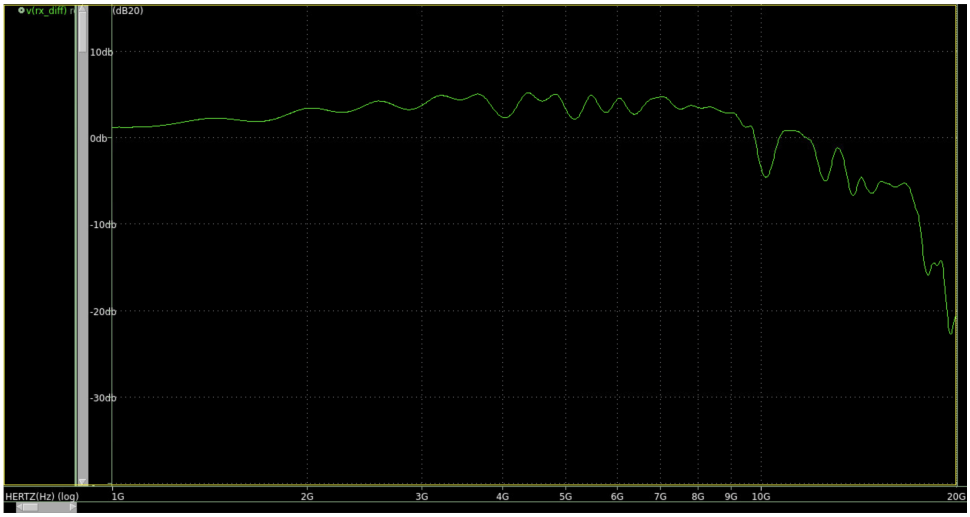


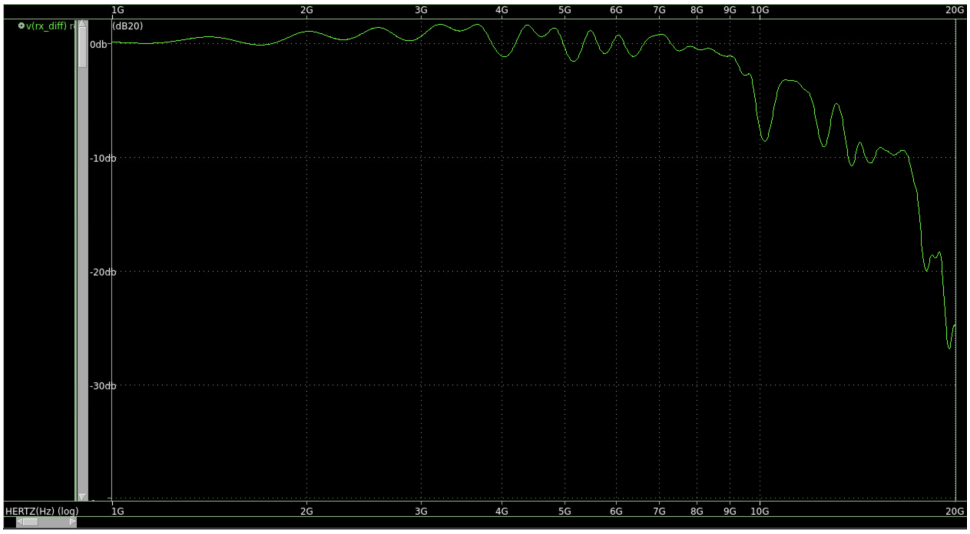
Figure 31: Shortest link equalized with CTLE.

5.3.3 Final Equalization Configuration

The CTLE and DLE equalization coefficients for both the longest and shortest links were combined in a simulation; however, we saw that the eye was no better than just using the CTLE alone, indicating that the DLE is redundant and only introduces power overhead. This indicates that the channel ISI is mostly dominated by post-cursors introduced by high frequency loss, and thus the main equalization benefit comes from resolving this. As such, in the final equalization settings for the link, we neglect the use of the TX DLE and employ a tunable two-pole/one-zero active CTLE circuit in order to equalize the channel's insertion loss. The peaking gain is extracted via AC simulation of the CTLE behavioral model, as shown in Fig. 32a & 32. The maximum peaking gain for the longest link is 5.2 dB, while for the shortest link it is 1.8 dB.



(a) CTLE AC response with zero at 1.4 GHz, lengths set to 0 inches to assess the peaking gain.



(b) CTLE AC response with zero at 2.25 GHz, lengths set to 0 inches to assess the peaking gain.

Figure 32: DLE only equalization for the shortest link.

5.4 Noise and Jitter

Here we evaluate the voltage noise and timing noise of the signaling system, highlighting those that were included in the link-level simulations and those that were ignored.

5.4.1 Voltage Noise

Below are the bounded and deterministic voltage noise sources present in the system. Note that FEXT is ignored, as TX and RX traces are on different signal layers. Additionally, we assume that 10% of the NEXT reaches the receiver. This also accounts for signals which reflect at the receiver due to termination mismatch with the stripline differential impedance. As such, the 1Vpp nominal signal that is launched will have a backward crosstalk amplitude of $V_{NEXT} = \pm 0.1 \cdot K_{B1} \cdot V_{launch} = 6.82$ mV.

The allowed value of the stochastic noise is determined after computing the effective gross margin which accounts for deterministic noise sources which were not already simulated. The allowed stochastic noise is thus the remaining amount permissible to achieve a BER of 10^{-14} as described in Sect. 5.5.

Other sources of voltage noise in the system can be attributed to systematic process variations, non-signaling related electromagnetic coupling from processors/clocks/power supplies, and device dependent noise such as thermal, flicker, and shot noise.

Bounded Voltage Noise					
Modeled in Simulation			*Not Modeled in Simulation*		
Parameter	*Value*	*Notes*	*Parameter*	*Value*	*Notes*
Intra-pair XTALK	68.2 mV	Computed using Eq. (8) as K_B1*Vpp	TX Offset	+/- 50 mV	Assumed +/- 5%
ISI	Included in eye diagram	Longest link	RX Offset	50 mV	RX aperture height
-	-	-	Inter-pair XTALK	0.0309 mV	Computed using Eq. (11) as K_B2*Vpp
-	-	-	Midplane Connector XTALK	6.9 mV	Computed using Eq. (10) as K_X,tot*Vpp
-	-	-	Adjacent NEXT	6.82 mV	Computed using Eq. (8) as 0.1*K_B1*Vpp

Figure 33: Bounded voltage noise sources present in the signaling system.

5.4.2 Timing Noise

Below are the bounded and deterministic timing jitter sources in the signaling system. For the deterministic jitter, we assume that the main source is in the CDR jitter. When using a dual-loop phase interpolator as discussed above, there is jitter as a result of the quantization error in the phase interpolator. This, is counted to be 10% of a unit interval. Then, various forms of crosstalk induced jitter also reduce our margin. Crosstalk induced jitter is determined as follows

$$CIJ = \frac{K_B V_{aggr}}{dV/dt} \tag{12}$$

where K_B is the crosstalk coefficient, V_{aggr} is the voltage of the aggressor signal (1 V), and dV/dt is the timing slope calculated via the 1 V signal swing and 30 ps rise time. Evaluating this for the various modes of crosstalk yield the values in Fig. 34.

Finally, some low jitter is estimated for the local crystals used to clock each switch chip, and for additional jitter within the timing distribution networks within the chip.

Bounded Jitter Sources					
Modeled in Simulation			*Not Modeled in Simulation*		
Parameter	*Value*	*Notes*	*Parameter*	*Value*	*Notes*
Intra-pair XTALK induced jitter	2.56 ps	Computing using Eq. (12) with KB1	CDR jitter uncertaintiy	+/- 9.345 ps	+/- 0.1*UI
ISI	Included in eye diagram	Longest link	Inter-pair XTALK	1.16 fs	Computed using Eq. (12) with KB2
-	-	-	Midplane Connector XTALK	260 fs	Computed using Eq. (12) with K_X,tot
-	-	-	Crystal Jitter	2 ps	(estimated)
-	-	-	PLL, CDR interpolation, Internal	2 ps	(estimated)

Figure 34: Bounded timing jitter sources present in the signaling system

The allowable random jitter is calculated to meet the BER specification, and is described in the next section.

5.5 Bit Error Rates

The maximum voltage noise and timing jitter are obtained from the eye-diagram of the longest signaling link, corresponding to the worst case scenario.

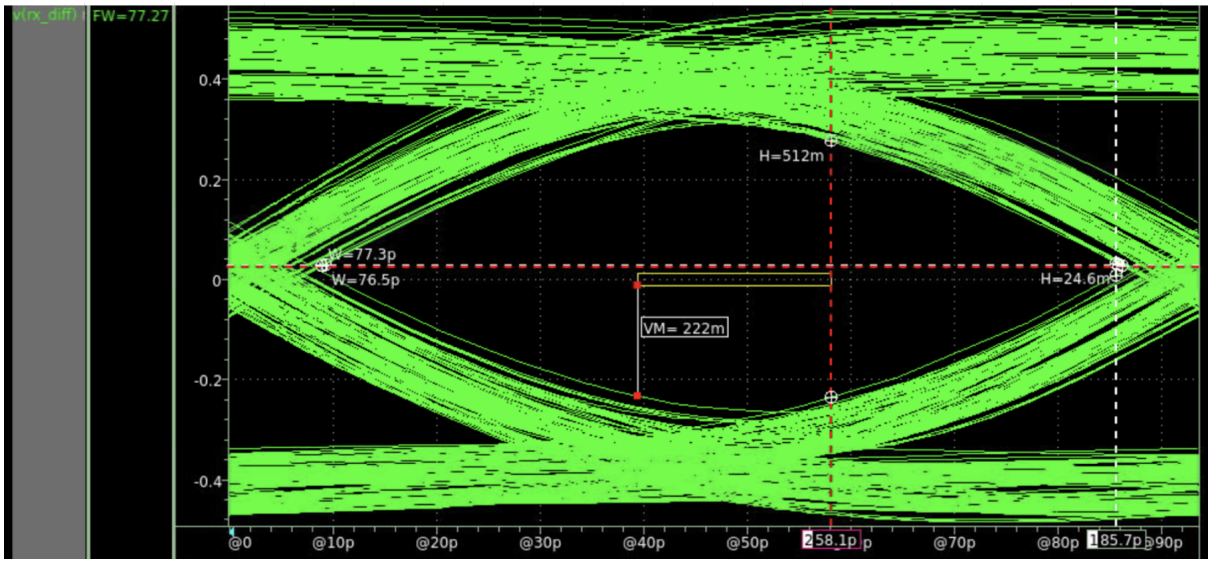


Figure 35: Longest link equalized with CTLE.

5.5.1 Stochastic Voltage Noise Limits

To determine the upper bound of the stochastic/random voltage noise permissible in the signaling system, we back-calculate what the noise can be in order to achieve an SNR that gives a design specification of BER $= 10^{-14}$. From the worst-case eye in Fig. 35, we see that the net margin of the eye is $V_{nm,\text{eye}} = 222$ mV. From the bounded noise source listed in Fig. 33, the total net margin accounting for the deterministic noise not included in the simulation is:

$$\begin{aligned} V_{nm} &= V_{nm,\text{eye}} - (V_{TX,\text{offset}} + V_{XT,\text{inter}} + V_{XT,\text{conn.}}) \\ &= 222 \text{ mV} - (50 \text{ mV} + 0.0309 \text{ mV} + 6.9 \text{ mV} + 6.82 \text{ mV}) \\ &= 158.82 \text{ mV} \end{aligned} \tag{13}$$

The stochastic noise voltage V_{rand} is back-calculated based on the complementary error function:

$$\begin{aligned} BER &= \frac{1}{2}\,\text{erfc}\left(\frac{V_{SNR}}{\sqrt{2}}\right) \\ &= \frac{1}{2}\,\text{erfc}\left(\frac{1}{\sqrt{2}} \cdot \frac{V_{nm}}{V_{rand}}\right) \end{aligned} \tag{14}$$

Given that erfc() is invertible, we determine the random noise voltage allowable to achieve the garget BER as:

$$V_{rand} = \frac{V_{nm}}{\sqrt{2} \cdot \text{erfc}^{-1}(2 \cdot BER)} = 21.5 \text{ mV, rms} \tag{15}$$

This yields a required SNR of 17.32 dB.

5.5.2 Stochastic Timing Jitter Limits

A similar procedure is used to determine the upper bound of the random timing jitter permissible in the system to achieve a specification of BER $= 10^{-14}$. Again, using the eye diagram in figure 35, we see that the net timing margin between the overlayed mask and the closest edge is $t_{nm,\text{eye}} = 26.7$ ps. Using

the bounded noise sources listed in Fig. 34, the total net timing margin accounting for the deterministic noise not included in the simulation is:

$$t_{nm} = t_{nm,\text{eye}} - t_{\text{bounded}} \tag{16}$$

The value for the timing noise margin is 13.6 ps. The BER as a result of timing error can be expressed as

$$BER = \frac{1}{2}\operatorname{erfc}\left(\frac{1}{\sqrt{2}}\frac{t_{nm}}{t_{rand}}\right) \tag{17}$$

and solving for t_{rand} is straightforward using the inverse error function.

$$t_{rand,rms} = \frac{t_{nm}}{\sqrt{2}\operatorname{erfc}^{-1}\left(2 \cdot 10^{-14}\right)} = 1.83 \text{ ps, rms} \tag{18}$$

This yields a required SNR of 20.6 dB.

5.6 Power Dissipation and Cost

As briefly discussed above, system cost was mainly influenced by the choice of signaling speed. Using the higher speed reduced the number of switch chips and connectors required in the system, and may have also saved cost on PCB area. A summary of our system cost is shown in Fig. 1, and a bill-of-materials of hardware components is shown in Fig. 36.

Each printed circuit board is manufactured out of a standard panel size, which has a fixed cost. Reducing the number of panels purchased reduces overall system cost. Since the dimensions of the line card were fixed, we could not make significant reductions to the panel count. The layer count and loss tangent chosen for the boards is discussed in Section 5. Mid-loss dielectrics were chosen to reduce cost. Overall, we need nine 14-layer mid-loss panels for the line cards and crossbar cards and one 8-layer mid-loss panel for the midplane.

Connectors are priced at 35 cents per differential pair, and we require a total of 64 4x8 orthogonal midplane connectors, 64 corresponding receptacles on the midplane, four 3x6 orthogonal midplane connectors, and four corresponding receptacles on the midplane.

Finally, the power consumption of our system is informed by the signaling rate. However, power is kept to a minimum in this signaling scheme by only using one switch chip per board. The majority of the system cost is attributed to power consumption, The associated costs for all twelve switch chips in total are as follows: $9360 for the 1.3 V 30A link rail, $14,400 for the 1.0 V 60 A core power rail, and about $240 for the 3.3 V 0.3 A peripheral and low-speed IO rail.

The total system cost is approximately $31,000, with about 22% of that cost dedicated towards PCBs, chips, and connectors and the remaining cost due to power consumption.

Hardware Bill-of-Materials			
Part	**Quantity**	**Cost per Unit**	**Total Cost**
Mid-loss 14-layer 18" x 25" PCB Manufacturing Panel	9	$249	$2,241
Mid-loss 8-layer 18" x 25" PCB Manufacturing Panel	1	$193	$193
Switch Chip	12	$300	$3,600
4x8 Orthogonal Midplane Connectors & Receptacles	64	$11.20	$716.80
3x6 Orthogonal Midplane Connectors & Receptacles	4	$6.30	$25.20
		Total Cost	$6,776

Figure 36: Hardware costs of system.

6 Conclusion

In this report, we have introduced our design for a 5 Tbps router system. We were able to meet all required specifications, while keeping cost to only $31,000, including power consumption. The link equalization is implemented in the most simple way that still maintains a large open eye. Due to area not being constrained, differential links in our busses are spaced far apart to reduce crosstalk, leaving plenty of voltage marin to meet the required BER.

If given the opportunity to improve our design, we would have liked to design an active CTLE circuit to see performance improvement. We could improve our equalization to open the eyes even further, which would push our BER down. Achieving the BER = 10^{-14} specification now would yield about 10 bit errors per day, which is quite low given that our signaling scheme builds in error correction. However, we may be able to reduce this further and reduce the amount of error correction overhead in the datastream, increasing throughput to some degree.

Doubling the bandwidth of this system is still within the capabilities of our switch chips, connectors, and PCB materials. However, we would need double the switch chips, which would increase our PCB layer count to at least 18 layers to accommodate more stripline routing. We could use higher density or larger connectors, such as the 6x12 orthogonal connectors from Amphenol's XCede line to prevent using an increased number of connectors.

References

[1] S. Palermo, "Lecture 5: Termination and TX driver design," https://people.engr.tamu.edu/spalermo/ecen689/lecture5_ee720_termination_txdriver.pdf, 2021.

[2] J. F. B. et. al., "A 10-gb/s 5-tap DFE/4-tap FFE transceiver in 90-nm CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 12, pp. 2885–2896, 2006.

[3] S. H. Hall and H. L. Heck, *Advanced Signal Integrity for High-Speed Digital Designs*. John Wiley & Sons, 2011.

[4] M. Hsieh and G. E. Sobelman, "Architectures for multi-gigabit wire-linked clock and data recovery," *IEEE Circuits and Systems Magazine*, 2008.

[5] S. Palermo, "Lecture 12: CDRs," https://people.engr.tamu.edu/spalermo/ecen689/lecture12_ee720_cdrs.pdf, 2025.