Title
Design of 60+Gb/s Serial-Link Transmitters Using Filter Techniques

Permalink
https://escholarship.org/uc/item/1h08466w

Author
Chen, Ming-Shuan

Publication Date
2015

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA

Los Angeles

Design of 60 Plus Gb/s Serial-Link Transmitters

Using Filter Techniques

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Electrical Engineering

By

Ming-Shuan Chen

2015
ABSTRACT OF THE DISSERTATION

Design of 60 Plus Gb/s Serial-Link Transmitters

Using Filter Techniques

by

Ming-Shuan Chen

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2015

Professor Chih-Kong Ken Yang, Chair

The demand for off-chip data bandwidth of serial-link transceivers have been pushed beyond 40-Gb/s/link. With the potential aggregate data rate, power consumption and area must be reasonable. State-of-the-art serial-link transmitters operating at data rates beyond 40-Gb/s can contribute up to 40-45% power consumption while the FFE is limited to 2 taps. The main design challenge comes from the fact that producing output signal levels using large devices at high frequency is very power hungry. It is critical to low the power of transmitter at high data rate to
enable high performance serial links.

This dissertation explores the potential of operating transmitters at data rates that approach the fundamental technology limit. Several circuit techniques are proposed to improve power efficiency and performance of serial-link transmitters. A LC-based feedforward equalizer (FFE) is proposed to eliminate the need of multiple MUXs. The design methodology of LC-ladder filter is proposed and applied to the FFE design to improve the bandwidth and the use of inductors. The serializer is power-optimized using a direct 4:1 multiplexer (MUX) at final stage with a novel 4:1 MUX design. An inverter-based digital-to-phase converter (DPC) using harmonic rejection filter is proposed to improve the phase linearity.

To verify the effectiveness of the proposed techniques, a 64-Gb/s transmitter prototype and an 8-bit DPC are designed and measured in 65-nm CMOS technology. The optimal use of on-chip inductors improve the bandwidth and power efficiency significantly. The proposed transmitter prototype, achieves a maximum data rate of 64.5 Gb/s and an energy efficiency of 3.1 fJ/b. The DPC demonstrates a maximum INL and DNL of 1.33 and 0.52 LSB while consumes a power of 4.3 mW at 1.5 GHz.
The dissertation of Ming-Shuan Chen is approved.

Brian Regan

Mau-Chung Frank Chang

Sudhakar Pamarti

Chih-Kong Ken Yang, Committee Chair

University of California, Los Angeles

2015
To my lovely wife E-Chin
# Table of Contents

ABSTRACT OF THE DISSERTATION ................................................................................ ii

Table of Contents ........................................................................................................ vi

Acknowledgments ......................................................................................................... x

VITA ............................................................................................................................... xii

CHAPTER 1 Introduction ............................................................................................ 1

1.1 Motivation ............................................................................................................. 3

1.2 Thesis Organization ........................................................................................... 8

CHAPTER 2 Background .......................................................................................... 11

2.1 Serial-Link Transmitter Basics ........................................................................ 11

2.1.1 Transmitter Basics ..................................................................................... 12

2.1.2 Transmitter Performance Metrics ............................................................... 14

2.2 Channel Equalization and Feedforward Equalizer (FFE) ............................ 19

2.2.1 Channel Equalization Basics .................................................................. 19

2.2.2 CTLE ........................................................................................................... 23

2.2.3 DFE ............................................................................................................. 25

2.2.4 FFE ............................................................................................................. 27

2.2.5 Requirements and Challenges of FFE at 60+Gb/s ................................... 32

2.3 Serializer Architecture and Multiplexer Design ........................................... 36
2.3.1 Serializer Architecture .................................................................................. 37
2.3.2 Multiplexer ....................................................................................................... 40
2.3.3 Design Challenges of 60+Gb/s Serializers ...................................................... 42
2.4 Digital-to-Phase Converter Basics ....................................................................... 50

CHAPTER 3  A 4-Tap, 64-Gb/s FFE Using LC-Ladder Filter ........................................ 55
3.1 Delay-Line-Based FFE Architecture ..................................................................... 56
3.2 LC-Ladder-Filter-Based Distributed Combiner Network ...................................... 59
3.3 FFE Delay Line Using LC-Ladder Filter ............................................................... 71
3.4 FFE Architecture and Performance Discussions ................................................ 76
3.4.1 FFE Architecture ............................................................................................ 77
3.4.2 Area Comparison ............................................................................................. 79
3.4.3 Non-1-UI Delay ............................................................................................... 80
3.4.4 Data Rate Limitations ..................................................................................... 82

CHAPTER 4 Design of a 64-Gb/s Serializer Using Direct 4:1 Multiplexing ............... 84
4.1 Transmitter Architecture ...................................................................................... 84
4.2 4:1 Multiplexer .................................................................................................... 87
4.3 Predriver ............................................................................................................... 94

CHAPTER 5 An Inverter-Based Digital-to-Phase Converter Using Harmonic Rejection .......................................................................................................................... 96
5.1 Phase Interpolator Phase Nonlinearity ................................................................. 97
References ............................................................................................................................. 158
Acknowledgments

Foremost, I would first like to thank my advisor, Professor Ken Yang, for the continuous help, support, and guidance that he gave to me throughout this journey. I find it really hard to express my gratitude for all the time he spent with me discussing my ideas, reviewing my papers, listening to my design problems, and always being ready with advice and the right feedback even under his busiest time. I feel myself lucky and proud for having him as my advisor.

I would also like to thank my dear friend and colleague, Amr Amin Hafez, who guides me in the beginning of my PhD study and explores with me the field of transmitter design. Without his great precedent works, I would have never been able to get my transmitter chip work in time.

I would also like to thank my committee members Prof. Frank Chang, Prof. Sudhakar Pamarti, and Prof. Brian Regan for their support. Special thanks to Prof. Frank Chang for his continuous support with the fabrication of all my chips and his great contribution to the EE department. I would not have made it without him.

I am lucky and thankful for contributing to the Intel 64Gbps transceiver project. It is my great pleasure to collaborate with Brian Casper from Intel, Prof. Elad Alon and Prof. Vladimir Stojanovic from UC Berkeley, Prof. Patric Chiang from Oregon State University, and Prof. Pavan Hanumolu from UIUC. I would like to particularly thank my dear friend, Yue Lu, and Prof. Elad Alon from UC Berkeley, for their generous supports in system and circuit design. Without discussing and absorbing their experience in 64G receiver design, I would not have made my chip work.

I spent very enjoyable three summers interned in Xilinx, Rambus, and Apple Inc., working with
talented SERDES designers whom I learned so much from. I would like to thank Jafar Savoj and Ken Chang from Xilinx, Kambiz Kaviani, Yohan Frans, Huy Nguyen, Kevin Cai, and Dinesh Patil from Rambus, and Raman Thiara, Daniel Friedman, Sanjeev Maheshwari, Rick Chen, and Wenbo Liu from Apple for these wonderful experiences. I gained a lot of great ideas and experiences from working with these people.

During my stay at UCLA, I was blessed to be among a group of wonderful people who made my stay in Los Angeles a wonderful memory and who were always there for me, helping me get over every technical and non-technical obstacle I have met. I would like to thank my dear friends and colleague, Yousr Ismail, for great discussion and friendship. I would like to thank E-Hung Chen, who spent a lot of time helping and guiding me in the beginning of my PhD life. I would like to thank our group members Ramy, Wonho, Tamer, and Henry for their useful discussion and friendships.

I would finally like to express my greatest gratitude toward my family; my parents, my sister, and my lovely wife for the unlimited love and support they gave to me throughout my entire life without which I would have never been even close to where I am today.
## VITA

<table>
<thead>
<tr>
<th>Year</th>
<th>Position</th>
<th>Institution</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>B.Sc, Electrical Engineering</td>
<td>National Tsing Hua University, Hsinchu, Taiwan</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>M.Sc, Electrical Engineering</td>
<td>National Taiwan University, Taipei, Taiwan</td>
<td></td>
</tr>
<tr>
<td>2008-2009</td>
<td>Chief Counselor, R.O.C. Marine Corps</td>
<td>Taiwan</td>
<td></td>
</tr>
<tr>
<td>2009-2010</td>
<td>Research Assistant, Graduate Institute of Electronics Engineering</td>
<td>National Taiwan University, Taipei, Taiwan</td>
<td></td>
</tr>
<tr>
<td>2010-2015</td>
<td>Research and Teaching Assistant</td>
<td>Electrical Engineering Department</td>
<td>University of California, Los Angeles, California, USA</td>
</tr>
<tr>
<td>2011</td>
<td>Circuit Design Intern</td>
<td>Xilinx Inc., San Jose, California, USA</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Circuit Design Intern</td>
<td>Rambus Inc., Sunnyvale, California, USA</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Circuit Design Intern</td>
<td>Apple Inc., Cupertino, California, USA</td>
<td></td>
</tr>
</tbody>
</table>

## PUBLICATIONS


CHAPTER 1

Introduction

The demand of off-chip data bandwidth has continuously pushed data rates of serial-link transceivers toward higher frequencies. Applications such as multi-core processors, ultra-high definition displays, and high-speed Ethernet have pushed the aggregate data bandwidth beyond 400 Gb/s. Figure 1.1 and 1.2 show bandwidth predictions of microprocessor [1] and Ethernet [2-3], respectively. Both of them indicate the bandwidth requirement to exceed 1 Tb/s within five years and the power dissipation in I/O to dominate the system.

Fig. 1.1 Bandwidth prediction for microprocessor system [1]
With the potential aggregate data rates of multiple Tb/s, area cost and power consumption of high-speed serial links must be reasonable. Although the bandwidth requirement can possibly be fulfilled by applying more channels and lowering per pin data rate, the increased number of transceiver pairs leads to higher number of I/O pins and area cost. Increasing the data rate per pin is more cost effective. However, as shown in the energy efficiency and data rate survey of recently published serial-link transceivers in Fig. 1.3, poor energy efficiency is observed at high per pin data rate, suggesting large overall power consumption. Data rates of conventional serial-link transceivers are limited to less than 40 Gb/s. This dissertation, targeting on improving the energy efficiency of 60+Gb/s serial-link transmitters, introduces circuit techniques that optimizes the bandwidth and power consumption by using classical filter theories. The motivation and organization of this dissertation are introduced in Section 1.1 and 1.2, respectively.
1.1 Motivation

A typical serial-link system comprises of a transmitter, a receiver, and a channel. The main function of a transmitter is to generate a single, high-speed output data stream from a number of low-speed data. Depending on the target channel responses, proper equalization, i.e., feedforward equalizer (FFE), is commonly required to minimize the effects of intersymbol interferences (ISI). Within a transceiver, a transmitter typically contributes to 30 - 35% of power consumption [4-7], while this proportion of power can rise to 40 – 50% as the data rate higher exceeds 40 Gb/s [8-9]. To illustrate the trend of transmitter power under different data rates, Figure 1.4 illustrates the energy efficiency versus the data rates of recently published serial-link transmitters with data rates higher than 20 Gb/s [5-25] with the performance of our precedent work [24] and this work [25].

Fig 1.3   Energy efficiency v.s. data rate of published serial-link transceivers
highlighted. As can be seen, although many transmitters operating at 20-32 Gb/s achieves sub-5-pJ/b energy efficiency, as the data rate exceeds 40 Gb/s, the energy efficiency can hardly be lower than 10 pJ/b. The information on whether a transmitter including an FFE or not is also included, which suggests that no publish papers have ever implemented FFE at data rate beyond 40 Gb/s. State-of-the-art serial-link transmitters require at least 2-taps of FFE to overcome the channel-induced ISI. Significant improvement in energy efficiency of FFE-embedded transmitters is needed in order to enable high performance platforms without being restrained by the I/O power dissipation.

![Fig. 1.4: Energy efficiency of published transmitters with data rate > 20 Gb/s](image)

One of the main challenges for high-speed transmitter design lies on how to produce sufficient output signal levels with reasonable power and bandwidth. The most power hungry blocks of a
transmitter are those operating at full data rate including the final-stage multiplexer (MUX), the subsequent predriver, and the output driver. Although recent publications demonstrate significant power saving using CMOS logic for MUX [26] and output drivers [4, 5, 18, 26, 27, 28], the maximum data rate is limited to 16 Gb/s for 65-nm technology. At higher data rate, using current-mode logic (CML) to meet the speed requirement for MUX becomes inevitable. Even with CML, the power is eventually limited by the self-loading drain capacitance at high frequency. To demonstrate this limitation, Fig. 1.5(a) shows a general CML predriver (N=1) and N:1 multiplexer (N > 1), of which the loading capacitor, \( C_L \), is obtained by the gate capacitance of the subsequent stage of output driver or predriver. Fig. 1.5(b) shows the corresponding power consumption under different N. As can be seen, the power increases linearly with the data rate at low data rates. As the data rate continuously increases, the power grows up rapidly and eventually goes to infinity.

Fig. 1.5: (a) A general multiplexer (N > 1) and predriver (N=1) circuit, and (b) the corresponding power consumption under different N.
self-loading capacitances become dominant. The maximum data rate is limited by the self-loaded capacitance.

Similar to the predriver and the multiplexer, the design of transmitter output network has similar issues but different concerns. Figure 1.6 shows a structure of general output driver network. To maintain good signal integrity and minimize the effect of reflection, both insertion and return loss responses of transmitter output network should sustain high bandwidth. Typical transmitters require the return loss response ($S_{11}$) to be lower than -10 dB across DC to the Nyquist frequency. Unfortunately, many components in the transmitter output network induce parasitic capacitance and limit the bandwidth. For example, for a transmitter with N-tap feedforward equalizer (FFE), N driver buffers, of which the current are digitally adjusted by external current digital-to-analog converters (DAC), are required to drive the transmitter output nodes. To sustain a large enough signal swing, the transistor size of these drivers are not small, leading to large drain capacitances. In addition, ESD devices are required to fulfill certain ESD protection levels, which in general contributes to at least 200 fF of capacitance. The output PAD can contribute to additional parasitic capacitances. The total output capacitance is typically higher than 500 fF, which limits the signal bandwidth to less than 15 GHz. The low bandwidth can induce additional loss, lower the equalized eye opening, cause reflection, and damage signal integrity.
One common way to resolve this power and bandwidth problem is to use apply on-chip inductors. Inductive peaking and transformer (T-Coil) peaking were proposed to broaden the bandwidth of predriver, multiplexer [20, 30], and output driver circuits [28, 31], which can theoretically improve the bandwidth by factors of 1.8 and 2.8, respectively. In practice, even with T-Coil peaking, which is essentially two inductors in series, the maximum return loss bandwidth (-10 dB) of a transmitter can hardly exceed 30 GHz and thus limits the maximum achievable data rate to no higher than 60 Gb/s. This observation suggests that more than two inductors are needed. Unfortunately, conventional design approaches rely on empirical tuning and can hardly handle circuits that contain more than two inductors. Since on-chip inductors occupy large area, the design should be carefully optimized to minimize the area cost. It is thus important to figure out a way to design and optimize the use of on-chip inductors for transmitter circuits.
In this dissertation, we introduce several circuit techniques based on classical filter theories to improve the power and performance of critical transmitter blocks. First of all, we explore the possibilities of optimizing the use of on-chip inductors in extending the bandwidth of a transmitter output network. An LC distribution network is proposed to distribute the output parasitic capacitance, which achieves a broadband insertion and return loss bandwidth that are required to fulfill the aggregate bandwidth requirement. As will be shown in the latter sections, the proposed approach provides with a way to optimally choose the passive components and minimize the inductor area cost. To reduce the power of FFE, a delay-line-based FFE that generates the required delay using LC-ladder LPF which avoids replicating the entire multiplexing tree and achieves more than 75% power saving for each additional tap is proposed. The serializer design is reviewed and a novel circuit structure is proposed to improve performance of the speed-bottleneck, the final-stage 4:1 MUX. A 64-Gb/s transmitter prototype design is presented to demonstrate the LC-ladder-filter-based FFE and the high-speed serializer design. Finally, we present an example of improving the performance of a digital-to-phase converter (DPC), which is an essential building block in serial-link transceivers, by applying proper filter techniques to clock signals.

1.2 Thesis Organization

The dissertation consists of seven chapters. Chapter 2 gives the necessary background on the
serial-link transmitters. This chapter begins with a brief discussion on the transmitter basics and important performance metrics of a serial-link transmitter. Following that, we briefly review the importance of channel equalization and present several commonly used equalizers including CTLE, DFE, and FFE. Then, we discuss about the design requirements and design challenges of operating the transmitter-side equalizer, FFE, at 60+Gb/s. Next, we discuss conventional serializer architecture and the design challenges of multiplexer circuits. This chapter ends with a brief review on the phase-tunable clock circuits, digital-to-phase converter (DPC).

Chapter 3 presents the design of a 4-tap, LC-ladder-filter-based FFE. We begin with a comparison of the conventional and the delay-line-based FFE structure, and demonstrate the improvement on power saving that the delay-line-based FFE structure can achieve. Following that, we turn to a discussion on the return loss issue and present our proposed LC-ladder-filter-based LC combiner network. The delay line of the FFE using LC-ladder filters is then presented and the overall FFE architecture is demonstrated. This chapter is summarized with a discussion on the potential issues of the proposed FFE.

Chapter 4 presents the design of our 64-Gb/s serializer. The architecture based on the design of our precedent paper [24] is presented. Following that, we present a novel 64-Gb/s, 4:1 MUX that improves both bandwidth and power. The design issues of using 4:1 rather than 2:1
multiplexing are discussed, followed by the design of the 32-to-16-GHz divider, which can potentially impact the performance of the 4:1 MUX. Finally, the design of the predriver is presented.

In Chapter 5, we present a wide-range, 7-bit DPC design using harmonic rejection (HR) filter. We begin with an introduction on sinusoidal phase interpolation, demonstrating that by using a sinusoidal wave in phase interpolation, a significant improvement on phase linearity can be achieved without lowering the clock amplitude. Then, the principle and operation of HR filters are introduced and our proposed high-precision DPC using HR filter is demonstrated. Following that, the circuit implementation of a prototype 1.5-GHz DPC is presented, which employs an inverter-based phase interpolator structure that achieves significantly lower power and lower area than conventional current-mode-logic-based phase interpolator design. The chapter is summarized with the introduction of nonlinear phase interpolation technique, which can further improve the phase linearity for applications that requires ultra high phase linearity.

A 64-Gb/s transmitter and a 0.1-1.5-GHz DPC prototypes are implemented in 65-nm to verify the results of this study. The measurement results are discuss in Chapter 6. Finally, we summarize this dissertation with a conclusion in Chapter 7.
CHAPTER 2

Background

This chapter provides with necessary background for serial-link transmitter design. An introduction on common transmitter architecture and performance metrics is presented in section 2.1. In the next two sections, two of the main transmitter functions, equalization and serialization, are discussed. Section 2.2 reviews the basic of channel equalization and several implementations of equalizers including continuous-time linear equalizer (CTLE), decision-feedback equalizer (DFE), and feedforward equalizer (FFE). This section is summarized with a discussion on the design requirement and challenges of our target 60+Gb/s transmitter-side FFE. In section 2.3, the serializer architecture and common multiplexer circuit are reviewed. Then, the design difficulties of 60+Gb/s serializers, including the timing constraints and the bandwidth limitations, are discussed. This section ends with an introduction on the quarter-rate architecture, which can relax the timing requirement and enable higher data rate. Finally, we review the basics of DPC, a critical building block for both serial-link transmitters and receivers, in section 2.4.

2.1 Serial-Link Transmitter Basics

This section introduces the basics of serial-link transmitter. The general transmitter
architecture and the role of transmitters are discussed in 2.1.1. The performance metrics of transmitters are presented in 2.1.2.

2.1.1 Transmitter Basics

The main function of the transmitter is to generate a full-rate data stream with sufficient voltage swing such that that after propagating through the lossy channel, the received signal swing is large enough to make the receiver capable of distinguish a 0 from a 1 without errors. A general transmitter architecture, as shown in Fig. 2.1(a), consists of a PLL (sometimes called clock multiplication unit (CMU)), an M:1 serializer, and an equalizing filter. Two of the main functions of transmitter are serialization and waveform shaping, is shown in Fig. 2.1(b). Serialization is completed with a serializer. Figure 2.1(b) shows examples of timing diagrams of the input and output data of a transmitter. Consisted of several multiplexing stages, the serializer takes a number of low-speed data ($D_1$ to $D_N$, with a data rate of $R_b/N$) and multiplex them into a single data stream with data rate of $R_b$. The multiplexing is implemented by selecting one of multiple input data by clocks with proper timing and frequency. To produce high-quality output data, the bandwidth of each multiplexing stage must be sufficient and low timing jitter is required for the clocks used for multiplexing. Serializing multiple low-speed data into high-speed one enables higher per pin data rate and lowers the cost on the number of pins and physical interfaces. The details of serializer and
Fig. 2.1: (a) Typical transmitter architecture, and (b) the corresponding timing diagram

After the full-rate data is generated, the signal is sent to an equalizing filter for waveform shaping. An equalizing filter shapes the output waveform and propagates the output data into the channel. To ensure the transmitter output demonstrating a sufficient voltage swing, one or multiple stages of predriver, which are essentially high-speed amplifiers with certain gain and bandwidth requirements, are applied to amplify the signal and to drive an output driver. The output driver interfaces the on-chip and off-chip circuits and delivers the output signal into the channel. In wireline communication system, the channel can be optical fibers, coaxial cables, or printed-circuit board (PCB) traces. All of them demonstrate loss at high frequency. The low-pass channel responses can significantly disperse the signal waveform and induce ISI, which needs to be equalized to avoid inducing errors. For common wireline transmitters, an FFE, essentially multiple
output drivers with programmable coefficients that are driven by data with different delay, is applied as part of equalizing filter to overcome the effect of ISI. The basics of channel equalization will be introduced in section 2.3.

2.1.2 Transmitter Performance Metrics

In wireline communication systems, one of the most important performance metrics is bit error rate (BER) [32]. BER is defined as the number of bit errors for a given unit time. Typical wireline specifications require BER to be lower than $10^{-12}$. BER depends on several factors including received signal swing, ISI, jitter, and noise. Within a transceiver, a transmitter shapes the received signal waveform and is thus critical for achieving a low overall BER.

The performance of a transmitter is measured by its output waveform. To evaluate the quality of transmitted signal, an eye diagram, which folds a time-domain waveform into one or several bit periods as shown in Fig. 2.2, is commonly used. Fig. 2.3 shows an example of a measured transmitter output eye [20] and highlights the measurements that are commonly used to evaluate the performance of a transmitter. These measurements include:

1) **Voltage Swing**: The voltage swing of transmitter output signals is critical because it determines the amplitude of received signal and sets a requirement for receiver sensitivity [33].
Since the power of transmitter is dominated by building blocks that operates at full data rate, increasing the required output voltage swing unavoidably leads to larger transmitter power consumption. Applying a smaller transmitter output voltage swing lowers the transmitter power but demands a better receiver design and can result in higher receiver power. The transmitter output voltage swing and the receiver sensitivity trade off with each other. Typical transmitter voltage swing falls in the range between 600 mV to 1 V (differential peak-to-peak). For example, the specified peak-to-peak differential voltage swing for PCI Express Gen. 3.0 is 800 mV [34].

![Eye Diagram Illustration](image)

**Fig. 2.2:** Illustration of eye diagram
2) **Inner Eye Opening**: The inner eye opening is an indication on the effective bandwidth and noise of a transmitter. Insufficient bandwidth of transmitter blocks such as MUX, predriver, and output driver can induce additional ISI and lower the inner eye opening. Although insufficient transmitter bandwidth can be modeled as part of the channel loss and compensated with equalizers, it still reduces the receiver eye opening and thus BER. Typical serial-link transceiver systems specify signaling/jitter compliance masks which measured transmitter eye diagrams must satisfy. The interested readers can refer to Table I of reference [35] for a summary of transmitter eye mask for several standards of electrical and optical communication systems.

3) **Jitter**: Jitter of transmitter output reduces the eye width of received signal and the timing margin for error-free sampling. Depending on the shape of the received signal, the transmitter
output jitter can vary the pulse width and induce voltage noise leading to a reduction of overall BER. The transmitter jitter can be categorized into deterministic jitter (DJ) and random jitter (RJ) [36]. The main difference between DJ and RJ is that DJ is confined within certain time while RJ is not confined and demonstrate a Gaussian distribution. The sources of transmitter DJ include data dependent jitter due to the finite transmitter bandwidth and supply induced jitter. The RJ of transmitter, on the other hand, comes mostly from the phase noise of the PLL. Generally speaking, the effect of RJ dominates DJ due to the unrestricted nature. To model the unconfined nature of RJ, typical standards estimate the effect of RJ by multiplying the rms jitter value with a factor of $\pm 7$ for a target BER of $10^{-12}$. As a reference, for the PCI Express Gen. 3.0 specification, the DJ and RJ of data jitter of transmitter should be lower than 18 ps and 2.23 ps, rms, respectively. Note that although both transmitter and receiver jitter degrade BER, transmitter jitter is more critical because it modulates the pulse width, which effectively produces amplitude noise [37].

4) Noise: The noise labeled in Fig. 2.3 include both electrical noise and ISI. The requirement of electrical noise of transmitters, in general, is less critical than receivers because of the high signal-to-noise ratio (SNR). As the data propagates through the channel, the noise generated by transmitter is heavily attenuated by the low-pass channel response. Therefore, the equivalent noise power is much lower than that of the receiver front-end. The ISI is discussed in the metric of inner
eye opening.

Except for the output waveform, the other commonly required measurement for serial-link transmitters is the return loss response. The return loss response is the measured single-port S parameter ($S_{11}$ or $S_{22}$) at transmitter output by a network analyzer. The return loss response indicates the amount of reflection due to impedance mismatch. Fig. 2.4 shows an example of a simulated return loss response of an inductorless transmitter. To minimize the impact of the reflection on signal integrity, common specifications require return loss responses to be lower than -8 to -10 dB across the entire frequency range from DC to the Nyquist frequency. Fig. 2.5 shows an example of return loss mask of PCI Express Gen. 3.0 [34]. Achieving the target return loss response specification is an important challenge for 60+Gb/s transmitters, which will be discussed in detail in section 3.2.

Fig. 2.4: Simulated return loss response of an inductorless transmitter
2.2 Channel Equalization and Feedforward Equalizer (FFE)

In this section, we review the basics of channel equalization, introduce several ways of implementing an equalizer, and discuss the design requirements and challenges of transmitter-side equalizer, FFE. Section 2.2.1 begins with an introduction on the wireline channel and equalization. Section 2.2.2 to Section 2.2.4 introduce several commonly used equalizers including CTLE, DFE, and FFE. Section 2.2.5 discusses the design requirements and challenges of transmitter-side equalizer, FFE, at data rates of 60+ Gb/s.

2.2.1 Channel Equalization Basics

Wireline channels, such as optical fibers, coaxial cables, twist-pairs, or PCB traces, demonstrate frequency-dependent insertion loss due to skin effect, dielectric loss, and coupling
loss. Depending on the quality of channels and the target data rate, these high frequency losses can filter out the high-frequency portion of the signal and limit the maximum achievable data rate. For example, Fig. 2.6 shows the frequency response of a 0.5-meter, twist-pair channel. As can be seen, the channel behaves as a low-pass filter, which demonstrates a 25-dB loss at 30 GHz. To connect the transmitter output to the channel, bondwires or bumps are necessary, which can cause additional 3 to 5 dB loss at 30 GHz. The other source of signal degradation is discontinuity. Ideally, a channel should demonstrate a characteristic impedance of 50 \( \Omega \) throughout the transmitter and the receiver. In practice, the characteristic impedance along the signal propagation path can be varied if there is any discontinuity in the physical geometry. Except for that, resistance variations on transmitter-/receiver-side termination can also deviate the impedance from 50\( \Omega \). These discontinuities can cause reflections, which degrade signal amplitude, reduce rise time, and increase jitter.

Fig. 2.6: An example of insertion loss (S\(_{21}\)) of a twist-pair channel
To evaluate the effect of channel loss and reflection, one effective way is to simulate the pulse response at the receiver input, and use the information to find out the impact of one particular pulse on the previous and subsequent bits, which are referred as intersymbol interference (ISI). To demonstrate this approach, Fig. 2.7 shows a pulse response at the receiver input of a 0.5-meter, twist-pair operating at 62.5 Gb/s. The pulse demonstrates a peak of 0.3 V followed by two glitches caused by reflections. To demonstrate the impact of the high-frequency channel loss, a zoom-in version is plotted in Fig. 2.8(a). The pulse is severely dispersed resulting in a width that is much wider than 1 UI. To evaluate the impact of the present bit to previous and former bits, Fig. 2.8(b) shows the sampled pulse response, which is obtained with an ideal receiver sampling at the frequency equal to the data rate. The sample with the maximum value is defined as the main cursor. The other samples that demonstrate residual values are referred as ISI. The ISI can be categorized into pre-cursor and post-cursor depending on whether the sampling point happens before or after the main cursor. One simple way to estimate the impact of ISI on eye opening and BER is to find out the worse-case (WC) eye opening under ISI by the equation [38]:

$$V_{\text{eye, opening}} = V_{\text{main}} - \sum_{k=-\infty}^{\infty} |V_{\text{ISI}, k}|,$$

where $V_{\text{main}}$ is the value of the main cursor and $V_{\text{ISI}}$ is the ISI. As can be seen, any ISI can
directly reduce the eye height of received signal or even close the eye. Taking the channel of Fig. 2.8 as an example, the simulated WC eye opening is -308 mV, indicating a completely closed eye.

![Fig. 2.7: The pulse response of a twist-pair channel at 62.5 Gb/s](image)

To minimize the effects of ISI and open the eye, equalizers are commonly applied at both transmitter and receiver sides. The performance of equalizers directly affect the eye opening and BER and is thus critical in serial link design. In the literature, equalizers are implemented in there
different forms, continuous-time linear equalizer (CTLE) [39-41], decision feedback equalizer (DFE) [40-43], and feedforward equalizer (FFE). The next three subsections briefly introduce CTLE, DFE, and FFE, respectively.

2.2.2 CTLE

A CTLE, usually implemented at the receiver front-end, is essentially a continuous-time high-pass filter that counters act the loss-pass channel response. Fig. 2.9(a) and (b) show one of most commonly used CTLE structures and the corresponding frequency response. The circuit is a differential pair with an additional source-degenerated resistor and capacitor pair. The degeneration RC pair inserts a low-frequency zero, $z_1$, which leads to a high-pass frequency response that flattens the overall frequency response across a wide frequency range. As long as the channel demonstrates linear frequency response (-20 dB/dec) at high frequency, the CTLE can eliminate the ISI completely. For example, for on-chip interconnections, since the frequency response of on-chip, resistive wires is dominated by a single low-frequency pole, it can be completed equalized with a single CTLE (See Appendix A). A detail discussion on how to design the CTLE can be found in [39].

The main advantage of CTLE is the low equalization power. Since a CTLE is not different than an amplifier except for the pole-zero doublet, as long as the receiver needs an amplifier, by
Fig. 2.9: (a) Source-degenerated CTLE, and (b) the corresponding frequency response replacing the amplifier with a CTLE, it does not induce additional power. Note that as technologies continue to scale down, the unit-current-gain bandwidth ($f_T$) of transistors for state-of-the-art technologies can exceed 200 GHz easily, which facilitates low-power amplifier design. In other words, it is energy efficient to apply multiple amplifiers at the receiver side as compared with increasing the transmitter output swing for advance technologies. Unfortunately, since few channels demonstrate linear frequency responses, CTLE along is typically insufficient to eliminate all the ISI. Additional equalizers, i.e., FFE and DFE, are required in most cases.

Since a CTLE is typically placed at the first stage of the receiver front-end, the gain and the noise performances are critical. Since a CTLE suppresses low-frequency signal and induces
additional noise, it degrades the signal-to-noise ratio unavoidably. However, since a CTLE amplifies high-frequency signal, it can potentially suppress the high-frequency noise of subsequent stages. Interested readers may refer to [33], which discusses the impact of CTLE gain and noise on receiver sensitivity, and [39], which presents an analysis on CTLE power/noise trade-offs.

2.2.3 DFE

Unlike CTLE, of which the equalization capability and the flexibility of CTLE are limited by the realizable frequency responses, DFE is a more sophisticated and effective way of equalization. A common DFE structure is shown in Fig. 2.10 (a). DFE eliminates the ISI by taking the decisions of previous bits and utilizing the decision to subtract the input signal with proper weightings. To preserve the previous decisions, a number of D Flip-flops (DFF) are required, and the feedback paths are implemented with amplifiers with variable gain, which is controlled by high-resolution DACs. The feedback signals are combined in a summer, which is then sampled by the decision making comparator. A DFE that cancels N subsequent ISIs is typically referred as an N-tap DFE. Fig. 2.10(b) shows the pulse response after an ideal 3-tap DFE being applied to our target twisted-pair channel. As can be seen, the three samples that follow the main cursor are completely cancelled out by the DFE, improving the worse-case eye opening ($V_{\text{eye,opening}}$) from -308 mV to -56 mV. DFE is very effective in canceling ISI and is the only possible way to deal with reflections.
Fig. 2.10: (a) An example of common DFE structure, and (b) the sampled pulse response after applying an ideal 3-tap DFE

[44]. DFE is usually required in stringent link environment with large high frequency loss and reflection.

Despite of all the advantages, one of the key design challenges of DFE is how to close the feedback loop within a timing margin of 1 bit period (UI). As shown in Fig. 2.10, after the comparator samples a data, it has to regenerate within a short period and drive the first amplifier that subtracts the input with an amplitude that is equal to the first post cursor. To cancel the first post cursor ISI completely, the feedback signal sent to the summer has to settle before the next decision. This requires a short setup time and clock-to-q delay for the sampler and a high bandwidth and low delay summer. Unfortunately, for channels with long tail pulse responses, many taps are required leading to a large number of feedback amplifiers. These amplifiers can induce
large parasitic capacitance, lower the bandwidth of summer, and induce large power consumption. For 65-nm CMOS technology, the speed of conventional 1-tap DFE is limited to 21 Gb/s [21] [44].

The other drawback of DFE is that it cannot handle precursor ISI since the value of the next bit is unpredictable. The precursor ISI is usually not small and can impact the eye opening significantly. For example, even with 3-taps of DFE, in pulse response of Fig. 2.10(b), the WC eye opening is -56mV and the eye is closed. Among all the ISI, the first precursor ISI has a value of 0.15 V which eats up 48% of the eye opening. Note that if the effect of precursor ISI can be cancelled out completely, the worse-case eye opening can be improved from -56 mV to 92 mV. This result indicates that cancelling the precursor ISI is very important. In the next section, we will introduce the basics of FFE, which is one of the most common ways of handling precursor ISI.

2.2.4 FFE

FFE, sometimes referred as finite impulse response (FIR) filter, is a commonly used equalizer that features good capability of eliminating ISI and low power cost (as compared with DFE). Although FFE can be implemented in both transmitter and receiver, it is more common to realize an FFE in the transmitter side for noise and linearity concerns. Figure 2.11 shows a common FFE structure. An FFE reduces the ISI by delaying data by multiple of bit periods ($T_b$) and summing them with weighted coefficients. The summation is usually achieved in current-domain by
applying multiple gm-based driving cells, and the coefficients are adjusted digitally by tuning the tail current of each driver with a digital-to-analog converter (DAC). The total number of driver cells is referred as the number of taps. An FFE with N drivers is commonly called an N-tap FFE. In general, the driver cell that has the largest current is called main tap. The other driver cells are referred as precursor or postcursor taps, depending on whether the driven data come latter or earlier than the data of the main tap. The precursor and postcursor taps eliminate precursor and postcursor ISI, respectively. Most transmitter-side FFE incorporates at least 1 pre-cursor tap and 1 post-cursor tap for channel equalization.

![Fig. 2.11 A conventional FFE structure](image)

To illustrate how an FFE equalizes the channel response, Fig. 2.12(a) and (b) compares the pulse responses before and after the channel of two designs, one applying a 3-tap FFE (with one pre-cursor and one post-cursor tap) while the other does not. As can be seen, instead of generating one single pulse, by delaying the data and creating two negative pulses before and after the original
pulse, a large amount of ISI is cancelled. The WC eye opening can be improved from -233mV to 20mV.

Fig. 2.12 Pulse responses (a) before the channel, and (b) after the channel for transmitters with no FFE / 3-tap FFE

Except for pulse responses, the other common way to see how an FFE works is to utilize the frequency response. The frequency response of an N-tap FFE with coefficient of $\alpha_1$ to $\alpha_N$ can be derived as:

$$H(j\omega) = \sum_{k=1}^{N} \alpha_k * e^{-j\omega T_p}$$
where $T_b$ is the bit period. The frequency response is repetitive for every $1/T_b$ due to the nature of FIR filter. Fig. 2.13 shows the frequency response from DC to the Nyquist frequency of the previous example. As can be seen, the FFE is effectively a high-pass filter with a 21-dB DC suppression that counteracts the low-pass channel response. Based on this observation, an FFE suppresses the DC component rather than amplifies the high-frequency component. Lowering the low-frequency signal components effectively enhances the effect of noise at the low-frequency portion. This phenomenon is called noise enhancement. At the transmitter output, since the signal-to-noise ratio (SNR) is large and is thus less of a concern, the impact of noise enhancement for a transmitter-side FFE is not severe. Note that the impact of noise enhancement can be more severe if the FFE is realized at the receiver.

To implementing an FFE, a circuit block that can generate delay of $T_b$ is required. In addition,
to add up the signals with different weightings, a summation circuit with programmable gain for each tap is needed. This leads to two possible implementations as shown in Fig. 2.14(a) and (b), respectively. The FFE structure of Fig. 2.14(a) is used for transmitter side, of which the required $T_b$ delay is generated by using DFFs. The structure of Fig. 2.14(b), on the other hand, produces the delay by using sample-and-hold (S/H) amplifiers. Realizing the FFE at the receiver side has the advantage of having the potential to adjust coefficients adaptively. However, since realizing high-speed DFFs is much easier than S/H amplifiers, it is more common to implement an FFE at transmitter. Other advantages of transmitter-side FFE include better noise performance due to the high SNR and facilitating the design of receiver-side clock recovery circuits.

![Fig. 2.14 Implementations of FFE on (a) transmitter side, and (b) receiver side](image)

Fig. 2.14 Implementations of FFE on (a) transmitter side, and (b) receiver side

Fig. 2.15 shows a common implementation of a summer for an N-tap FFE, which consists of N differential pairs that achieves precise summation in the current domain. The weighting of each
tap is adjusted by tuning the tail current with additional biasing circuits of which the current are set by current digital-to-analog converters. The output of the FFE summer is connected to the pin and the channel. In order to satisfy the impedance matching requirement, the termination resistors, $R_{t}$, are set to 50 Ohm. Although it is possible to have an additional output driver stage to lower the impact of parasitic capacitance of FFE driver buffers on the output bandwidth [45], nonlinearity of the output driver can distort the signal. The other common ways of implementing an FFE summer is to utilize voltage-mode drivers, which can potentially save four times of power in output drivers. The finite inverter FO4 delay, however, limits the data rate of FFE using voltage-mode drivers to less than 20 Gb/s. Interested readers can refer to [46] and [47] for more information regarding to voltage-mode FFE.

![Common implementation of FFE summer](image)

**Fig. 2.15 Common implemenation of FFE summer**

### 2.2.5 Requirements and Challenges of FFE at 60+Gb/s

As discussed in the previous sections, CTLE, DFE, and FFE all have different characteristics.
Although the requirement on each equalizer is channel dependent, typical transceivers apply at least 3-taps of FFE at the transmitter side, at least one stage of CTLE, and several taps (3 to 15 taps) of DFE at the receiver side. As we increase the data rate, the equalization requirements is higher due to the higher channel loss as suggested in the frequency response of Fig. 2.6. To evaluate the FFE requirement for our 60+Gb/s transmitter, we develop a system evaluation tool that derives the receiver WC eye opening based on the pulse response of Fig. 2.7. Here, we assume the data rate is equal to 64-Gb/s and the receiver applies 3-tap of DFE (which is feasible at 64-Gb/s by using dynamic latch technique as shown in [43]). Then, we sweep the number of FFE taps and evaluate the corresponding receiver eye opening. The coefficients is derived based on the eye-max algorithm as shown in [38], which optimizes the receiver eye opening at the target sampling time. An alternative solution is to use least-mean-square (LMS) algorithm, which results in slightly smaller receiver eye opening. Fig. 2.16 show the receiver eye opening under different FFE taps, which also include the results of applying different pre-cursor taps. As can be seen from this figure, using one precursor tap is the most efficient solution because it leads to better eye opening under the same total FFE taps. As we increase the number of FFE taps, the eye opening will continue to grow up and eventually saturate at around 4 taps. These result suggests that 4-taps of FFE, of which one is pre-cursor tap and two is post-cursor taps, is required for our 60+Gb/s transmitter design.
Unfortunately, the 4-tap FFE requirement poses severe difficulties at 60+Gb/s transmitter design. First of all, as a full-rate DFF can hardly work at such a high data rate, the transmitter has to apply DFFs at a lower data rate to produce the required delay and to utilize multiple multiplexers, as shown in Fig. 2.17, to produce full-rate data streams. As will be discussed in the next section, the power consumption of multiplexer at high data rate is very high due to the large self-loading capacitance. Except for the multiplexers, several predrivers and output drivers are also required for generating the target voltage swing. Since a total of N data streams are required, all of these power required to generate one full-rate data stream have to multiply by a factor of N to realize an N-tap FFE. This problem fundamentally limits the number of FFE taps. As demonstrated in Fig. 1.4., for state-of-the-art transmitters operating at data rate beyond 40 Gb/s, the FFE is limited to 2 taps. Note that this problem can be resolved by applying a delay-line-based FFE, which will be
Except for the large power consumption, the other design challenge of high-speed FFE is the limited insertion and return loss bandwidth at the transmitter output node due to the drain capacitance of driver buffers. To produce large signal swing, wide transistors are required for the FFE driver buffers, which results in large drain parasitic capacitance. As multiple of driver buffers are required as shown in Fig. 2.15, the parasitic capacitance is getting larger and we increase the number of FFE taps. This parasitic capacitance, combined with ESD devices and output PAD, can lower the bandwidth of transmitter significantly. Fig. 2.18(a) and (b) shows a simple model of the transmitter output network and the corresponding $C_{par}$ and bandwidth under different FFE taps. As can be seen, for a 4-tap FFE, the corresponding bandwidth is lower than 14 GHz, which can
Fig. 2.18 A possible FFE Implementation at high data rate

deteriorates both insertion and return loss response significantly. This problem can be improved by distributing $C_{\text{par}}$ using multiple inductors, which will be discussed in detail in section 3.2.

2.3 Serializer Architecture and Multiplexer Design

A serializer is an essential building block that enables high-speed serial communication. A serializer takes a large number of low-speed and combines them into a high-speed one. The essential building blocks of serializers are multiplexers (MUX) and latches. Multiplexers utilizing clock signals to select the desired output from one of multiple input to produce output stream with a higher data rate. Latches are placed in front of multiplexers to provide delay for input data and to ensure proper timing for the multiplexing. This section introduces the basics of serializer design. We begin with an introduction on the most common serializer architecture, 2:1 multiplexing tree, in section 2.3.1. Section 2.3.2 introduces the key building block of a serializer, the multiplexer. In
section 2.3.3, we discuss the design challenges of the serializer of our target 60+Gb/s serial-link transmitter and present a quarter-rate architecture that can potentially achieve a higher data rate and lower power consumption.

### 2.3.1 Serializer Architecture

One of the common architecture is to use a 2:1 multiplexing tree as shown in Fig. 2.19. The serializer structure consist of multiple 2:1 MUX stages. The data rate of the output of each 2:1 MUX stage is increased by a factor of two, and each 2:1 MUX stage drives another 2:1 MUX stage until it produces the full-rate data. The 2:1 MUX stage and the corresponding waveforms are shown in Fig. 2.20(a) and (b), respectively. Here, $D_A$ and $D_B$ are data generated by 2:1 MUXs of previous stages and they have the same transition timing because the previous MUXs have the same clock input. The 2:1 MUX takes the clock and two data signals and implements the multiplexing function, which can be expressed as:

$$D_{out} = D_{in,1} \cdot CK_{in} + D_{in,2} \cdot \overline{CK_{in}},$$

where $D_{in,1}$ and $D_{in,2}$ are the logic values of the two data stream. To ensure proper timing relationship between $D_{in,1}/D_{in,2}$ and $CK_{in}$, two latches, $L_1$ and $L_2$, are inserted to minimize the impact of the delay of the frequency divider and the clock-to-q delay of the previous MUX, and
one additional latch, \( L_3 \), is inserted to delay the data of \( D_{in,2} \) by half UI to provide better timing for the clock.

![Diagram of 2:1 MUX tree architecture](image)

**Fig. 2.19** Conventional 2:1 MUX tree architecture

![Diagram of 2:1 MUX stage and waveforms](image)

**Fig. 2.20** The 2:1 MUX stage and the corresponding waveforms

This serializer architecture is ubiquitously used for the following reasons. First of all, for all MUXs operating at the same data rate, this architecture needs only one pair of complementary clocks, which facilitates the design of the corresponding frequency divider. The pulse width of the
MUX output is determined by the duty cycle of the clocks, which should be designed to be 50%. Any duty cycle distortion can lead to unbalanced eye width, which can degrade the BER. As will be discussed later, this requirement is easier to achieve as compared with the other possible approach, which controls the pulse width by using the phase difference between multiphase clocks. The quality of output data using this approach typically has better quality and is less unbalanced. One of the main drawbacks of this architecture is the large number of latches, which is unavoidable due to the fact that only one pair of complementary clocks are available.

One way to reduce the number of latches is to apply multiphase sampling architecture as shown in Fig. 2.21 [24]. In this architecture, instead of spacing the timing of MUX inputs with a latch, the timing is adjusted by utilizing clocks with multiple phases during the multiplexing of the previous stage. For example, in Fig. 2.21, the required timing delay of half UI at the final 2:1 MUX is achieved by utilizing the four-phase clocks at the previous stage. The additional latch that required to generate the half-UI delay is thus eliminated. Overall, this approach can eliminate up to 43% of latches required, which saves power and area significantly.

The main concern of this approach, however, is that multiphase clocks are required at low frequencies, which increases the design difficulties of the frequency dividers. The conventional way of implementing a multiphase divider using multiple static dividers has the problems of phase
uncertainty because the initial condition of the dividers are unknown. This problem can be resolved by inserting additional reset signal or applying multipath injection-locked divider. Interested reader can refer to [48] for a multiphase frequency divider without phase uncertainty problem.

Fig. 2.21 Serializer architecture using multiphase sampling

2.3.2 Multiplexer

A MUX takes multiple input data and generates a single high-speed data stream by selecting which data to output depending on the information of the clock. A MUX is effectively a logic gate that realizes sum of products. Fig. 2.22 shows two of the most common MUX implementations. The MUX of Fig. 2.22(a) uses two transmission gates (T-gate) controlled by a pair of complementary clocks to select the output value. A T-gate MUX has the advantages of being low-
power, low-area, and compatible with CMOS logics. T-gate MUXs are widely used at low data rates and small dimensions of gate can be used to minimize the power consumption. The drawback of the T-gate MUX is the long charging/discharging delay. In 65-nm CMOS technology, the speed of T-gate MUX is limited to design that has output data rate below 16 Gb/s.

![T-gate MUX and CML MUX](image)

Fig. 2.22 Common implementations of MUX: (a) T-gate MUX, and (b) CML MUX

To operate at data rates beyond 16 Gb/s, a different logic family is required. One of the most common implementations is to use current-mode logic (CML) as shown in Fig. 2.22(b). Here, the input data and clocks switch the current flow and the output data is generated by transferring the current into voltage through the loading resistor, $R_L$. This circuit structure manifests several advantages. First, when the transistors are flowing current, it is operated in the saturation region, which suggests better driving strength and smaller transistor sizes. Thus, a lower output
capacitance and higher bandwidth can be achieved. In general, the bandwidth of CML MUX is determined by the output RC time constant, which can be increased by choosing a smaller resistance of $R_L$ with the cost of larger current and power. Even with a small loading resistance and large power, the bandwidth of CML MUX is fundamentally limited by the self-loading capacitance, which is determined by transistor $f_T$. A common way to extend the bandwidth further is to apply on-chip inductors with the cost of area [9] [30]. In 65-nm CMOS technology, a CML MUX without and with peaking inductors can operate up to 24 and 40 Gb/s, respectively.

2.3.3 Design Challenges of 60+Gb/s Serializers

The serializer design is challenging as the output data rate goes beyond 40 Gb/s because of the stringent timing constraints and limited bandwidth. This section discusses the timing and bandwidth requirement in section 2.3.3.1 and 2.3.3.2, respectively. An architecture that applies quarter-rate, directly 4:1 multiplexing to relax the timing constraint and the power issues, is presented in section 2.3.3.3.

2.3.3.1 Timing Constraints

A serializer consists of clocked circuits. Similar to digital systems, the timing constraints must be satisfied at all times to guarantee the correct functionality. The critical timing path and the
corresponding timing diagrams of an example 2:1 MUX stage is shown in Fig. 2.23. Here, the transition timing of the MUX output, \( D_a \), is determined by the delay of the frequency divider from \( CK_{in} \) to \( CK_{in,2} (t_{div}) \) and the clock-to-q delay of the 2:1 MUX \( (t_{ck-q}) \). Since the latch needs to samples \( D_a \) with the clock \( CK_{in} \), both setup and hold time constraints must be satisfied. These requirement fundamentally set the operating range of the MUX stage. A complete derivation of the timing constraints can be found in [49], which shows that as the data rate increases, the reduced bit period leads to lower timing margin and limits the maximum achievable data rate. In reality, as all the delays and setup/hold time of latches are subject to PVT variations, it is difficult to cover all PVT corners with a fix design.

Fig. 2.23 Critical timing path and the corresponding timing diagram of the MUX

One solution to this problem is to make the phase of the latch input tunable, which can be achieved by the architecture of Fig. 2.24. Here, a digital-to-phase converter (DPC) takes \( CK_{in} \) and
generates the output clock with tunable phase. With the assistance of the DPC, it is possible for designers to correct the phase relationship across PVT corners. This DPC, however, consumes a large amount of power because it operates at a high frequency. For example, a DPC needs to provide with a 32-GHz clock if used in a 64-Gb/s transmitter. At a frequency of 32 GHz, it is inevitable to incorporate on-chip inductors to form an LC tank to sustain sufficient gain and output swing, which occupy a large area.

The other possible solution is to apply additional delay matching buffers or phase matching loop. These method, however, has similar power and area issues. Interested readers can refer to [8], [16], and [19] for more details.

2.3.3.2 Bandwidth Limitations

A serializer consists of a large number of MUXs and latches that operate at high data rates.
As the data rate increases, the bandwidth requirements for these MUXs and latches are increased, which can consume a large power consumption and fundamentally limit the maximum achievable data rate.

To illustrate this problem, Fig. 2.25 shows the block diagram of a serializer using conventional 2:1 MUX tree targeting on generating a 64-Gb/s data stream. As can be seen, a large number of MUXs and latches operating at data rates of 16, 32, and 64 Gb/s are required. As discussed in the previous section, since low-power T-gate MUXs can barely operate at data rates of 10 Gb/s, CML MUXs and latches are unavoidably the only choices at these data rates. To estimate the power consumption of CML MUXs, a simple N:1 MUX model based on the circuit shown in Fig. 2.26(a) is used. The power required for driving a loading capacitance, $C_L$, can be derived as:

$$Power = \frac{0.5 \cdot \omega_{3dB} \cdot V_{ov} \cdot A_{DC} \cdot C_L}{1 - N \cdot A_{DC} \cdot \frac{\omega_{3dB}}{\omega_T}},$$

where $A_{DC}$ is the target DC gain, $\omega_{3dB}$ is the target 3dB bandwidth, $\omega_T$ is the unit-current-gain frequency of the NMOS transistors, $V_{ov}$ is the overdrive voltage of the NMOS transistors, and $N$ is the fan-in of the MUX. Based on this equation, the power consumption versus data rates curves for different Ns are plotted in Fig. 2.26(b). As can be seen, the curve is linear at low data rate, and
as the data rate increases, the power consumption grows up drastically and as the denominator term approaches 0, resulting in infinite power consumption. One easy way to interpret this power curve is that, as the target data rate rises, the bandwidth of a multiplexer is limit by the self-loading capacitance eventually, which corresponds to the denominator term of the power equation. Despite derived for CML MUXs, CMLs latch will demonstrate similar trend. We can therefore see that as multiple stages of MUXs and latches operating at data rates beyond 30 Gb/s are required, a huge power consumption is inevitable. The only possible way to reduce the power consumption is to reduce the number of MUXs and latches, and right-shift the power curves of Fig. 2.26(b)

Fig. 2.25 Conventional serializer architecture for 64-Gb/s transmitters
Fig. 2.26 (a) Conventional CML MUX circuit for power estimation, and (b) the power consumption under different multiplexing ratio

2.3.3.3 Quarter-rate Architecture

Based on our discussions, the main design challenges for high-speed serializers include severe timing constraints and large power consumption due to the limited bandwidth. To resolve these problems, researchers proposed an architecture that uses quarter-rate multiplexing at the final stage. [22] [49] [50]. Figure 2.27(a) shows an example of 64-Gb/s serializer using quarter-rate architecture. Here, a 4:1 MUX rather than multiple 2:1 MUXs is incorporated to generate the full-rate, 64-Gb/s data directly from four 16-Gb/s data. As shown in the timing diagram of Fig. 2.27(b), the full-rate data is generated by ANDing two consecutive clocks that has the same frequency while separates by 90-degree in phase to implement a direct 4:1 MUX. Multiphase multiplexing, as discussed in section 2.3.1, is applied to further reduce the number of 16-Gb/s latches. To
demonstrate the advantages of using direct 4:1 MUX at final stage, Fig. 2.28 compares this architecture with a conventional 2:1 MUX tree architecture. One important advantage of applying a 4:1 MUX at the final stage is that, since the data to be multiplexed is at a data rate of 16 Gb/s rather than 32 Gb/s, the bit period is doubled, which provides with additional timing margin. It is therefore possible to generate the full-rate data across PVT corners without additional phase tuning mechanism. The other advantage is that this architecture eliminates three 32-Gb/s latches, two 32-Gb/s MUXs, and a large number of latches at 16 Gb/s by skipping 32 Gb/s and applying multiphase multiplexing. This suggests significant power and area saving as these high-speed building blocks consume significant amount of power consumption and need peaking inductors due to the severe bandwidth limitation.

Fig. 2.27 (a) Quarter-rate, direct 4:1 MUX architecture, and (b) timing diagrams
Despite all the advantages of the quarter-rate architecture, there are some design challenges involved using a 4:1 MUX at final stage. First of all, although a single 4:1 MUX replaces with a large number of high-speed latches and MUXs, the design of a 4:1 MUX, as suggested in Fig. 2.26(b), is much challenging than a 2:1 MUX due to the larger self-loading capacitances. Even with a state-of-the-art 4:1 MUX [49], the maximum data rate can barely achieve 48 Gb/s. The other important design concern is that, unlike the conventional 2:1 MUX tree structure, of which the pulse width is determined by the duty cycle of clock, the pulse width of the output of 4:1 MUX is determined by the phase difference between the two consecutive clocks. In other words, four clocks with no quadrature phase errors are required. Any phase deviation between 0 and 90 degree clocks can result in an uneven eye width for odd/even data, which can reduce the worst-case eye.
opening. As discussed in the previous section, it is much easier to sustain a 50% duty cycle than ensure the phase relationship between a 4-phase clocks. Additional phase tuning to ensure the phase relationship of quadrature clocks is necessary.

In this dissertation, a serializer based on the quarter-rate architecture is proposed. To overcome the design issues of the 4:1 MUX, we propose a high-speed, low-power 4:1 MUX that operates at data rates beyond 60 Gb/s. The problem of quadrature phase mismatch is addressed in the clock generator. The proposed 4:1 MUX and the corresponding clock generator are going to be presented and discussed in Chapter 4.

2.4 Digital-to-Phase Converter Basics

A digital-to-phase converter (DPC) generates one or more clock signals with digitally-controlled output phases from multi-phase input clocks. DPCs have many applications in wireline and wireless communication systems. Except for being applied into a serializer to calibrate the timing of a MUX stage, a very common application of DPC is in a phase interpolator (PI)-based clock and data recovery circuit [10] [51-55] in serial-link receivers as shown in Fig. 2.29(a). A digital loop filter (DLF) and the DPC forms a feedback loop to determine the optimized sampling phase of the sampling clock, \( CK_{out} \). Another common application of a DPC in receivers is
monitoring the size of the received data eye (eye-opening monitor circuit, EOM) [49] [56-57]. Figure 2.29(b) shows an example in which by sweeping the reference voltage of $V_{H}$ and $V_{L}$ and the sampling timing of $CK_1$ and $CK_2$, the EOM can capture a two-dimensional map of the eye diagram without external test equipment. An EOM can also be used to provide information for an adaptive equalizer or a clock-and-data recovery circuit (CDR) to determine the optimized equalizer coefficients and sampling phase [57].

Fig. 2.29 Applications of DPC in (a) PI-based CDR of serial-link receiver, (b) on-chip eye-opening monitor (EOM)

In these applications, important characteristics of a DPC are phase linearity, resolution, and frequency range. In a CDR application, the linearity and resolution impact the jitter performance of the received clock. The phase quantization steps introduce added jitter as the output clock is tracking a desired frequency and phase [58-61]. The minimum step-size is determined by both the resolution and the nonlinearity. In closed-loop systems, severe nonlinearity can potentially cause
a loop to false-lock at local minima. A common target for PCI Express and USB 3.0 systems is a max-to-min step ratio of 2.5 for 6-bit phase resolution [62-63]. The frequency range is an important metric because data links often need to satisfy multiple data rate standards or different input reference frequencies [62-65]. While it is generally not difficult to achieve the DPC phase linearity requirement for a single clock frequency, designs that support a large frequency range can become very challenging.

There are many ways to implement a DPC. One possible way is to implement a delay-locked loop (DLL) and to use a multiplexer to select the clock with the desired [66-67] as shown in Fig. 2.30(a). This approach demonstrates excellent phase linearity because the delay variations between different delay stages are usually small. One of the main drawbacks is that a large number of voltage-controlled delay lines (VCDL) is needed to realize a high resolution DPC, which can be very power hungry. The other drawback is that the frequency range is limited by the tuning range of the delay of VCDLs and it is not suitable for DPCs operating at both high- and low-frequencies. In practice, the resolution of a DLL-based DPC can hardly exceed 4 bit.

A more common DPC architecture a phase interpolator (PI) as shown in Fig. 2.30(b). The architecture involves a multiphase clock generator, an N:2 MUX, and a phase interpolator. In this architecture, the MUX selects two clocks with consecutive phases from an N-phase clock generator
Fig. 2.30 Applications of DPC in (a) PI-based CDR of serial-link receiver, (b) on-chip eye-opening monitor (EOM)

and a digital PI mixes them to produce the digitally controlled output phase. By proper design of the phase selector, the DPC achieves full 360-degree phase range. This range enables continuous phase rotation that is often required in CDR applications using DLLs [67]. The time resolution of the DPC is determined by the performance of the PI and the number of phases generated by the clock generator. The phase linearity of conventional PI is not satisfactory. Two possible implementations of PIs are shown in Fig. 2.31. The CML-based PI, as shown in Fig. 2.31(a), manifests good linearity and better bandwidth. The inverter-based PI, as shown in Fig. 2.31(b), has lower power but the linearity is notoriously poor. Unfortunately, for DPCs that uses a CML-based PI, the resolution can hardly meet 6-bit.
Fig. 2.31 (a) CML-based and, (b) inverter-based PI structure

To overcome this problem, a DPC design using harmonic rejection (HR) filter to improve the PI linearity is proposed. The HR filter corresponds with the passive low-pass filter form a bandpass filter that creates sinusoidal waves for phase interpolating, which significantly improve phase linearity. An inverter-based PI structure is further applied to improve the power and compatibility. The proposed DPC is going to be presented and discussed in Chapter 5.
CHAPTER 3

A 4-Tap, 64-Gb/s FFE Using LC-Ladder Filter

As discussed in section 2.2, as the data rate exceeds 40 Gb/s, the design of the FFE is more difficult. First, it is difficult to generate 1-UI delay at high speed. Many designs replicate the multiplexing tree for each delayed signal which requires multiple high-speed MUXs operating at full rate. Even though recent research indicates that running a single MUX at 60 Gb/s is possible [16], the power consumption increases linearly with the number of taps and can become impractical. Second, signals with different delay are usually combined in current domain by multiple output drivers. These drivers need to provide sufficient output current and when combined with ESD can lead to large parasitic capacitances. The total parasitic capacitance degrades the bandwidth and return loss of the transmitter considerably. A common solution is to introduce inductors to tune out the impact of the capacitance at the cost of area.

This chapter introduces an FFE and output-combiner implementation using LC-ladder filters that reduces the power consumption per tap and improves the bandwidth with minimized number of inductors. Section 3.1 introduces the delay-line-based FFE architecture, which allows significant power saving with the cost of area. Section 3.2 discusses the implementation of the
output combiner network and how the use of LC-ladder filter reduces the number of inductive elements while maintaining desired response. Section 3.3 applies the filter implementation to the delay element. Section 3.4 ends this chapter by comparing the area cost of the approach and discussing the potential design issues.

3.1 Delay-Line-Based FFE Architecture

At high data rate, generating delayed data can be power hungry. The intuitive way of generating delayed data with full-rate D-flipflops (DFF) leads to climbing power for bit periods below FO4 inverter delay (roughly 30 Gb/s for 65-nm CMOS technology). Rather, a possible solution is to implement multiple serializers and to generate the delay data at lower speed as shown in Fig. 3.1(a) [4] [13] [21] [68]. This approach extends the maximum data rate of FFE since a MUX has a higher bandwidth limit. However, because all of the building blocks, including the N:1 MUX, the clock buffer, and the predriver, have to be duplicated, the power cost for each additional tap is still large. As shown in Fig. 3(a), even with the proposed low-power serializer (to be discussed in detail in Section IV.A), a power of 48 mW/tap is required for a 64-Gb/s design in our 65-nm standard CMOS technology. Although this problem can potentially be alleviated by sizing each serializer according to the coefficients so as to reduce the power accordingly, this solution is limited in the range of different channels that the equalizer can be applied. Moreover,
as the devices meet the minimum size constrain of a technology, the improvement in power consumption is limited. As shown in Fig. 1.4, for data rates above 40 Gb/s in a similar technology, the maximum number of taps including the data tap is often only two.

Fig. 3.1  Circuit structures and power comparison of FFE implemented using (a) multiple MUXs, and (b) delay lines

One alternative approach to extend the number of taps that exhibits good potential is to use a delay-line-based FFE structure as shown in Fig. 3.1(b) [9] [55] [69-72]. The delay is generated by an analog delay line. Since the only building block that has a tap dependent power is the analog delay line, the power per tap can be reduced if the power consumed by the analog delay line is
lower than the entire multiplexing tree. Instead of using a large number of power-hungry CML buffers [9], section 3.3 proposes a design that consumes only 12 mW per delay element at 64 Gb/s. The tables in Fig. 3.1 compare the power consumption of the delay-line-based FFE with the conventional approach. Since one set of serializer is still required for the delay-line-based FFE, the power benefit is apparent for a large number of FFE taps.

With the low power per tap of the delay-line-based structure, it is conceptually possible to implement an FFE with a large number of taps under a reasonable power budget. In reality, the existing delay-line-based FFE structures pose several design issues that limits the maximum number of taps. First, most linear circuits that are capable of providing with up to 1-UI delay involve passive elements, i.e., inductors, and area of the implementation is a big concern. Designs such as [55], [70], and [71], need at least 28 inductors for 3 taps. Second, increasing the number of taps requires more stages of output drivers that output signals with different weights. The device sizes of these drivers need to be large enough to provide sufficient output swing for highly attenuating channels. These drivers introduce large parasitic capacitances at the output node, and when combining with parasitic capacitances of ESD devices and the pad, these capacitances can limit the bandwidth and deteriorate the return loss response significantly. Finally, because the delay lines are inherently linear circuits and have finite bandwidth, each delay stage can reshape the
pulse responses and induce additional ISI. For the last several taps of which the output signals have to pass through several stages of delay lines, this effect is more severe. For the latter taps, the increased ISIs can lessen the capability of equalization, which limits a practical number of taps to less than 6.

In the following sections, we presents a new FFE architecture based on an LC-ladder filter to relax the above inherent issues. The FFE incorporates an output combiner network and multiple analog delay lines, which are presented in section 3.2 and 3.3, respectively. Section 3.4 compares the area of the proposed approach and discusses several design issues of the proposed delay-line-based FFE.

3.2 LC-Ladder-Filter-Based Distributed Combiner Network

As mentioned previously, the total capacitance at the output of the transmitter can limit the data bandwidth. For instance, a CML output driver as shown in Fig. 3.2(a) has a pole, $p_{\text{TX, out}}$, at the output node:

$$p_{\text{TX, out}} = \frac{1}{(R_L||Z_o)\cdot C_{\text{par}}},$$

where $R_L$ and $Z_o$ (transmission line impedance) are generally matched and designed to be 50 ohms and the lump parasitic output capacitance is $C_{\text{par}}$. At high data rate, the pole $p_{\text{TX, out}}$ disperses
the signal and introduces ISI. Although the loss due to the pole can be modeled as part of channel and compensated with an equalizer, the additional loss unavoidably reduces the eye opening. The pole also deteriorates the return loss response ($S_{11}$) at high frequency. Typical transmitters require the magnitude of the return loss response at the Nyquist frequency to be lower than -10 dB [28][34]. To meet the specification with a single pole system, the frequency of $p_{\text{TX, out}}$ has to be $>3x$ higher than the Nyquist frequency. Such high $p_{\text{TX, out}}$ is very difficult since the ESD portion alone of $C_{\text{par}}$ to satisfy human body model (HBM) protection level of 1600V needs to be $>220$ fF [73-74]. The other major source of $C_{\text{par}}$ is the output capacitances of FFE output drivers. The total device size depends on the required signal swing and is proportional to the number of taps. Figure 3.2(b) shows the frequency of $p_{\text{TX, out}}$ under different $C_{\text{par}}$. For a 4-tap FFE design, when $C_{\text{par}}$ is

![Small-signal model of conventional CML driver](image1)

![Frequency of $p_{\text{TX, out}}$ under different $C_{\text{par}}$](image2)

Fig. 3.2 (a) Small-signal model of conventional CML driver, and (b) the frequency of $p_{\text{TX, out}}$ under different $C_{\text{par}}$
higher than 500 fF, the corresponding bandwidth is lower than 13 GHz, which corresponds to an 8.6-dB insertion loss and a -0.65-dB return loss at Nyquist frequency for 64 Gb/s.

Before discussion on the LC-ladder-filter approach, we first examine the effectiveness and limitations of distributing $C_{\text{par}}$ in an LC delay line as shown in Fig. 3.3(a) [55] [69-73]. To emulate the behavior of an ideal transmission line (T-line), the inductors are designed based on the characteristic impedance of transmission line:

$$Z_o = R_L = \frac{L}{\sqrt{C}}$$  \hspace{1cm} (3.1)

where $Z_o$ and $R_L$ are designed to be 50 ohm and $C_{\text{par}}$ is distributed uniformly. However, since lumped LC networks result in ripple in the frequency response, a large number of segments are needed to mimic a good T-line leading to large area cost even if assuming that $C_{\text{par}}$ can be divided accordingly. To illustrate, we plot the return loss responses for different number of segments in Fig. 3.3(b) for a target $C_{\text{par}}$ of 560 fF. To achieve lower than -10 dB $S_{11}$ at 32 GHz, more than 4 inductors (8 for differential signaling) are required.

To extend the bandwidth of the return loss response, our approach instead distributes $C_{\text{par}}$ based on the classical filter theory. To begin with, the circuit of the LC delay line of Fig. 3.3(a) resembles a 9th-order, doubly-terminated LC-ladder low-pass filter (LPF) as shown in Fig. 3.4(a). Even
Fig. 3.3 (a) Output combiner network using LC delay line approach, and (b) the return loss response of RC-only driver and several LC delay lines with different number of inductors though the LC-ladder filter has non-uniform inductances and capacitances, the frequency response is accurately defined. We can thus use the LC-ladder filter to optimize the return loss response of the LC distribution network.

Fig. 3.4 (a) The circuit structure of a 9th-order, doubly-terminated LC-ladder filter, and (b) its associated two-port model

To outline this approach, first, the LC-ladder LPF is modeled as a two-port network as shown
in Fig. 5(b), of which the first port corresponds to the channel and the second port, corresponds to the last-tap FFE driver. Assuming low-loss LC elements in the network, the return loss of transmitter ($S_{11}$) can be obtained by $S_{21}$ as:

$$|S_{11}| = 1 - |S_{21}|^2$$  \hspace{1cm} (3.2)

where $S_{21}$ can be accurately obtained by designing the LC network as an LC-ladder filter. It is important to note that since most LC-ladder filters\(^1\) are symmetric, $S_{21}$ and $S_{12}$ are equal. This suggests that optimizing the transfer function from the last-tap FFE driver to the transmitter output ($S_{12}$) equivalently optimizes the return loss response ($S_{11}$). To implement a target $S_{12}$ as an LC-ladder filter, the L and C values can be designed by using filter tables that included LC values for a normalized filter ($R_L=1$, $\omega_{3\text{dB}}=1$ rad/s) as shown in Fig. 3.5(a) [29], and applying the frequency and impedance scaling equations:

$$L_i = \frac{R_L L_i'}{\omega_{3\text{dB}}}$$  \hspace{1cm} (3.3)

\(^1\) Based on the filter tables of [75], the doubly-terminated LC-ladder filter is symmetric for odd order Butterworth and Chebyshev response. However, the LC-ladder filter may not be symmetric for other responses such as Bessel response.
\[ C_i = \frac{c_i'}{R_L*\omega_{3dB}} \]  

(3.4)

where \( L_i' \) and \( C_i' \) are normalized inductances and capacitances, \( \omega_{3dB} \) is the target cut-off frequency, and \( R_L \) is the termination resistance (50 ohm for the output-combiner network). Since a portion of \( C_{par} \), i.e., the capacitances due to ESD devices, can be flexibly placed, \( C_{par} \) can be distributed based on these values to synthesize a target \( S_{21} \) and thus \( S_{11} \) responses. An example is shown in Fig. 3.6, where the capacitances due to the pad and the four taps of equally sized FFE drivers are separated by four inductors and the remaining capacitances due to the ESD diodes are distributed based on the LC-ladder filter design. Note that we attempted to minimize adding capacitance that is not inherent to the design since the total capacitance impacts the achievable \( \omega_{3dB} \).
Fig. 3.5 (a) LC-ladder filter table for Butterworth, Chebyshev, and Bessel responses, and (b) component values after applying frequency and impedance scaling ($R_L=50$, $\omega_{3dB}=52$ GHz)

Fig. 3.6 Design of the proposed LC-ladder-filter-based output combiner network

Different types of filters result in different LC values and demonstrate different frequency and return loss responses. Tables of different filter types are available [75]. Some filter types result in
inductances and capacitances that are either too large or too small to be useful. As general guidance, we target filters with large normalized capacitances since filters with higher $C_i'$ lead to designs that are capable of handling higher $C_{par}$ under the same $\omega_{3dB}$. In other words, for a target $C_{par}$, a higher cut-off frequency and wider bandwidth can be achieved. Smaller values of $L_i'$ are also preferred to reduce area but not as important since the increase in area cost is not proportional to an increase in inductance of an on-chip inductor. Figure 3.5(b) lists the LC values of five different filters that target on the same cut-off frequencies of 52 GHz: Butterworth, Chebyshev type-I with 0.01, 0.1, and 0.5-dB ripple, and Bessel LPFs. As can be seen in Fig. 3.7(a), Chebyshev filters demonstrate much larger total capacitances than Butterworth and Bessel filter. As the ripple of Chebyshev filter increases, the resulting total capacitance is higher, which indicates that Chebyshev filter with a large passband ripple is very effective in distributing $C_{par}$. To compare the frequency and return loss responses, Fig. 3.7(c) and (d) show the $S_{21}$ and $S_{11}$ of the five different types of filter responses, respectively. As compared with Bessel and Butterworth LPFs, Chebyshev LPFs demonstrate steeper $S_{21}$ and $S_{11}$ around the cut-off frequency, making them suitable for broadband design. As expected and plotted in Fig. 3.7(c) and (d), there is a strong relationship between the peak values of $S_{11}$ and the passband ripple of $S_{12}$. The return-loss, $S_{11}$, requirement (-10 dB) constrains the passband ripple (0.46 dB) which constrains the total capacitance that can be distributed within the
structure. In our design, we choose a 9th-order Chebyshev LPF with 0.1 dB passband ripple with LC values shown in Fig. 3.8(a). Figure 3.8(b) plots the simulated $S_{11}$ of our proposed LC-ladder filter. The $S_{11}$ demonstrates a ripple response with maximum of -16.4 dB in the passband. The return loss response differs from the ideal $S_{11}$ in Fig. 3.7(a) due to the finite Q of inductors and is lower than -10 dB up to 57 GHz. For comparison, the $S_{11}$ of the LC delay line under the same $C_{par}$ is also plotted in Fig. 3.8(b). As can be seen, applying LC-ladder filter improves the bandwidth of return loss response by up to 2.7x.

![Graphs showing filter responses](image)

Fig. 3.7 (a) Total capacitances of different types of filter for a target cut-off frequency of 52 GHz, (b) peak values of $S_{11}$ of Chebyshev LPF for different $S_{21}$ passband ripple requirements, and (c) $S_{11}$ and (d) $S_{22}$ of different types of filter responses.
Fig. 3.8 (a) LC values of the proposed LC-ladder-filter-based combiner network, and (b) comparison of the simulated return loss responses for a uniform LC delay line and the proposed approach

Except for frequency and return loss responses, we compare the pulse responses between the LC-ladder filter approach and the uniform LC. Since these structures have essentially multi-input and single-output, the pulse responses of the taps closer to the output cannot be easily derived. Instead, we compare them based on simulation. The results in Fig. 3.9(a) and (b) for the design of Fig. 8(a) show that the LC-ladder-filter approach achieves comparable or even better pulse response than LC delay line approach. As part of the comparison, we also mimic the impact of a channel by simulating with a single pole filter that has 8.6-dB loss at Nyquist frequency. The resulting pulse responses after the channel for the two designs are plotted in Fig. 3.9(c) and (d), respectively. Fig. 3.10(a) lists the peak value, the worst-case (WC) ISI, the WC eye opening, and
the WC eye opening assuming a 3-tap DFE is applied, for each of the pulse responses\(^2\). As expected, the proposed LC-ladder LPF demonstrates higher WC eye openings for the last tap because the transfer function of the last tap is optimized. For the remaining taps, the proposed LC-ladder LPF exhibits comparable or even better responses than LC delay line approach. These results confirm that the LC-ladder filter approach is an effective solution.

It is important to note from Fig. 3.9 that as we apply the LC network to distribute Cpar, each LC segment induces an additional delay of Δt. The delay values are listed in Fig. 3.10(b). The delay plays an important role as part of the entire FFE design as discussed in the next subsection. Finally, similar to other ESD distribution approaches [76], the series resistance of inductors can degrade the ESD performances and must be sufficiently low. In our design, the series resistance of the inductors is approximately 1.4 Ohm. To achieve better ESD performance, a wider wire width of inductors can be selected to lower the series resistance with the cost of larger area.

\(^2\) We assume the receiver samples at the peak. The ISI is estimated by the summation of absolute values of the following samples with 64 GHz clock. The WC eye opening is derived by subtracting the peak value with the WC ISI.
Fig. 3.9 Pulse responses (a) and (b) are without a channel for the proposed LC-ladder filter, and uniform LC delay line. Pulse responses (c) and (d) are after a channel.

<table>
<thead>
<tr>
<th></th>
<th>Proposed LC-ladder LPF</th>
<th>LC Delay Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tap1 (V)</td>
<td>0.662</td>
<td>0.690</td>
</tr>
<tr>
<td>Tap2 (V)</td>
<td>0.659</td>
<td>0.688</td>
</tr>
<tr>
<td>Tap3 (V)</td>
<td>0.655</td>
<td>0.651</td>
</tr>
<tr>
<td>Tap4 (V)</td>
<td>0.664</td>
<td>0.642</td>
</tr>
<tr>
<td>WC ISI (V)</td>
<td>0.356</td>
<td>0.443</td>
</tr>
<tr>
<td>Tap1 (ps)</td>
<td>0.360</td>
<td>0.353</td>
</tr>
<tr>
<td>Tap2 (ps)</td>
<td>0.366</td>
<td>0.354</td>
</tr>
<tr>
<td>Tap3 (ps)</td>
<td>0.345</td>
<td>0.370</td>
</tr>
<tr>
<td>Tap4 (ps)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WC Eye opening w/ DFE (V)</td>
<td>0.305</td>
<td>0.246</td>
</tr>
<tr>
<td>WC Eye opening w/ DFE (V)</td>
<td>0.300</td>
<td>0.335</td>
</tr>
<tr>
<td>WC Eye opening w/ DFE (V)</td>
<td>0.288</td>
<td>0.297</td>
</tr>
<tr>
<td>WC Eye opening w/ 3-tap DFE (V)</td>
<td>0.319</td>
<td>0.272</td>
</tr>
<tr>
<td>Tap1-2 Delay (ps)</td>
<td>6.38</td>
<td>5.97</td>
</tr>
<tr>
<td>Tap2-3 Delay (ps)</td>
<td>6.20</td>
<td>5.86</td>
</tr>
<tr>
<td>Tap3-4 Delay (ps)</td>
<td>5.27</td>
<td>6.88</td>
</tr>
</tbody>
</table>

(a) (b)

Fig. 3.10 (a) Table of the peak output voltage, worst-case (WC) ISI, worst-case (WC) eye opening without DFE, and worst-case (WC) eye opening with 3-tap DFE for both the proposed LC-ladder LPF and a uniform LC delay line, and (b) delay for both structures.
3.3 FFE Delay Line Using LC-Ladder Filter

The analog delay line is a critical block in the delay-line-based FFE design. An ideal analog delay line has to provide sufficient delay while preserves the signal waveform. For linear circuits, the delay and bandwidth trade off with each other inevitably. Lowering the delay line bandwidth results in larger delay while introducing additional ISI. On the other hand, designing a high bandwidth delay line leads to insufficient delay.

First, instead of requiring a large delay for the delay line, the proposed architecture leverages the inherent delay within the LC-ladder-filter-based combiner. For 64 Gb/s, the combiner introduces an additional delay that is roughly equal to 0.4 UI $^3$ ($\Delta t = 6.2$ ps) between each consecutive tap. By propagating the analog delay line in the opposite direction as the combiner filter as shown in Fig. 3.11 [69-71], the required delay of each analog delay segment is reduced to $T_b - \Delta t$.

---

$^3$ Here, we target on total delay of 16.7 ps for data rate of 60 Gb/s to cover the 50-64 Gb/s data rate.
In a 65-nm technology, a CML buffer can be designed to have a delay of only 2.6 ps for a bandwidth of 52 GHz. To build a delay line of 0.6 UI (10.5 ps) requires up to 4 such stages leading to large power dissipation. Our design uses buffered LC segments [55] [72] to implement the delay. We implement the LC segments using a filter design perspective similar to the design of output-combiner. The LC values are sized based on equation (3.3) and (3.4). Unlike the combiner network, $R_s$ and $R_l$ do not need to be 50 ohm in this circuit, and the resistances can be sized according to the bandwidth requirement. Another key difference is that a designer does not have the flexibility of ESD capacitance to spread within the LC network to adjust the filter capacitances.

Similar to the output combiner network design, we implement the LC network as a doubly-terminated, 5th-order Chebyshev LC-ladder LPF with 0.1 dB passband ripple and cut-off frequency of 55 GHz. We first recognize that the bandwidth of the LC network can be improved by splitting...
the components of the total loading of each stage, $C_L$, within the LC network. $C_L$ comprises of the loading of the buffer of the next delay line segment ($C_{DL}$) and the output driver ($C_{Drv}$). In our architecture, as shown in Fig. 3.12(a), two-inductor, doubly-terminated LC distribution network is used as an analog delay line. $C_{Drv}$ is moved to node $V_X$ in between the two inductors. The equivalent small-signal model is shown in Fig. 3.12(b). We chose a 5\textsuperscript{th}-order Chebyshev LPF because the capacitance $C_3$ is 1.7x higher than $C_1$ and $C_5$, which is close to the ratio between $C_{Drv}$ and $C_{DL}$ because in practice, $C_{Drv}$ is determined by the required signal swing and is typically larger than $C_{DL}$. By using the inherent $C_{Drv}$ and $C_{DL}$ as the filter elements instead of adding capacitances, we maintain high bandwidth without needing additional area for inductors. Figure 3.13(a) and (b) shows the simulated frequency and pulse responses for node $V_{out}$ and $V_X$, respectively. The delay line achieves a bandwidth of 51.5 GHz at node $V_{out}$ and 65 GHz at node $V_X$. Overall, splitting $C_{DL}$ and $C_{Drv}$ with a Chebyshev LPF extends the bandwidth by 2.1X as compared to a uniform LC delay line approach. Each delay line consumes 12 mW which is the same power as in [55] while the delay line of [55] targets a data rate of 40 Gb/s.
Fig. 3.12 (a) Proposed LC-ladder-filter based delay element that splits $C_{DRV}$ and $C_{DL}$, and (b) the corresponding small-signal model.

Fig. 3.13 (a) Simulated frequency responses, and (b) pulse responses of node $V_X$ and $V_{out}$ for the proposed LC-ladder-filter-based delay line.

While using two inductors per segment is area efficient, we explore the impact of using a higher order filter on bandwidth and delay. Figure 3.14(a) shows the circuit examples of $3^{rd}$, $5^{th}$, and $7^{th}$-order LC-ladder LPFs. The corresponding delay of Chebyshev filters with 0.1-dB ripple under different cut-off frequencies are plotted in Fig. 3.14(b). For high-order filters, since the delay...
is larger due to the large number of LC segments, the required bandwidth for achieving the same delay is higher. As can be seen, to achieve a target delay of 10.5 ps, the required bandwidths for the 3rd, 5th, and 7th-order filters are 26, 56, and 87 GHz, respectively. Although the 3rd-order LPF manifests the advantage of requiring only one inductor, the low bandwidth can lead to significant ISI. The 9th-order LPF, on the contrary, results in a design with very high bandwidth. However, the lower filter termination for the high bandwidth dissipates more power. Based on the above observations, a 5th-order LC-ladder LPF with two inductors is a good compromise between area, bandwidth, and power for our target UI.

![3rd-, 5th-, and 7th-order LC-ladder LPF](image)

**Fig. 3.14** (a) 3rd-, 5th-, and 7th-order LC-ladder LPF, (b) the delay of the three filters under different bandwidth
Although we choose the same design of delay lines for all different taps, one important design issue of the analog delay line is that, as the current of the combiner is programmable, the gate capacitance \( (C_3) \) and thus the delay are dependent on the coefficients. This effect, however, is minimized in this design because not only is the equalizer not very sensitive to delay variations, but also a double-terminated LC-ladder filter has low sensitivity to component variation [75]. Simulation shows that changing the current of the output driver from 0 to 10 mA varies the capacitance value from 50 to 75 fF. As illustrated in Fig. 3.15, this change corresponds to less than 1-ps delay variation.

![Graph showing delay variations for different CDrv](image)

**Fig. 3.15 Delay variations for different CDrv**

### 3.4 FFE Architecture and Performance Discussions

In this section, we present the overall FFE architecture and discuss several design
considerations of the proposed FFE. The FFE architecture and the area comparison to other published architecture are presented in section 3.4.1 and section 3.4.2, respectively. In section 3.4.3, we discuss and analyze the effect of having non-1-UI delay of FFE. This section ends with a discussion on the practical limit on the achievable data rate of the proposed FFE architecture.

3.4.1 FFE Architecture

The overall architecture of the proposed LC-ladder-based FFE is shown in Fig. 3.16(a). To cover a wider range of channel responses, the polarity of the last tap is made programmable. The combiners of the first three taps and the last tap are shown in Fig. 3.16(b) and (c), respectively. The FFE coefficients are adjusted by controlling the current of the combiner. For the last-tap combiner, the width of the differential pairs is designed to be half of that of other combiners, reducing the maximum coefficient of the last-tap by a half.

Figure 3.17 illustrates the simulated pulse responses for different taps. The first tap demonstrates a differential signal swing of 670 mV. The tap spacing from tap 1 to tap 4 is 16.1, 

\[ \text{tap spacing} = 16.1 \]

\[ \text{differential signal swing} = 670 \text{ mV} \]

\[ \text{coefficients} \]

\[ \text{current control} \]

\[ \text{width of differential pairs} \]

\[ \text{half of other combiners} \]

\[ \text{maximum coefficient reduced by a half} \]

\[ \text{achievable voltage swing of last tap halved} \]

\[ \text{For the pulse response simulation, we use the same driver buffer for all four taps. In our design, the achievable voltage swing of the last tap is halved of other taps.} \]

77
Fig. 3.16 (a) Architecture of the proposed FFE and, the driver buffers of (b) the first three taps and (c) the last tap.

14.9, and 14.8 ps, respectively. The signal swing for each following tap is attenuated by approximately 1.2 dB due to a large signal gain that is < 1. The lower swing for latter taps indicates reduced eye opening when the number of pre-cursor taps is increased. To demonstrate the effect of decreased magnitude of latter taps, Fig. 3.18 illustrates the system simulation results of the WC eye opening under different delay line gain for a typical backplane channel. As can be seen, the benefit of having precursor taps diminishes as the delay line gain is lower than 0.9. This comes
from the fact that for FFEs with precursor taps, the low delay-line gain degrades the magnitude of main tap and thus the eye opening, which can eat out the improvement of cancelling precursor ISI. This effect combined with the cost of power and area limit the number of taps to less than 6 and the number of pre-cursor taps to less than 2. This problem is caused by insufficient bandwidth and gain due to the low transistor $f_t$. The problem can be resolved by reducing the data rate and increasing the DC gain of the delay lines by 10%, respectively.

![Fig. 3.17 Simulated pulse responses of the FFE](image)

**3.4.2 Area Comparison**

The area cost of the architecture is dominated by inductors and can be seen from Fig. 3.16(a).

Each additional tap requires 3 x 2 inductors (the number is doubled due to differential signaling), of which 2 x 2 inductors are required for the analog delay line and 1 x 2 inductors for the output combiner. Figure 3.19 compares the number of inductors between the proposed approach and other
publications including using multiple MUXs, CML buffers, and LC delay line in [9], [13], and [55]. In the design using multiple MUXs or using CML buffers as delay elements, in order to operate at high data rate, inductive peaking for each stage is necessary. As can be seen, by using LC-ladder-filter, the required number of inductors of the proposed approach is comparable or lower than other approaches while maintaining sufficient bandwidth for both analog delay lines and the LC combiner network.

Fig. 3.19 Various implementations of the delay line and output combiner and comparison of the number of inductors for each approach

3.4.3 Non-1-UI Delay

One of the fundamental problems of the delay-line-based FFE is that because the delay lines
are inherently linear circuits, the frequency and pulse responses are subject to PVT variations, and the resulting delay can deviate from 1 UI. Although accurate 1-UI delay can be achieved by incorporating varactors inside analog delay lines and implementing a delay-locked loop (DLL) \[55\], the excess capacitances can lower the achievable bandwidth of the analog delay line. In addition, a DLL needs a replica delay line \[55\] which increases both area and power cost. The impact of small delay deviations on eye opening is not significant. To demonstrate, we simulate the eye openings under different delays for 31 different channels are shown in Fig. 3.20(a). The equalization includes a 4-tap FFE (1 pre-cursor tap and 2 post-cursor taps) and a 3-tap DFE. The coefficients of FFE and DFE are derived based on eyeMax algorithm of \[31\]. Figure 20(b) shows that the worst-case eye-opening reduction for the range of channels under delay variations of ± 0.2
UI to be less than 3%.

Fig. 3.20 Analysis of the sensitivity of the FFE to delay variation across (a) 31 channels and their corresponding frequency responses, (b) Statistics of worst-case eye reduction for 0.2-UI delay deviation

### 3.4.4 Data Rate Limitations

One advantage of the proposed LC-ladder-filter-based approach is the flexibility in porting the design into a different data rate. By simply applying the frequency and impedance scaling equations (3.3) and (3.4), the required LC values can be derived. Note also that the proposed approach is suitable for high-speed design because of the small inductance values. To demonstrate the feasible range of the proposed approach, Fig. 3.21 plots the required inductances for $L_{D,2}$ and $L_{D,4}$ of the analog delay line and $L_2, L_4, L_6,$ and $L_8$ of the output combiner under different data rate. Assuming that the maximum inductance is limited to 1.2 nH, the minimum data rate will then be limited by the inductor of analog delay line to 24 Gb/s. Based on (3.3), the inductance of analog
delay lines can be reduced by choosing a smaller $R_L$. Thus, the fundamental minimum data rate is limited by the $L_4$ and $L_6$ of the output combiner, which is equal to 14 Gb/s.\footnote{It may be better to use DFF-based design at such low data rates for better power and area.}

---

\textbf{Fig. 3.21 Required inductances of the FFE for different data rates}
CHAPTER 4

Design of a 64-Gb/s Serializer Using Direct 4:1 Multiplexing

The achievable data rate using conventional 2:1 MUX tree architecture, as discussed in section 2.3, are limited by the timing constraints and the large number of MUXs and latches. One solution is to apply quarter-rate architecture and to replace the last two stages of 2:1 MUXs with a single 4:1 MUX relaxes the timing constraints, eliminates several latches and MUXs, and reduces the power. This approach, however, requires a direct 4:1 MUX that supports full-rate data.

This chapter introduces a complete serializer design that generates 64-Gb/s full-rate data from 64-bit, 1-Gb/s inputs with an energy efficiency of ~1.6pJ/b. The serializer design focuses on the final-stage 4:1 MUX and the first frequency divider that generates the required four-phase clocks for the 4:1 MUX. The architecture of the entire transmitter, including the serializer, the PRBS generator, and the FFE, is introduced in section 4.1. The key building blocks of the serializer, the final-stage 4:1 MUX and the first divider, are presented in section 4.2. The design of the 64-Gb/s predriver, which interfaces the FFE and the 4:1 MUX, is presented in section 4.3. The remaining parts of the serializer are discussed in [49] and [80].

4.1 Transmitter Architecture
The architecture of the transmitter prototype is shown in Fig. 4.1. It consists of a clock multiplication unit (CMU), a parallel PRBS generator, a 64:1 serializer, and a 4-tap FFE. To broaden the measurable range of data rate, the CMU is clocked with an external clock with wide range of frequency rather than a PLL. The external clock is transformed into differential clocks through an on-chip balun and sent to the divider chain. The first divider is an LC-loaded static divider that generates quadrature phase clocks needed for the 4:1 MUX and the subsequent divider. The remaining dividers take high-speed clocks from the previous divider and generate low-frequency, multiphase clocks for the MUXs and subsequent dividers. The dividers are based on the design in [77].

---

**Fig. 4.1 Block diagram of the transmitter prototype**
The parallel PRBS generator generates 64-bit, 1-Gb/s data with programmable length of $2^7$-1, $2^{15}$-1, $2^{23}$-1, and $2^{31}$-1. To minimize the power and the cost, the PRBS generator is designed based on a procedure that minimizes the number of DFFs as described in [78] and implemented with the standard-cell design approach. The 64-bit PRBS data is sent to the 64:1 multiplexing tree and serialized into one single 64-Gb/s output. In the 64:1 multiplexing tree design, the architecture similar to the design of [49] is applied. First, the 64:16 MUX is implemented by the conventional 2:1 MUX-tree structure to minimize layout complexity. Following that, the subsequent 16:8 and 8:4 MUXs are serialized with multiphase clocks (16-phase, 4-GHz and 8-phase, 8-GHz clocks), which reduce the number of latches by up to 46%. At the last stage, we skip 32 Gb/s and apply a direct 4:1 MUX structure to generate the full-rate, 64-Gb/s data. This structure allows us to eliminate three 32-Gb/s latches, two 32-Gb/s 2:1 MUXs, and the corresponding clocking circuits. In addition, the quarter-rate structure relaxes the timing constraints of the final MUX and eliminates the requirement for an additional phase tuning loop.

The output of the 4:1 MUX is the full-rate data which is buffered by a predriver and sent to the proposed LC-ladder-filter-based 4-tap FFE. The predriver manifests high bandwidth and moderate gain. It increases the voltage swing and lowers the capacitance loading of the 4:1 MUX. In our design, using one predriver stage minimizes the total power. In the next sections, we
describes the key building blocks of the 64-Gb/s serializer, the final-stage 4:1 MUX and the first divider.

4.2 4:1 Multiplexer

The 4:1 multiplexer takes 4 quarter-speed data as input and serializes them into a differential high-speed output signal. The self-loaded drain capacitance at the output limits the signal bandwidth of a MUX. To reduce the large self-loading capacitance, we propose a 4:1 MUX design as shown in Fig. 4.2 [22]. The MUX consists of inductively-peaked loads and four unit-cells, each consisting of two pulse generators that are controlled by an input data pair and two clocks of which the phases are separate by 1 UI. The pulse generators produce data-dependent, 1/UI current pulses, which are combined in current and transformed into voltage through the loads. Each unit cell comprises of 2 series devices to produce the current with the source of the bottom device connected to ground. The inputs are driven by rail-to-rail signal generated with CMOS logic. This arrangement maximizes the overdrive voltage and minimizes the required device sizes leading to small self-loading capacitance. In this design, the capacitance contributed by each pair of driver is approximately 8 fF. The load resistance of the $R_L$ is made tunable by PMOS switches to cover PVT variations as shown in Fig. 4.3(d). Switches are added in series with parallel legs of resistors on top of the inductor. In our technology, the parasitic capacitances of the switches can contribute
up to 30 fF of capacitance on both drain and source sides. As compared with a conventional design [32], by keeping the output isolated from $R_L$ using the inductor $L_{D,1-2}$, the impact of the parasitic capacitances is halved and bandwidth improves by up to 18%.

The pulse generator is shown in Fig. 4.3(a) and consists of two logic gates that output $V_F$ and $V_R$ driving the stacked transistor $M_6$ and $M_7$ respectively. The pulse generator implements the logic function as follows:

$$I_{out} = \overline{CK_{in,1}} \overline{CK_{in,2}} \cdot D_{in}.$$  \hspace{1cm} (4.1)

The timing diagram of the operation of the pulse generator is shown in Fig. 4.3(b). First, when $CK_{in,1}$ and $CK_{in,2}$ are high, both $V_R$ and $V_F$ are discharged to low and $I_{out}$ is zero. When $CK_{in,1}$ goes low, then depending on $D_{in}$, $V_F$ either stays connected to ground or is pre-charged to $V_{DD}$. If $V_F$ is
pre-charged to $V_{DD}$, as $CK_{in,2}$ falls low, both $M_6$ and $M_7$ are turned on and the current $I_{out}$ flows to the output. After one bit period, the current is turned off as $CK_{in,1}$ rises high and discharges $V_F$. Both $V_R$ and $V_F$ are reset to low as $CK_{in,2}$ goes high again. The proposed 4:1 MUX has several features. First, the waveform of the current pulse depends mostly on the rising edge of $V_R$ and the falling edge of $V_F$, which corresponds to transistor $M_2$ and $M_3$, respectively. As compared with other published 4:1 MUX such as [22], [49], and [50], the proposed design avoids stacked devices in the critical paths. The two transition edges are further sharpened by skewing the P:N ratio of the logic gates and sizing $M_2$ and $M_3$ larger, which reduces the rise and fall time of $V_R$ and $V_F$ by 13% and 43%, respectively. The non-stacked, skewed structure reduces the equivalent logical effort of the gate and achieves a higher fan-out under the same timing requirement leading to lower power. Figure 4.3(c) compares the power consumption of different MUXs. By embedding the latching, eliminating additional latches, and easing the timing requirement, the proposed MUX structure achieves >35% power than previous designs [49].

While having a number of design advantages, a design using a 4:1 MUX is sensitive to quadrature phase mismatch. As shown in Fig. 4.4(a), due to the short bit period, any phase mismatch between I and Q clocks ($\Delta \phi_{IQ}$) can vary the eye width of even/odd data considerably.
Fig. 4.3 (a) Circuit diagram of the pulse generator, (b) timing diagram of its operation, (c) power comparison with other designs, and (d) conventional and proposed tunable resistor design.

The eye diagrams of Fig. 4.4(b) demonstrate the eye reduction due to quadrature phase mismatch.

Phase calibration of the quadrature clocks is needed. In our design, the divider after the on-chip balun for the input clock is implemented as a static divider with an LC resonator similar to the quadrature clock generator for an on-chip oscillator as shown in Fig. 4.5. The locking range is extended by a 2-bit capacitor array for coarse tuning and a pair of varactors that are driven by two separate 7-bit DACs for fine tuning. To adjust the quadrature alignment of the output phases, the control voltages of the varactors allow each of the quadrature output phases to be adjusted by up
Fig. 4.4 The 4:1 MUX design (a) timing diagram under quadrature phase mismatch, (b) output eye diagram under a 5-ps quadrature phase mismatch.

Fig. 4.5 The balun and first divider design to 26° independently. Although not implemented in this prototype, closing the phase control loop
can be done with feedback as shown in [81].

A second issue with a current MUX without a tail current source is that the output signals can exhibit large unwanted ripple when the input clocks are skewed in duty cycle. Figure 4.6(a) illustrates the mechanism of the ripple. The falling edge of current pulse $I_{out,1}$ depends on the rising edge of $CK_{in,1}$ while the rising edge of the next current pulse $I_{out,2}$ depends on the falling edge of $CK_{in,3}$. Duty cycle distortion creates an overlap or underlap between the two current pulses and causes a periodic data-dependent current ripple. The problem can become more severe as the rise/fall time of $V_R$ and $V_F$ are unequal due to PVT corners. Figure 4.6(c) and (d) show the eye diagram under 40% and 60% duty cycle distortions, respectively. In our design, the duty cycle distortion is minimized by biasing the DC voltage of the input of clock buffers with resistive feedback, coupling the clock signal capacitively, and inserting a cross-coupled pair of inverters after the clock buffers as shown in Fig. 4.6(b), which results in less than 4.4% duty cycle distortion across PVT corners. The rise/fall time of $V_R/V_F$ demonstrate less than 3.3% variations across PVT corners, resulting less than 0.1% pulse width variations. Except for that, the common-mode rejection of the predriver and the subsequent analog delay lines also reduce the common-mode portion of the ripple considerably. Simulation shows that the MUX, the predriver, and the first delay line output demonstrates worst-case common-mode ripples of 13.63, 11.84, and 7.5 mVrms,
respectively. The performance of the MUX is summarized in Table

Fig. 4.6 (a) Timing diagram of the 4:1 MUX with duty cycle distortion, (b) clock drivers for the 4:1 MUX, and output eye diagrams with (c) overlap and (d) underlap

<table>
<thead>
<tr>
<th>Differential Signal Swing</th>
<th>Bandwidth</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>880 mV_{P2P}</td>
<td>40.1 GHz</td>
</tr>
<tr>
<td>(t_{rise}) of (V_R) (20%(\rightarrow)80%)</td>
<td>(t_{fall}) of (V_F) (20%(\rightarrow)80%)</td>
<td>(t_{rise/fall}) of (D_{out}) (20%(\rightarrow)80%)</td>
</tr>
<tr>
<td>7.9 ps</td>
<td>8.3 ps</td>
<td>6.9 ps</td>
</tr>
</tbody>
</table>

Table 4.1 Performance Summary of the Proposed 4:1 MUX
4.3 Predriver

The predriver buffers the output of the MUX to the output driver. This buffering permits (a) lower fanout for the MUX without excessive power and (b) sizing of the driver devices for the desired output current independent to the bandwidth of the MUX. In this design, the common-mode rejection of the predriver also helps reducing the common-mode ripple from the MUX. The predriver is implemented as a differential pair with resistive and inductive loading to provide a small gain and extend the bandwidth. A design issue from using standard CMOS technology is the variations of the loading resistance leading to both gain and bandwidth variations. To satisfy the requirement across PVT, design can be over-designed by 30%.

Figure 4.7(a) shows our predriver design that allows tunable load resistance. Switches are added in series with parallel legs of resistors on top of the inductor. In our technology, the parasitic capacitances of the switches can contribute up to 30 fF of capacitance on both drain and source sides. Hence by keeping the output isolated from \( R_L \) using the inductor \( L_{D,1-2} \), the impact of the parasitic capacitances is halved and bandwidth improves by up to 18%. The inductor is implemented as a spiral inductor and demonstrates inductance of 440 pH, Q of 8.8, and lump capacitance of 4 fF. To further extend the bandwidth by an additional 12%, we apply series-peaking inductors, \( L_{S,1} \) and \( L_{S,2} \) to separate the drain capacitance of the predriver with the gate
capacitance of the subsequent stage. Figure 4.7(b) and (c) compares the performance difference for different arrangements [36]. The bandwidth improves from 39 GHz to 55 GHz, which is further extended to 65 GHz by applying series peaking inductors.

Fig. 4.7 (a) Proposed predriver design, (b) conventional and proposed tunable resistor design, and (c) comparison of the predriver frequency responses for the proposed design with and without series inductor.

Fig. 4.8 (a) Proposed wire inductor design, (b) layout view of the proposed inductor design, (c) the simulation of the quality factor of a single wire and the proposed wire inductor
CHAPTER 5

An Inverter-Based Digital-to-Phase Converter Using Harmonic Rejection

Digital-to-phase converters (DPC) are widely used in both serial-link transmitters and receivers. These applications include transmitter-phase-calibrating loop, clock-and-data recovery, and eye-opening monitor circuit. In most of these applications, a DPC with high resolution and high phase linearity is desired. Conventional DPCs utilize phase interpolators (PI) to increase the resolution. The phase linearity, however, is severely limited by the performance of the PI.

In this chapter, we propose a power and area efficient DPC design by improving the PI design to provide high resolution across a wide frequency range and PVT variations. By applying the proposed harmonic rejection (HR) filter technique, significant linearity improvement is achieved by canceling out the 3rd- and 5th-order harmonics of the phase interpolated signal. An inverter-based combiner structure is adopted to further improve the power and area of the DPC.

In section 5.1, we begin with a discussion on the sources of nonlinearity and indicate that a sinusoidal waveform at the output of the PI can provide high linearity and good immunity to noise. Following that, we introduce a circuit approach that applies a harmonic rejection (HR) filter in a
PI design to produce a sinusoidal output across a broad range of frequencies in section 5.2. This chapter is summarized with a detailed description of circuit implementation of this approach in section 5.3.

5.1 Phase Interpolator Phase Nonlinearity

A common implementation of a digital PI is to apply two arrays of differential pairs with inputs 45-degrees separated in phase. The output phase is controlled by switching the current ratio between the differential pairs of the two arrays. The nonlinearity of a digital PI has both systematic and stochastic components. PIs are similar to a digital-to-analog converter in its sensitivity to the random mismatch of the digitally-controlled unit elements. With proper sizing or calibration to reduce the random mismatch, the linearity of a conventional PI with RC loading has been shown to systematically depend on the rise time (RC time constant) of the interpolator [79] [80]. There is a trade-off between linearity and noise immunity such that a large RC time constant can be used for high linearity but with the filtered signal amplitude results in reduced noise immunity.\(^6\)

\(^6\) As the signal becomes small, systematic coupling of clocks can result in offsets on the crossing edge of clock and deteriorate the phase linearity
Additionally, for a range of input frequencies, the RC load’s pole must track the input/output frequency. Lastly, the RC load is subject to PVT variation. With the use of digital phase interpolation, some of the systematic component can be removed by injecting the proper digital weight from each of the two inputs. Instead of using two weighting coefficients, \( \alpha \) and \( 1-\alpha \), the summation of the two coefficients needs not to be one [81] [82]. However, since the mapping of the input code to the weighting coefficient depends on the RC load, it is very sensitive to PVT and input frequency. Linearity can easily degrade from a design optimized for a single operating point. Even when assuming perfect matching, simulation shows that for an 8-bit DPC using RC-loaded PI, decreasing the frequency by a factor of 2.5 increases the DNL from 0.45 to 1.27 LSB. On the other hand, increasing the frequency by the same factor degrades the magnitude by 39%. In general, the realizable resolution of DPC using RC-loaded PI is limited to about 6 bits.

One can improve upon the linearity by using different load structures. The design in [83] substitutes the RC load with an LC resonator and achieves a measured DNL of \(< 0.2\) LSB targeting 5-bit resolution hence indicating the potential to achieve 7-bit resolution. In another example [84], the authors propose to improve the linearity by controlling the crossing time with a current-integration circuit. The approach has achieved 7-bit resolution with a DNL of 0.75-LSB.

Figure 5.1 shows a circuit model of a PI and can be used to compare the nonlinearity of
various waveform shapes. The model includes a combiner and a filter. The combiner consists of two buffers and an ideal adder that mixes the two input signals, $CK_{in,1}$ and $CK_{in,2}$, with weighting coefficients of $\alpha$ and $1-\alpha$, respectively.\(^7\) The weighting coefficients $\alpha$ and $1-\alpha$ are controlled by input codes and usually hold a linear relationship. The combiner’s output passes through a filter to generate the final output. The type of filter can be a low-pass filter (LPF), integrator, or bandpass filter (BPF) to model the effect of different loads.

Fig. 5.1 General phase interpolator equivalent model

By using an ideal square wave inputs, the output waveforms using a LPF\(^8\), an integrator, or a

\(^7\) As previously noted, the summation of the two coefficients needs not to be one to compensate for any inherent AM-to-PM nonlinearity of the buffers.

\(^8\) The LPF here is implemented by a single-pole RC filter with 3-dB cut-off frequency equal to the clock frequency.
BPF under three different $\alpha$ (0, 0.5, and 1) are shown in Fig. 5.2(a), (b), and (c), respectively. The RC-filtered square wave of Fig. 4(a) corresponds to a LPF. The piecewise linear (PWL) wave corresponds to an integrator, and the sinusoidal wave, a BPF. Ideally, for the case of $\alpha=0.5$, the crossing point should locate at the midpoint between that of $\alpha=0$ and 1. However, a significant deviation can be observed for the square wave as compared with the other two waveforms. The simulated output phase versus $\alpha$ curve for different waveforms are shown in Fig. 5.2(d), respectively. As we can see, the square wave exhibits a deviated phase-$\alpha$ curve, which suggests a high INL/DNL. The PWL output waveform demonstrates an ideal linear curve, implying zero INL/DNL. Finally, the sine wave demonstrates a phase-$\alpha$ curve that almost overlaps with the ideal curve of the PWL, indicating an excellent linearity. The maximum INL of RC-filtered square wave, PWL waveform, and sine wave are 4.30, 0, and 0.46 degree, respectively.
Fig. 5.2 Output waveforms for $\alpha=0, 0.5, 1$ using (a) LPF (RC-filtered square wave), (b) integrator (piecewise linear wave), and (c) BPF (sine wave) given square wave input, (d) phase v.s. $\alpha$ curves for the three different waveforms.

Figure 5.3 shows the INL versus amplitude curve for RC-filtered square wave and sine wave, respectively. As can be seen, although increasing attenuation of LPF improves the linearity of the square wave, the decreased output amplitude is subject to noise or signals coupled into the output. We have also indicated the simulated level of the signal coupled from gate-drain capacitance of the subsequent inverter in Fig. 5.3, which can be as large as 120mV. This coupled signal can induce phase offset and deteriorate the phase linearity as the signal amplitude is small. On the contrary, sine wave achieves much better linearity without sacrificing the immunity to noise and unwanted
coupled signal as the fundamental tone is well-preserved.

![INL versus output signal amplitude curve for RC filtered square wave and sine wave](image)

**Fig. 5.3 INL versus output signal amplitude curve for RC filtered square wave and sine wave**

### 5.2 Phase Interpolator Using Harmonic Rejection Filter

This section presents a circuit technique that applies a harmonic rejection (HR) filter in a PI design to produce a sinusoidal output across a broad range of frequencies. The principle of HR filter is introduced in 5.2.1. Section 5.2.2 and 5.2.3 discuss the design considerations of the analog filter and the frequency range. Section 5.2.4 presents the simulation results of a HR filter-based PI including INL, DNL, and the frequency range.

#### 5.2.1 Principle of Harmonic Rejection Filter

Previous publications [79] have shown improved linearity by increasing the time constant of RC-loaded PI so that the filtered PI output waveform resembles that of an integrator. However,
integrators and heavily filtered signals usually have small signal amplitude or consume large amounts of power. A pure sinusoidal output waveform can be generated using a band-pass filter. As shown by [83], an LC resonator with high quality can serve as an effective filter. This approach, however, requires large area for the passive inductor and is suitable only for high frequencies. Moreover, even though the resonant frequency can be adjusted through varactors, adjustments across a large frequency range lower the quality factor of the LC tank and degrade the phase linearity. Op-amp based filter [85] can alternatively be used. These designs usually contain multiple op-amps, which are power hungry and bandlimit the clock frequency, and are only suitable for low-frequency designs.

This paper proposes a harmonic rejection (HR) filter. HR filter has been proposed in [86] to improve the linearity of RF mixers and it has been extensively used in RF transceiver circuits. Inherently a digital filter, HR filter reveals better rejection to harmonics than analog filters. The block diagram of HR filter is shown in Fig. 5.4(a). The HR filter is implemented by combining 0-, 45-, and 90-degree input clocks with a ratio of 1, $\sqrt{2}$, and 1, respectively. The summation node of the HR filter is modeled as an analog filter which optionally performs any additional frequency shaping. The corresponding PI architecture and the equivalent model are shown in Fig. 5.4(b) and (c), respectively. The principle of the harmonic rejection is illustrated in Fig. 5.5 using phase
diagrams. Figure 5.5(a) shows the phase diagram of the fundamental tone. Here, the three vectors are separated by 45 degree with a length ratio of $1: \sqrt{2}: 1$. For the third-order harmonic (HD3), on the other hand, the phase differences between consecutive vectors are tripled, leading to a phase diagram of Fig. 5.5(b). As can be seen, due to the particular phase and magnitude relationships, at HD3, the three vectors will cancel out with each other completely. The same phenomenon also happens to HD5 as shown in Fig. 5.5(c). This results in a perfect rejection on HD3 and HD5 leaving essentially the fundamental frequency.

Fig. 5.4 (a) Block diagram of HR filter, (b) HR filter-based PI architecture, (c) HR filter-based PI equivalent linear model
Fig. 5.5 Phase diagram of harmonic rejection filter of (a) fundamental tone, (b) HD3, and (c) HD5

The transfer function of the HR filter, $H(j\omega)$, can be represented as

$$H(j\omega) = 1 + \sqrt{2}e^{-j\omega \frac{f_{clk}}{8}} + e^{-j\omega \frac{f_{clk}}{4}}$$  \hspace{1cm} (5.1)$$

where $f_{clk}$ is the clock frequency. The corresponding frequency response is shown in Fig. 5.6.

Because the HR filter is actually a finite impulse response (FIR) filter with unit delay equal to 1/8 of the clock period, the frequency response of an HR filter repeats every $8f_{clk}$. For this reason, the HR filter rejects $8N+3$- and $8N+5$- order harmonics. Figure 5.6 also shows conceptual signal magnitudes under different harmonics. In general, higher-order harmonics demonstrates lower magnitudes. Therefore, after passing through the HR filter, the only harmonics remain are HD7 and HD9. Since the magnitudes of the higher harmonics are usually small, the output is close to an ideal sinusoidal wave.
Another advantage of HR filter is the small power overload. From the frequency response of Fig. 5.6, other than filtering out HD3 and HD5, the HR filter demonstrates 1.63-dB attenuation on the fundamental tone. The source of this attenuation can be observed in Fig. 5.5(a). Although the final output has a phase of $\Phi_2$, the vectors of $\Phi_1$ and $\Phi_3$ are not completely in-phase. Therefore, the power penalty can be derived to be $\frac{2 + \sqrt{2}}{2\sqrt{2}}$ (around 1.20$^9$). In practice, as compared with the advantages it provides, the signal power penalty is not significant.

### 5.2.2 Analog Filter Design Consideration

Since the HR filter already removes much of the undesired harmonics, additional filtering is

---

$^9$ The power penalty in real implementation can be increased if the circuits in the front do not scale proportionally.
not needed. In a circuit implementation, the summation node often introduces additional filtering as shown as the analog filter block in Fig. 5.4(a). The analog filter is effectively in series with the HR filter to provide further rejection to harmonics. Figure 5.7 compares the requirements of the analog filter with and without the HR filter. For a RC-loaded PI, the equivalent filter to reject HD3 would have a very sharp frequency response. With an HR filter and residual harmonic component of HD7 and above, the amplitude of the harmonics is low requiring a low order LPF. Based on our simulation results, the analog filter can be designed by controlling the inherent single-pole at the output of the PI by adding capacitance.

Fig. 5.7 Frequency response of the analog filter (a) without HR filter, (b) with HR filter

Figure 5.8 shows the simulated DNL of PI as a function of the total harmonic distortion (THD) of the output signal. We have assumed that the only harmonic component contributing to THD is HD7. Based on this figure, to achieve a resolution of 8 bit, the HD7 amplitude has to be 25 dB
(5.6%) lower than the fundamental tone to achieve an 8-bit resolution. Generally speaking, the requirement is not difficult to achieve for a single frequency. In case we need a wide frequency range, the design of the LPF can be completed by recursively simulating the magnitude of HD7 and adjusting the pole frequency until the worst-case THD and DNL requirements are achieved.

Fig. 5.8 Simulated DNL as a function of different total harmonic distortion (THD)

5.2.3 Frequency Range

An advantage of the HR filter-based PI is the filter notches scales with the operating frequency while a RC-loaded PI would need its frequency response calibrated to the operating frequency. Equation (5.1) shows the tracking of the frequency response with $f_{\text{clk}}$. To illustrate the impact of scaling the operating frequency, we assume that the analog filter is a single-pole LPF and the transfer function is

$$H(s) = \frac{1}{1 + \frac{s}{p_1}} \quad (5.2)$$
where $p_1$ is the frequency of the pole and is fixed. The clock frequency, $f_{clk}$, is related to $p_1$ by the variable $K_{freq}$ as

$$f_{clk} = K_{freq} p_1 \cdot$$ (5.3)

Figure 5.9(a) shows the frequency responses that are normalized to the input frequency under different $K_{freq}$. Lowering the clock frequency does not affect the attenuations on HD3 and HD5. As expected, the attenuation on HD7 decreases with lower frequency. For the three different $K_{freq}$ of 0.25, 0.5, and 1, the attenuations on HD7 are 7.72, 12.86, and 18.62 dB, respectively. Since the amplitude of HD7 of input clock is typically 40 dB below the fundamental tone, reasonable linearity is still achieved for low frequency.

![Figure 5.9](image)

(a) Overall frequency responses, (b) Output waveforms for different input frequencies

Note however, that the frequency response of the subsequent analog filter is still fixed. Thus, the operating frequency cannot be extended beyond a small range. Figure 5.9(b) shows the output waveforms of HR filter based-PI under low and high frequencies. At low frequency, the HR filter-
based PI also behaves similarly to an RC-loaded PI such that the sharp edges caused by a small
time constant degrade the linearity. At high frequency, the output amplitude reduces with
increasing frequency leading to greater noise sensitivity.

5.2.4 Simulation Results

Two PIs with and without HR filter are compared in simulation. The conventional RC-loaded
PI is designed using current-mode logic (CML) buffers loaded with a single-pole RC LPF. The
LPF has a 3-dB cut-off frequency equal to the input clock frequency which corresponds to the case
of $K_{\text{freq}}=1$ in Fig. 5.9. Figure 5.10 (a) and (b) show the simulated waveforms and FFTs under the
two cases, respectively. As can be seen, by applying the HR filter, the waveform approximates a
pure sinusoidal wave and the magnitudes of HD3 and HD5 have been decreased dramatically. With
HR filter, the overall THD has been improved from 12.6% to 0.7%. Figure 5.10(c) and (d) shows
the simulated INL and DNL. With the HR filter, the INL and DNL have been improved from 2.44-
to 0.28-degree and 0.50- to 0.05-degree, respectively. These results show that a very high
resolution is now achievable with the aid of HR filter.
Fig. 5.9 HR filter-based PI (a) overall frequency responses, (b) output waveforms for different input frequencies.

Fig. 5.10 Simulation results of PI (a) output waveform, (b) FFT, (c) INL, (d) DNL.
To compare the frequency range, we plot the DNL under different frequencies ($K_{\text{freq}}$) and the corresponding frequency range in Fig. 5.11. As can be seen, as the clock frequency decreases, the DNL increases due to the increase of the residual HD7 harmonic. We define the lower bound of the frequency range as the frequency that results in a DNL of 0.5 LSB (0.703 degree for 8-bit resolution). We define the upper bound as the 3-dB cut-off frequency of the analog filter and is equal to $f_{\text{clk}}$ ($K_{\text{freq}}=1$) for both cases. Based on these definitions, the HR filter combined with a single-pole LPF improves the frequency range from 33% to 78%, and the ratio between the upper and lower bound of frequencies improves from 1.28 to 4.4.

Fig. 5.11 DNL versus $\alpha$ curve and the frequency range

5.3 Architecture and Building Blocks
The proposed 8-bit inverter-based DPC architecture is shown in Fig. 5.12. The DPC is separated into two parts—multiphase clock generators and DPC core. The multiphase clock generator consists of two frequency dividers to generate the required 8-phase clocks. The clocks are the inputs to the DPC core to generate output with digitally-controlled phase. The proposed HR filter is implemented by the summation of three phase-shifted paths. Each path includes a phase selector and a PI core. The phase selector selects two phases to be interpolated based on 8-bit control signals from a controller block. Each of the three paths selects different phases for proper harmonic rejection. The phases are inputs to PI cores which are nonlinear inverter-based combiners. Based on the decoded 32-bit thermometer control signals, it generates the output clock throughout nonlinear phase interpolating. To ensure the magnitude relationship between the three different paths, the sizes of the inverters and resistors of the different cores are sized accordingly. The outputs of the three paths are summed and filtered with a LPF to provide additional attenuation of the higher-order harmonics. The control logic is designed properly to generate the control signals for the phase selectors and the nonlinear PI cores.
This architecture has a couple of notable features. The architecture is open-loop leading to fast settling response to changes in the input code. The short delay is important in close-loop timing recovery applications since loop latency can degrade the loop dynamics and jitter performance. Second, the CMOS circuits can be designed using a standard-cell design flow with other digital logic leading to short design time.
The following subsections discuss each of the key building blocks. Section 5.3.1 introduces the phase selector and inverter combiner design. Section 5.3.2 presents the HR filter and the subsequent low-pass filter. The nonlinear phase interpolator core is presented in section 5.3.3.

5.3.1 Phase Selector and Inverter Combiner

Figure 5.13(a) depicts the circuit structures of the phase selector. The phase selector is an 8:2 MUX that selects the proper phases for the nonlinear PI core. With the appropriate select signals, the 8:2 MUX can be implemented as two 4:1 MUXs and covers 360-degree phase range. As shown in Fig. 5.13(b) and 5.13(c), the 4:1 MUX is designed using four tri-state inverters enabled by the select signals. The series transistors $M_2$ and $M_3$ provide good isolations between the input and output nodes when the inverter is OFF. Note that since the output signal is sensitive to signal couple, linearity degrades if the input is coupled to the output through a gate-drain capacitance.

The same tri-state inverter is used as a combiner in the HR filter and PI core to combine the three or two clock signals. An inverter-based combiner structure adds the two signals with weighting controlled by the sizes of the inverters. Unlike CML-based interpolators, this structure requires no level converters to provide nearly CMOS swings at the output and consumes no static power. Furthermore, due to the aggressive scaling of CMOS processes and the corresponding
improvements in $f_T$, the speed of an inverter is sufficient for most applications. It is well-known that using inverters in a PI structure results in poor phase linearity due to the time-varying nonlinear turn-on resistance. The authors in [87] have proposed to introduce additional capacitance and make the inverter output not fully charged/discharged so that the inverters can operate in a more linear region. However, the reduced signal swing is more sensitive to noise and filtering pole frequency is still sensitive to the PVT variation of the transistors.

In this DPC design, we use a combiner structure with current-limiting resistors in series with the inverter’s outputs as shown in Fig. 5.14(a). The resistors limit the charging/discharging current of the inverters and reduce the variation of the total resistance and desensitize the sensitivity to
variation. The equivalent small-signal model is shown in Fig. 5.14(b). The weighting coefficient $\alpha$ can be derived as

$$
\alpha = \frac{R_{on,B}(t) + R_B}{R_{on,A}(t) + R_{on,A}(t) + R_A + R_B}
$$

(4)

When used in the PI and HR filter, the current-limiting resistors are considered part of the output filter in the DPC equivalent circuit in Fig. 5.1 and Fig. 5.4(a). Hence the design of the resistance value is determined based on the phase linearity and THD requirement.

Fig. 5.14 Proposed inverter-based combiner with improved linearity using series resistors (a) circuit structure, (b) equivalent model

5.3.2 HR Filter and LPF
The HR filter and LPF design are shown in Fig. 5.15(a). The weighting factors of the three different inputs are controlled by sizing the width of inverter and the conductance of the current-limiting resistors with a ratio of 1:1.4:1. The number 1.4 instead of the ideal √2 has been chosen to achieve good matching in layout. The small 1% gain mismatch still provides 50 dB rejections to HD3 and HD5 [86]. To address the nonlinearity of the inverter, we incorporate an additional resistor $R_2$ and capacitor $C_1$, which combined with other passive devices including the parasitic capacitance $C_p$ of the inverters form an equivalent 3rd-order LPF. The parameters of $R_1$, $R_2$, $C_1$, and $C_2$ and the structure are shown in Fig. 5.14(a). The frequency response of the LPF, as shown in Fig. 5.15(b), has a 3-dB cut-off frequency of 1.0 GHz. For a clock frequency of 1 GHz, the LPF provides 26-dB rejection on HD7. The rejection is about 9-dB higher than a single-pole RC LPF.

![Fig. 5.15 Inverter-based HR filter design](image)

(a) Circuit structure, (b) Frequency response of the LPF
The performance of the proposed inverter-based HR filter can be shown by plotting the FFT. Figure 5.16 shows three different cases—without the LPF, with the LPF, and with the LPF plus HR filter. As can be seen, the combination of the HR filter and the LPF has improved the THD substantially at 0.75 GHz. The THD has been improved from 35.1% (without LPF) to 15.0% (with LPF), and further to 1.7% (with LPF and HR filter). The overall simulated INL and DNL are 1.19 and 0.30 degree, respectively. When operating at half and double of the frequency (0.375 and 1.5 GHz), the resulting INL are 6.93 and 0.98 degree and DNL, 1.64 and 0.13 degree.

![Fig. 5.16 FFT simulation of HR filter for three different cases: without LPF (left), with LPF (middle), with LPF and HR filter (right)](image)

5.3.3 Nonlinear Phase Interpolation Core

As mentioned in Section II, the phase linearity can potentially be improved by using nonlinear mapping between the input code and the interpolation weighting factor, $\alpha$. Figure 5.17(a) and (b) show an example of nonlinear weighting and mapping, respectively. If the relationship between
phase and $\alpha$ is predictable, the PI nonlinearity can be calibrated by a fixed mapping to reduce systematic INL/DNL. In an implementation, however, the relationship is sensitive to the waveform shape of the interpolated signal and hence is subject to PVT variation. Furthermore, as the waveform changes with clock frequency, the amount of improvement will be reduced.

Fig. 5.17 (a) $\alpha$ versus code curve for linear and nonlinear phase interpolation, (b) nonlinear coefficient extraction by inverse-mapping

To illustrate the effectiveness of nonlinear phase interpolation under frequency deviations, Fig. 5.18 shows the transfer curves of RC-loaded and HR-filter based PI under different $\alpha$, respectively, where $\alpha=0.5$ and 2 correspond to an input frequencies that are halved and doubled, respectively. For an RC-loaded PI, a significant improvement on linearity can be achieved by a making a nonlinear mapping based on the transfer curve of $\alpha=1$. However, as the clock frequency changes, the improvement decreases accordingly. For HR-filter based PI, on the other hand, if we map the
curve based on the assumption that the waveform is sinusoidal, as frequency increases, the deviation of the transfer curve is not increased as the waveform resembles to a sinusoidal wave at high frequency. This behavior can be observed from the transfer curves of $\alpha=1$ and 2, which almost overlap. In other words, the linearity of HR-filter based PI is improved instead of deteriorated as the frequency increases, which suggests a better immunity to PVT and frequency variations. Note that for an RC-loaded PI, the problem can be improved by incorporating an additional bandwidth control loop. However, it increases cost and complexity and is generally not preferred.

![Graphs showing phase versus $\alpha$ for different cases](image)

Fig. 5.18 Phase versus $\alpha$ curve under the cases of $\alpha=0.5$, 1, and 2 for (a) HR-based PI, (b) RC-loaded PI.
The PI core in this design uses nonlinear code mapping. As shown in Fig. 5.19(a), the nonlinear PI core incorporates with 32 unit cells, each consisting of two tri-state inverters followed by a resistor. The weighting coefficient $\alpha$ is adjusted by selecting the numbers of cells connected to the input of $CK_{in,1}$ and $CK_{in,2}$. By using non-uniform sizing for the resistances, we realize the nonlinear mapping. Since the turn-on resistances of the inverters are small as compared to the resistances of the current-limiting resistors, we find that varying the inverter sizing is not needed. The coefficient is determined as

$$\alpha = \frac{\sum_{i=1}^{K} G_i}{\sum_{i=1}^{N} G_i}.$$  \hspace{1cm} (5)

where $G_i$ is determined by the code to weight relationship through simulation. Figure 5.19(b) shows the coefficient table for both linear and nonlinear phase interpolation. Because of the sinusoidal waveform, the deviation from linear weighting is not large. Simulation shows that, by applying nonlinear PI, the INL and DNL can be further reduced from 1.19 to 1.04 and 0.30 to 0.19 LSB, respectively. The resistance values for generating the nonlinear coefficient are shown in Fig. 5.19(c). The resistors of nonlinear PI core comprise of small segment of unit resistors $R_{\text{unit}}$ to achieve good matching. Simulation shows that the overall coefficient deviation due to quantization error of the resistance is less than 0.1%. Monte Carlo simulation shows that the deviation of DNL...
causing by device mismatch is less than 0.04 LSB for a ±10% deviation on the resistance of \( R_{\text{unit}} \).

![Diagram](image)

Fig. 5.19 (a) Proposed nonlinear phase interpolator code, (b) coefficient table for linear and nonlinear phase interpolation, (c) resistance table for nonlinear phase interpolation

Note that we implemented the nonlinear phase interpolation to explore the amount of improvement beyond the inherent improvement of the HR filter-based PI. While the improvement is noticeable, it comes at the cost of substantial increase in area. Many applications may not need this enhancement.
CHAPTER 6
Experimental Results

This chapter demonstrates the details of the experimental results of two chips. The first chip is a 64-Gb/s transmitter prototype that embeds the LC-ladder-filter-based FFE and the serializer discussed in the chapter 3 and 4, respectively. The second chip is a 1.5-GHz, 8-bit DPC discussed in chapter 5. Both chips are fabricated in 65-nm CMOS technology and tested in chip-on-board assemblies. Section 6.1 and 6.2 presents the experimental results of the 64-Gb/s transmitter prototype and the DPC, respectively.

6.1 Measurement Results of the 64-Gb/s Transmitter Prototype

The 64-Gb/s transmitter prototype described in the previous chapters has been designed and fabricated in 65-nm CMOS technology. Figure 6.1(a) shows the die photo. The transmitter core occupies 1.20 x 1.00 mm². The active area is 0.35 mm², of which about 40% is occupied by the LC-ladder-filter-based FFE, 34% by the balun\(^{10}\) and the first divider, and 26% by the serializer.

\(^{10}\) The area of balun is roughly equivalent to that of an LC oscillator.
and the predriver. The photograph of the test setup is shown in Fig. 6.1(b). The transmitter is wire-bonded to the target channel, which is a 16 mm trace on a Nelco N4000-13 PCB. After the trace, the signal is measured through an edge-mounted 50-GHz connector. The overall channel loss is 9.5 dB at 32 GHz. The transmitter consumes 199-mW power from a 1.2-V supply, which corresponds to an energy efficiency of 3.1 pJ/b. The power breakdown of the transmitter is shown in Fig. 6.1(c). The power consumed by the FFE, serializer, and clocking are 39%, 27%, and 34%, respectively.

![Fig. 6.1 (a) Die photograph, (b) photograph of the testing board, and (c) the power breakdown](image)

The transmitter operates from 50 to 64.5 Gb/s. The range of data rate is limited by the locking range of the divider. The measured 64-Gb/s, single-ended eye diagram for PRBS $2^7$-1 pattern before and after the equalization at data rate of 64 Gb/s are shown in Fig. 6.2(a) and (b), respectively. As can be seen, the FFE opens the eye even without using a receive side DFE with an eye opening of 25mV (differentially 50 mV). The measured differential peak-to-peak output
voltage swing for unequalized eye is 600 mV\textsuperscript{11}. To demonstrate the capability of the transmitter to calibrate quadrature phase mismatch, Fig. 6.2(c) and (d) show the eye diagrams for two worst-case conditions. The first eye minimizes the phase difference between I and Q clocks while the second eye maximizes it. As can be seen, by adjusting the quadrature phases, the eye width of even/odd eyes can be adjusted.

Figure 6.3(a) shows the measured single-ended, 64-Gb/s pulse responses of the FFE. Since the maximum coefficient of the last tap is half of the remaining taps, the coefficients of the remaining three taps are set to be half of the maximum. From the first tap to the last tap, the pulse responses demonstrate delay of 17.8, 17.2, and 16.5 ps for each consecutive tap. The comparison between measured and simulated pulse responses are shown in Table 6.1. The measured pulse responses demonstrate lower amplitude and higher delay, suggesting lower signal bandwidth.

\textsuperscript{11} The 600-mV swing of unequalized eye is limited by the driving strength of one driver buffer. When more than one driver buffer flow current, the achievable swing can go as high as 850 mV and is limited by the DC voltage of transmitter output nodes.
Fig. 6.2 Measured 64-Gb/s, single-ended eye diagrams (a) before equalization, (b) after equalization, (c) with minimized difference between I/Q phases, and (d) with maximized difference between I/Q phases.

Figure 6.3(b) shows the measured peak-to-peak amplitude under different FFE coefficients. The FFE output drivers are implemented with 6-bit current DACs to provide sufficient tuning. The maximum DNL is approximately 1.4 LSB, which suggests a resolution of between 4 and 5 bit. The measured and simulated return loss responses are compared in Fig. 6.4. The peak value within the passband is -14.1 dB, which is acceptable for most standards. The performance comparison with
other published transmitters is illustrated in Table 6.2. This work presents a transmitter including
an FFE that shows excellent power efficiency, a large signal swing, and at a data rate that pushes
the limits of the 65-nm technology. Note that if the power of the FFE is excluded, an energy
efficiency of 2.1 pJ/b can be achieved, which is comparable to the transmitter design of [24].

Fig. 6.3 (a) Measured pulse responses for 4 different taps, (b) measured peak-to-peak amplitude
of pulse response with different tap-weight coefficient

Fig. 6.4 Measured and simulated return loss response
Table 6.1 Comparison of simulated and measured pulse responses

<table>
<thead>
<tr>
<th>Tap</th>
<th>Peak-to-peak Amplitude</th>
<th>Delay Between Taps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulation</td>
<td>Measurement</td>
</tr>
<tr>
<td>Tap1-Tap2</td>
<td>126 mV</td>
<td>148 mV</td>
</tr>
<tr>
<td>Tap2-Tap3</td>
<td>141 mV</td>
<td>130 mV</td>
</tr>
<tr>
<td>Tap3-Tap4</td>
<td>162 mV</td>
<td>115 mV</td>
</tr>
</tbody>
</table>

Table 6.2 TX performance comparison with previous publications

<table>
<thead>
<tr>
<th></th>
<th>This work</th>
<th>[24]</th>
<th>[9]</th>
<th>[16]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>65nm CMOS</td>
<td>65nm CMOS</td>
<td>40nm CMOS</td>
<td>65nm CMOS</td>
</tr>
<tr>
<td>FFE</td>
<td>4-tap</td>
<td>None</td>
<td>2-tap</td>
<td>None</td>
</tr>
<tr>
<td>Data Rate</td>
<td>50 - 64.5Gb/s</td>
<td>32.7-48.4Gb/s</td>
<td>39.8-44.6Gb/s</td>
<td>60Gb/s</td>
</tr>
<tr>
<td>Area</td>
<td>0.35mm²</td>
<td>0.4mm²</td>
<td>2.2mm²</td>
<td>2.1mm²</td>
</tr>
<tr>
<td>Power</td>
<td>199mW</td>
<td>88mW</td>
<td>870mW</td>
<td>450mW</td>
</tr>
<tr>
<td>Data Swing (Differential)</td>
<td>600-850mV_{pp}</td>
<td>300mV_{pp}</td>
<td>1050mV_{pp}</td>
<td>500mV_{pp}</td>
</tr>
</tbody>
</table>

*Active area only ** Depends on FFE coefficients

6.2 Measurement Results of the Digital-to-Phase Converter

The 8-bit DPC described in the chapter 5 has been designed and fabricated in 65-nm CMOS technology. Fig. 6.5(a) shows the die photo. The DPC core occupies 0.26 x 0.23 mm². The area is
dominated by the resistors in the nonlinear PI cores (36%). Figure 6.5(b) shows the power consumption versus frequency. Since the building blocks are inverters, the curve is linear. When operating at 1.5-GHz, the DPC consumes 4.30 mW from a 1.2 V supply. The power breakdown is listed in Fig. 6.5(c).

![Die photo, power consumption versus frequency, power breakdown](image)

Fig. 6.5 (a) Die photo, (b) power consumption versus frequency, (c) power breakdown

An external measurement of the DPC’s phase linearity requires an oscilloscope with < 1.5ps of phase accuracy. Instead, we have implemented an on-chip phase detector as shown in Fig. 6.6(a) [79]. The DPC output is XORed with a reference clock with fixed phase and then filtered with a LPF. The phase information can then be obtained from the DC output voltage. In order to operate the XOR with linear phase detection, the phase of the reference clock is selectable from a set of four different phases using a 4:1 MUX. Figure 6.6(b) shows the output voltage versus phase curves under four different phases. The curves are linear across a wide phase range and the linear regions
overlap for different phases. Thus a linear relationship is guaranteed by properly selecting the reference phase.

Fig. 6.6 (a) On-chip phase measuring circuit, (b) output voltage versus phase curves under different inputs

The DPC operates across a frequency range of 0.1-1.5 GHz. The measured output clock waveform for 1-GHz output and the transfer function for 0.5-GHz output are shown in Fig. 6.7(a) and (b), respectively. The input clocks at 4 GHz have RMS jitter of 3.11 ps and peak-to-peak jitter of 24.43 ps. The output clock at 1 GHz has similar jitter performance of 3.37 ps (RMS) and 24.44 ps (peak-to-peak). As shown in Fig. 6.8, the measured maximum INL and DNL at 500-MHz output 1.33 LSB (1.87 degree) and 0.52 LSB (0.73 degree) respectively. The maximum DNLs occur at code 32/96/160/224, where one input clock of the PI core switches between two different phases. The phase switching of DPC input degrades the phase linearity by inducing unequal signal
couplings between two consecutive codes. The measured INL/DNL curves are periodic for every 90 degree (64 codes) indicating that the source of nonlinearity is not only due to the DPC. The errors are due to unevenly spacing of the 8 clock phases at the input of the DPC. We found a layout mismatch between the group of clocks with 0-, 90-, 180-, 270-degree phase and the other group of clocks with 45-, 135-, 225, 315-degree phase. Despite the mismatch, the measured max-to-min step size ratio is still 1.89, which is better than the common target 2.5.

Fig. 6.7  DPC Measurement results (a) output clock waveform, (b) transfer function at 0.5-GHz output
To isolate the performance of the DPC, we exclude the effect of phase mismatches by normalizing the LSB with respect to the measure phase differences between every 32 codes instead of assuming an ideal LSB of 360 degree/256. The resulting INL and DNL results for cycle1 and cycle2 of Fig. 6.8(a) are plotted in Fig. 6.9(a) and (b), respectively. From the measurements, the maximum INL and DNL for cycle1 are 0.5 LSB and 0.16 LSB respectively and for cycle2 are 0.99 LSB and 0.45 LSB respectively.
To demonstrate the frequency range of the DPC, Fig. 6.10 shows the measured INL/DNL under different frequencies. The DPC demonstrates good INL/DNL across the frequency range of 0.1-1 GHz. As the frequency goes beyond 1 GHz, the INL/DNL increases considerably. The degraded linearity comes from the decreased signal amplitude due to over filtering of the LPF and systematic coupling of clocks. The DPC remains monotonic until 1.5 GHz, which is the upper limit of the operating frequency. As the frequency drops below 0.1 GHz, the linearity degrades quickly due to insufficient filtering of higher-order harmonics.

![Figure 6.10 Measurement results of (a) DPC INL versus clock frequency, (b) DPC DNL versus clock frequency](image)

Table 6.3 shows the performance comparison with other DPC designs. The proposed inverter-based design dissipates considerably less power and occupies lower area. The INL/DNL is
comparable or better than other published results [52-55][62-63][81-84] while operating across a wide frequency range.

<table>
<thead>
<tr>
<th></th>
<th>This Work</th>
<th>[79]</th>
<th>[88]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>65nm CMOS</td>
<td>0.13um CMOS</td>
<td>0.35um CMOS</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>1.2V</td>
<td>1.2V</td>
<td>3.3V</td>
</tr>
<tr>
<td>Resolution</td>
<td>8-bit</td>
<td>13-bit</td>
<td>8-bit</td>
</tr>
<tr>
<td>Operating Frequency</td>
<td>0.1GHz – 1.5GHz</td>
<td>0.5GHz – 1.5GHz</td>
<td>50MHz - 250MHz</td>
</tr>
<tr>
<td>DNL/INL</td>
<td>0.73/±<a href="mailto:1.87degree@0.5GHz">1.87degree@0.5GHz</a></td>
<td>0.04/±4.32degree@1GHz</td>
<td>1.41/±2.81degree@125 MHz</td>
</tr>
<tr>
<td>jitter</td>
<td>3.37ps rms</td>
<td>4.1ps rms</td>
<td>5.1ps rms</td>
</tr>
<tr>
<td>Power (mW)</td>
<td>4.3mW @1.5GHz</td>
<td>15mW</td>
<td>110mW</td>
</tr>
<tr>
<td>Area (mm²)</td>
<td>0.060</td>
<td>0.48</td>
<td>1.156</td>
</tr>
</tbody>
</table>

Table 6.3 DPC performance comparison
CHAPTER 7

Conclusion

This dissertation explores the potential of operating serial-link transmitters at data rates that approach the fundamental technology limit. We showed that the power of transmitters demonstrate a nonlinear relationship with the data rate that rises up drastically as the performance is fundamentally limited by the self-loading capacitance. The maximum achievable data rate is thus limited by the power consumption.

To avoid climbing the power wall, passive devices, especially inductors, are necessary to improve the bandwidth and to reduce the power consumption. We presented in this dissertation several circuit techniques that addresses the power issues of the key elements of transmitter circuits, including FFE, serializer, and DPC, by optimizing the use of passive devices with filter theories. In the FFE design, we proposed a delay-line-based FFE structure that generates the required 1-UI delay with linear circuits with large group delays rather than replicating the entire serializer to reduce the required power consumption per tap. We demonstrated a delay-line design using an LC-ladder low-pass filter that achieves broadband operation with moderate power consumption. We presented an LC combiner network that optimizes the return loss response of
transmitter by distributing the parasitic capacitance at transmitter outputs with LC-ladder low-pass filters. We showed that by using LC-ladder filter design, the LC values of both the delay lines and the LC combiner network and thus the bandwidth and area cost are optimized.

In the serializer design, we employ a quarter-rate, direct 4:1 multiplexing architecture to reduce the number of high-speed latches and MUXs and to relax the timing constraints. To enable the quarter-rate architecture, a novel 4:1 MUX circuit is proposed, which mitigates the effects of the self-loading capacitance of the 4:1 MUX by eliminating the current source of the current drivers. The pulse generator of the 4:1 MUX is optimally arranged and sized to avoid stacked devices and to lower the rise/fall time. The design issues of the 4:1 MUX were discussed and resolved by the proposed clock generator and the predriver.

In the DPC design, we demonstrated that high DPC phase linearity can be achieved by applying phase interpolation with sinusoidal waveforms. To produce sinusoidal waveforms, harmonic rejection filters were introduced, which can improve the phase linearity and extend range of frequency. An inverter-based phase interpolator is presented to further improve the power and area cost. We presented the ideal of nonlinear phase interpolation, which can further eliminate the remaining INL/DNL.
One of the main efforts in this dissertation is to apply filter theories to transmitter circuits including FFE and DPC. One of the main advantages of designing circuits as filters is the accuracy and effectiveness to estimate and to optimize the performance and the cost. With the assistance of filter theories, porting a design into a different data rate can be achieved by applying frequency scaling. When combining with technology parameters such as transistor $f_T$ and overdrive voltage, designers can simply estimate the performance and the fundamental limitation of technologies. Inductors can be applied to designs that have insufficient bandwidth because they create higher-order filters and improves the performance accordingly. These features enable us to achieve data rates as high as half of transistor $f_T$ ($f_T \approx 130\text{GHz}$ in 65-nm CMOS technology). Note that for state-of-the-art Fin FET technologies, the transistor $f_T$ is dropped significantly to enable higher transistor $r_o$ and thus gain. The techniques proposed in this dissertation have a good potential in resolving the insufficient bandwidth problem of transceiver designs using Fin FET.

In this dissertation, we demonstrated that classic filter design improves the use of passive delay-line of transmitter-side equalizer. In fact, this technique can be extended to many applications. One example is to apply the passive delay line to the receiver-side front-end. The receiver front-end shares similar design issues as the transmitter output combiner networks as it also requires ESD devices. Thus, the proposed LC combiner design can be applied to the receiver front-end to
distribute ESD diodes and improve the bandwidth and return loss. Note that since the LC combiner network provides with delayed data without any additional cost, it is possible to realize a receiver-side FFE by taking data from different nodes of the LC combiner network with different delay and summing them up with programmable coefficients similar to [12].

Another extension is to apply the passive delay line to transceivers with sophisticated coding such as PAM4 and PAM8. These applications require transmitters with different driver design. For example, a PAM4 transmitter requires a drivers that is inherently a 2-bit digital-to-analog converter (DAC). The proposed passive delay line is passive in nature, suggesting improved channel responses. As long as the lump capacitance meet the requirement of the LC-ladder filter, the insertion and return loss response can be improved regardless what kinds of drivers are used. The other application is to apply the technique to transmitters with voltage-mode drivers, which typically have higher parasitic output capacitance because both PMOS and NMOS are required.
Appendix

A Low-PDP and Low-Area Repeater Using CTLE for On-Chip Interconnections

The presence of multiple cores, shared on-chip memories, and specialized functional units in large digital integrated systems leads to using repeaters for wide data busses that pass data between blocks. To maintain high performance, high-throughput, low-latency, and low-power on-chip links are needed to transmit across highly dissipative wire channels. The low-pass frequency response of the wires disperses the pulse response and induces inter-symbol interferences (ISI). The most commonly used approach to handle this problem is to insert inverters as repeaters to split the large RC length into smaller segments with sufficient bandwidth [89]. Many circuits that reduce the number of repeaters and power have been proposed in literature [90-92]. Most of these approaches eliminate repeating stages to reduce routing complexity and to achieve low power consumption. Clocked comparators have been proposed at the receiver in order to recover the low swing, equalized data. These approaches have not been widely adopted since the latency is often limited by the clocking constraint. An added advantage of inverter repeaters is their compatibility with standard-cell design methodology. Also, the single-ended signaling enables higher throughput.
This chapter presents a repeater circuit with built-in equalization that preserves advantages of an inverter repeater while achieving 44% lower PDP and 46% lower area. In Section A.1, the trend and challenges of on-chip interconnections under the technology scaling are reviewed and discussed. Section A.2 discusses the model of on-chip wires and presents the channel prototype that we use to verify our design. The proposed repeater that overcomes the ISI problem with a passive CTLE is presented and discussed in Section A.3. In Section A.4, we describes the detail of our transceiver prototype that transmits 16 x 4-Gb/s data over 16 parallel channels, and the experimental results are presented in Section A.5.

A.1 On-Chip Interconnections: Trends and Challenges

The speed of digital circuit is constrained by both gate and wire delays. Despite the fact that gate delays have been decreasing in deep-submicron processes due to the reduced transistor size, the decreased dimension has been impacting the performance of wires drastically. Figure A.1(a) compares the FO4 delays and wire delays across different technologies. As can be seen, on the contrary to the decreased FO4 delays, wire delays have been increased considerably for advanced technologies. We can thus predict that, in the upcoming future, wire instead of gate delays will become the main performance bottleneck that fundamentally limits the performance in advanced technologies.
Fig. A.1 (a) FO4 and wire delay under different technology, (b) cross-session of different technologies

This trend can be understood from comparing the cross sessions of two different technologies, as shown in Fig. A.1(b). Since scaling the width and height of wires simultaneously reduces the cross-session area and increase the resistance essentially, the height is scaled slower, resulting wires that are taller than wider. The sidewall ($C_c$) instead of bottom-plate ($C_{bot}$) capacitances dominates, resulting in an increased wire capacitances for new technologies because the space is scaled faster than the thickness. Figure A.2(a) and (b) show the intrinsic wire delay as a function of the absolute length and the length normalized to the gate length across different technologies. As can be seen, even including the fact that technology scaling reduces the required distances of

---

12 The ratio of sidewall capacitance to bottom-plate capacitance is roughly 3:1 for modern technologies.

13 The intrinsic wire delay is defined as the delay of wires driven by ideal voltage sources.
routing, the wire performance is still degrading with technologies [89].

![Graph](image)

Fig. A.2 Intrinsic wire delay under (a) different length, (b) with length normalized to gate length

To understand the impact of wires, Ho et al. in [89] propose to categorize wires into three different groups—local, semi-global, and global, and evaluate them respectively. Local wires connect gates within modules, and the delay of which is less a problem because the distances are limited to a dozens of gates and they scaled closely with technologies. Semi-global wires connect modules with routing distances ranging from 0.1 up to 2 mm. As more and more modules are integrated on chip, the required routing complexity is ever-growing. Semi-global wires therefore suffer drastically from the poor wire characteristics. Finally, global wires connect top-level modules using top-metal layers. The required outing distances can be as far as a few mm, which induce large latencies that pose severe difficulties for both system and circuit designers [93-94].

To summarize, both semi-global and global wires will be impacted by the technology scaling.
Although various ways have been proposed to improve the performance of global interconnections [94-104], none of them are practical for semi-global interconnections since these approaches require the use of differential signaling and high-gain clocked comparators. These requirements double the cross-sectional area and constrain the latency to exactly one bit period. The design must be completed in full-custom design flow, which leads to large design efforts.

### A.2 On-Chip Wires: Problems and Conventional Solutions

An accurate model of on-chip wires is essential to evaluate the performance of on-chip links. The characteristics of on-chip wires are different from off-chip interfaces in that they are thinner and narrower and are thus highly-resistive. One of the most accurate ways is to model on-chip wires as RC distribution networks with an infinite number of RC segments. Ignoring wire inductances are acceptable for most applications. Applying a large number of RC segments, however, increase the time required for simulation. To reduce the effort of running simulations, Antinone et al. [105] compare the distributed RC models using different number of segments and conclude that five segments is satisfactory for most applications. Based on this result, a small-signal model as shown in Fig. A.3 is used in our design. Here, $R_{\text{Drv}}$ and $C_{\text{Drv}}$ is the on-resistance and parasitic capacitance of the inverter, and $R_{\text{wire}}$ and $C_{\text{wire}}$ are the lumped resistance and capacitance of the wire. The transfer function consists of 6 poles with the dominant pole $p_1$ much
lower than the others as shown in the pole-zero plot in Fig. A.4. The delay of the wires is dominant by \( p_1 \), which can be derived as:

\[
p_1 = \frac{1}{0.5 R_{\text{wire}} C_{\text{wire}} + R_{\text{drv}} (C_{\text{wire}} + C_{\text{drv}})}.
\]  

(A.1)

As can be seen, the on-resistance and the drain capacitance of the driver inverter corresponding with wire resistance and capacitance lower the frequency of \( p_1 \), which leads to high latency and limits the data rate. Although sizing the inverter larger to reduce \( R_{\text{drv}} \) mitigates the problem somehow, the frequency of \( p_1 \) is fundamentally limited by the intrinsic wire delay, \( 0.5 R_{\text{wire}} C_{\text{wire}} \), which depends on the length, the thickness, and the width of wires. As suggested by Fig. A.2(a), since both wire resistance and capacitance increase with length, the intrinsic delay is quadratic related to the length, which makes latency a big concern for long-distance on-chip links. In addition, the low-pass response due to \( p_1 \) induce ISI and limits the data rate and thus the throughput. State-of-the-art on-chip links with more than 3-mm distances can hardly exceed cross-sectional throughput of 2 Gb/s/um [89-92] [94-103].
To illustrate the impact of the low-pass channel response, a channel prototype, targeting for emulating semi-global wires of digital systems, is designed. Figure A.5 depicts the cross-sectional view of the prototype. The objective is to transmit 4-Gb/s data through single-ended, ground-shielded M4 wires across 5-mm length with a total width of 1um per channel. The frequency response of the channel prototype with length of 1 and 5 mm are plotted in Fig. A.6. As can be seen, for the wire with 5-mm length, the channel induces -17.2-dB loss at 2 GHz, which can induce a latency of 400 ps and limit the data rate to less than 1 Gb/s.
The quadratic delay-length relationship makes long-distance on-chip wires problematic for designers. To overcome this problem, two of the most common approaches are shown in Fig. A.7. The first approach is to insert several inverters as repeaters to break the long wire into shorter pieces as shown in Fig. A.7(a). The optimal number of repeaters for a given wire channel can be derived as shown in [89], and by inserting repeaters, the delay-length relationship can be improved from quadratic into linear. In our target channel, breaking the 5-mm wire into five 1-mm pieces minimizes the delay. As can be seen from Fig. A.6, the channel loss for a 1-mm wire is only -1.6 dB, which is significantly better than the 5 mm one. Inverter repeaters, however, cause additional capacitance loading and thus induce additional power consumption. In addition, repeaters occupy additional area and increase layout complexity.

The other approach [90-104], as shown in Fig. A.7(b), is to transmit data without interruption
and to overcome the ISI by applying equalization. This approach does not require additional repeating stages and thus demonstrates the advantages of low power and layout complexity. Equalization, which is usually accomplished with transmitter-side pre-emphasis, reduces the signal swing unavoidably. Differential signaling and a high-gain, clocked comparator are required. These requirements double the layout area and limit the latency to one clock cycle due to clock constraints. In addition, the design must be completed in full-custom design flow, which leads to large design efforts and can be impractical.

![Image](image1.png)

(a) Repeater Approach

(b) Equalized Link

Fig. A.7 Conventional ways of solving the delay problem: (a) Repeater approach, and (b) equalized link

A.3 Proposed On-chip Link Using Passive Equalizer

The proposed on-chip link is shown in Fig. A.8. As discussed in the previous section, for the target 5-mm distance, when using inverter repeaters, 5 stages are required for minimized delay. Without repeaters, the channel introduces -17.2dB loss at 2GHz. The proposed design utilizes a passive continuous-time linear equalizer (CTLE) as the receiver front-end. To recover the
equalized signal back to full swing, an additional inverter, INV1, is inserted in between the CTLE and the wire driver INV2. This inverter adds a modest delay but maintains proper logic polarity. By not requiring a clock, this simple approach directly replaces inverter repeaters requiring fewer repeating stages leading to lower power dissipation. For the 5-mm channel, we apply two repeating stages, each driving a 2.5-mm segment that has 6.9dB loss at 2GHz. The channel response is shown in Fig. A.9, which demonstrates a bandwidth of 1.0GHz. The equalizer is designed to compensate the channel loss and extend the bandwidth from 1.0 to 2.5GHz. The added bandwidth further reduces latency.

Fig. A.8 Proposed on-chip link architecture
Fig. A.9 Frequency response before and after equalization

Fig. A.10 shows the design of the passive CTLE, which consists of a resistive divider (formed by $R_B$ and $R_C$) that attenuates low-frequency signals and a coupling capacitor $C_C$ that passes high-frequency signals. The coupling capacitor $C_C$ is implemented by stacking a metal-oxide-metal capacitor ($C_{MOM}$) on top of an NMOS capacitor ($C_{NMOS}$) to minimize area. The dc level of the CTLE output is biased by $V_{bias}$, which is generated by a shared bias generator consisting of a decoupling capacitor ($C_D$) and feedback inverters that consumes little static current (30uW/link).

For non-dc balanced data, an additional inverter followed by resistors $R_B$ and $R_C$ is required for biasing purpose. The frequency responses of the unequalized channel ($V_{in}$), the passive CTLE ($V_X/V_{in}$), and the overall channel ($V_X$) are plotted in Fig. A.10. The passive CTLE inserts a zero, $z_1$, to cancel out the channel-induced dominate pole, $p_1$. The bandwidth is extended to the frequency of the CTLE-induced second pole, $p_2$, which depends on the product of $z_1$ and the inverse...
of the dc gain. Ideally, the bandwidth can be further improved by choosing a low dc gain. One of the key advantages of this design is to leverage \( INV_1 \) as a single-ended receiving amplifier. By reducing the voltage swing by only 6.6dB (~0.45X), the structure dissipates little static current.

Figure A.11(a) illustrates the power consumption of \( INV_1 \) under different pre-emphasis showing the increase in power with lower voltage swing (lower \( A_{DC} \)). This trade-off limits the bandwidth improvement. We choose \( A_{DC} \) to be roughly 6dB to cover sufficient PVT variation. Figure A.11(b) compares the power consumption of inverter repeater and proposed repeater under different activity factors (\( \alpha \)). The proposed repeater consumes less power when \( \alpha \) is higher than 0.054.

![Fig. A.10 Proposed passive-CTLE-based repeater design](image)

\[
A_{DC} = \frac{R_B}{R_C + R_B}
\]

\[
A_{AC} = \frac{C_C}{C_C + C_L}
\]

\[
A_{tot} = A_{DC} \cdot A_{AC}
\]

\[
z_1 = \frac{1}{R_C \cdot C_C}
\]

\[
p_2 = \frac{z_1}{A_{DC}}
\]
Fig. A.11 (a) INV1 static power under different pre-emphasis, and (b) power consumption under different activity factor

This passive CTLE has several advantages over other types of equalizers. First, since the channel response can be accurately extracted and modeled, a single zero accurately inserted can equalize the channel as effectively as an FFE or DFE. Second, the CTLE in combination with the receiver amplifier consumes little power (47uW) for equalization. Third, the design is robust to PVT variation since an incomplete pole-zero cancellation only results in a slightly higher ISI or peaking. Simulation shows that varying $z_1$ frequency by $\pm 30\%$ impacts the eye opening by less than 7%. Finally, Fig. A.12 compares the layout of the proposed repeater and the inverter repeater, which shows that while the CTLE repeater occupies 15% larger area per stage compared to inverter repeaters, it only requires 2 stages which leads to 46% lower area.
A.4 Experimental Results

To verify the proposed circuits, a prototype that transmits 16 x 4Gb/s data over 16 parallel 5mm wires is implemented in 65nm CMOS technology. The transceiver structure is shown in Fig. A.13. The required input sequence for testing is generated by a 16b, $2^{27}-1$ PRBS generator. To verify the correctness of the transceived data, a bank of quarter-rate samplers, operating at 1GHz, is incorporated which retimes and sends the data to the 64b, 1GHz parallel PRBS checker for BER measurement. Additional channels are included to measure the delay, power, and signal waveform. A channel using inverter repeaters is included for comparison.
Fig. A.13 The architecture of prototype IC with 16 parallel channels

Fig. A.14 illustrates the eye diagrams of the unequalized ($V_{\text{unEQ}}$) and equalized ($V_{\text{EQ}}$) signals after a 2.5mm channel. Fig. A.15(a) and (b) shows the voltage waveforms of the input ($V_Y$) and output ($V_Z$) of a channel and the BER bathtub curve, respectively. The higher peak-to-peak jitter is due to residual ISI and can be completely removed by either reducing the segment length by 15% or a 28% lower data rate. Fig. A.16(a) and (b) shows the measured PDP of the proposed link under different voltage swings ($V_{\text{EQ}}$) and the comparison between the two approaches under different supply voltages, respectively. The measured power and delay are 968uW and 286ps respectively. When compared to using inverter repeaters, the power, delay, and PDP are 29%, 22%, and 44% lower. Figure A.17(a) shows the die micrograph. Figure A.17(b) compares the energy efficiency per length and throughput with other state-of-the-art papers. A table summarizing the performance of this work and that of other state-of-the-art papers is shown in Table A.1.
Fig. A.14 Measured eye diagrams of unequalized ($V_{\text{unEQ}}$) and equalized ($V_{\text{EQ}}$) outputs after a 2.5mm channel

Fig. A.15 (a) Measured eye diagrams of input ($V_Y$) and output ($V_Z$) and, (b) measured BERT bathtub curve
Fig. A.16 (a) Simulated and measured PDP under different voltage swings and, (b) measured PDP under different supply voltages

Fig. A.17 (a) Die micrograph, and (b) comparison of energy efficiency per length and throughput of state-of-the-art on-chip links
### Table A.1 Performance Summary

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>90nm</td>
<td>65nm</td>
<td>65nm</td>
<td>65nm</td>
<td>65nm</td>
</tr>
<tr>
<td>Link Data Rate</td>
<td>4Gb/s</td>
<td>90Gb/s</td>
<td>3Gb/s</td>
<td>64Gb/s</td>
<td>64Gb/s</td>
</tr>
<tr>
<td>Channel Data Rate</td>
<td>4Gb/s</td>
<td>10Gb/s</td>
<td>3Gb/s</td>
<td>4Gb/s</td>
<td>4Gb/s</td>
</tr>
<tr>
<td>Length</td>
<td>10mm</td>
<td>6mm</td>
<td>10mm</td>
<td>5mm</td>
<td>5mm</td>
</tr>
<tr>
<td>Throughput (Gb/s/um)</td>
<td>2</td>
<td>2.56</td>
<td>0.75</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Energy Efficiency</td>
<td>356fJ/b</td>
<td>1044fJ/b</td>
<td>95fJ/b</td>
<td>169fJ/b</td>
<td>121fJ/b</td>
</tr>
<tr>
<td>Delay</td>
<td>250ps</td>
<td>100ps</td>
<td>333ps</td>
<td>365.4ps</td>
<td>285.8ps</td>
</tr>
<tr>
<td>FOM (pJ/b*ps/m)</td>
<td>1.78</td>
<td>11.31</td>
<td>1.42</td>
<td>2.47</td>
<td>1.38</td>
</tr>
</tbody>
</table>

### A.5 Conclusion

On-chip, highly resistive, semi-global and global wires pose severe difficulties for interconnection design. This work presents an improved repeater circuit that achieves a lower power, delay, and area design by applying passive CTLE. The design preserves the advantages of the inverter repeater including being single-ended and having high-throughput, and can potentially be integrated into standard-cell library. Designed and measured in 65nm CMOS technology, the proposed repeater achieves 44% lower power-delay product (PDP) while occupies 46% lower area.
References


Front-End and 14-Tap Decision Feedback Equalizer in 28 nm CMOS," IEEE J. Solid-State

[18] C. Menolfi et al., "A 28Gb/s Source-Series Terminated TX in 32nm CMOS SOI," in


[20] H. Wang et al., “A 40-Gb/s Transmitter with 4:1 MUX and Subharmonically Injection-
Locked CMU in 90-nm CMOS Technology,” in Symp. VLSI Circuits Dig. Tech. Papers
(VLSI’09), June 2009.

[21] H. Wang et al., “A 21-Gb/s 87-mW Transceiver with FFE/DFE/Linear Equalizer in 65-

[22] P. Chiang et al., " A 20-Gb/s 0.13-um CMOS Serial Link Transmitter Using an LC-Pll to
Directly Drive the Output Multiplexer," IEEE J. Solid-State Circuits, vol. 40, no. 4, pp. 1004-1011,


[25] M.-S. Chen et al., "A 50-64 Gb/s Serializing Transmitter with a 4-tap, LC-ladder-filter-
based FFE in 65-nm CMOS," in Proc. IEEE Custom Integrated Circuits Conf. (CICC’14), pp. 1-4,
Sept. 2014.


[34] PCI Express Based Specification Revision 3.1 [Online]. Available:


2006.


[84] A. Agrawal et al, “A 19Gb/s Serial Link Receiver with Both 4-Tap FFE and 5-Tap DFE


[93] T. Krishna et al., "Breaking the on-chip latency barrier using SMART," in Proc. of High


