Novel approach for serial data link in 28nm CMOS FDSOI and beyond

Mohammed Tmimi

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

Spécialité : **Nano Electronique et Nano Technologies**

Arrêté ministériel : 25 mai 2016

Présentée par

Mohammed TMIMI

Thèse dirigée par **Philippe GALY**, et
codirigée par **Philippe FERRARI, Stefano D'AMICO, Jean-Marc DUCHAMP**

préparée au sein du **Laboratoire RFIC**
dans **l'École Doctorale Electronique, Electrotechnique, Automatique et Traitement du Signal (EEATS)**

Nouvelle approche pour lien série en technologie FD-SOI 28 nm CMOS avancée et au-delà

Thèse soutenue publiquement le **16 décembre 2020**,
devant le jury composé de :

Monsieur Jean-Baptiste BEGUERET
Professeur, Université de Bordeaux, Rapporteur

Monsieur Andrea MAZZANTI
Professeur, Univérsité de Pavia, Rapporteur

Madame Anne-Laure BILLABERT
Maitre de conférences HDR, Le Cnam/ESYCOM Paris, Examinatrice

Monsieur Sylvain BOURDEL
Professeur, Univérsité de Grenoble Alpes, Président

Monsieur Philippe GALY
Directeur technique, HDR, STMicroelectronics, Directeur de thèse

Monsieur Philippe FERRARI
Professeur, Univérsité de Grenoble Alpes, Co-directeur de thèse, invité

Monsieur Stefano D'AMICO
Maitre de conférences, Univérsité de salento, Co-encadrant de thèse, invité

Monsieur Jean-Marc DUCHAMP
Maitre de conférences, Univérsité de Grenoble Alpes, Co-encadrant de thèse, invité

Monsieur Philippe CATHELIN
Docteur Ingénieur, STMicroelectronics, invité

"Real knowledge is to know the extent of one's ignorance."

- Confucius

Abstract

The global internet traffic exceeded the zettabyte marker in 2016, initiating the zettabyte traffic era. Since then, internet traffic proliferated with a compound annual growth rate of 26%; and is expected to continue its astronomical growth rate. This perpetual growth has significant implications for networking technologies and their present limits. Researchers anticipated these limits and managed to stay ahead of the curve by innovating and optimizing all data transfer levels. In that context, this work focuses on on-chip data transfer, acknowledging that communication energy efficiency is one of the integrated circuits near future bottlenecks, as the gap between the computation energy and on-die IC energy grows.

Evidently, improvements have to be made to the existing links solutions; higher data rates must be reached while considering the energy efficiency and the circuit complexity. Furthermore, with the increasing data rates, signal integrity problems arise due to channel imperfections. Although transistor scaling provided higher density packing of devices along with faster transistors, it did not benefit the interconnections performance since it resulted in higher wires density. Wires are more sensitive to their environment than active devices, that is, closer wires are more sensible to crosstalk and longer delay due to the wire's intrinsic delay. Delay is a critical metric for data transmission, and lower controlled delay is favored. In this work, we developed a high data-rate low delay solution for long-range on-chip serial links. The developed solution is complementary to the massively employed existing solutions. We believe it will help solve some of their issues and extend the existing Network on chips (NoC) architectures lifetime.

We start this work by introducing the standard and emerging on-chip interconnect solutions, then discussing their advantages and challenges. Clearly, each solution is suitable for specific applications. The chosen RF interconnects technique is most suitable for our requirements, mainly due to low delay, high available bandwidths, and CMOS process compatibility/friendliness. This approach requires transmitting the data at high frequencies instead of the baseband, that is, up-converting the data signal before transmitting it through the transmission lines. In practice, transmission lines behave differently at baseband and high-frequencies. In particular, both distortion and delay are much lower at high-frequencies. These two properties are essential for our work; low distortion implies that high signal integrity is reached without equalization or error-correcting codes, up to 14 Gbps in the proposed study. At least four times lower than baseband delay, the high-frequency low delay property signifies that long distances across the chip can be crossed in less time.

We believe this approach is most beneficial for distances longer than a couple of millimeters and up to twentieth millimeters.

Bandwidth at higher frequencies (60 GHz in our case) is a valuable commodity. To take full advantage of the available bandwidth, we choose to use duobinary modulation to double the data rate within the same bandwidth. This spectrum compression relaxes the RF components constraints such as linearity; The chosen modulation also allows for simple demodulation where a simple envelope detector is used to recover the 14 Gbps data.

A 10 Gbps prototype chip was designed and fabricated in the advanced 28 nm FD-SOI technology from STMicroelectronics. This technology is introduced in the third chapter; afterward, we explained the design process of the 10 Gbps transceiver (composed of a transmitter, a receiver, and a 4.6mm channel). The simulation results showed that we reached a higher data rate (at least double) than the state of the art, for a smaller area and a comparable energy efficiency. The post-layout simulation resulted in a BER lower than 10^{-12}. The measurement results will be published in future works.

We also proposed to use the same approach for interposer channels to connect chiplets with minimal delay. We study its application for a 130 nm BiCMOS technology passive silicon interposer, but it can be used for active ones too. We connected two 28 nm FD-SOI chiplets at a 7-mm distance and achieved a BER lower than 10^{-12} with a 7 ps/mm delay in simulations.

Résumé

Dans le cadre de l'échange massif de données numériques, la solution du lien série est largement utilisée dans les systèmes électroniques. Dans ce cadre, il existe une course permanente pour accroître le débit de transfert de données. Notamment les efforts portent sur l'amélioration de l'efficacité énergétique du système et l'optimisation des canaux de transmission. Cependant la contrainte physique du canal de transmission est une donnée majeure dans cette approche de transmission de données à haut débit.

Les méthodes standard de transmission intra-puces point à point utilisent la bande de base, le délai de transmission dans cette bande se situe autour de 40 ps/mm, acceptable pour des distances courtes inférieures au mm. Or, pour un lien de quelques mm, la solution standard d'utiliser des routeurs n'est plus optimale quant à la consommation et au temps de transfert dus à la propagation du signal en bande de base. En conséquence, un changement de paradigme est nécessaire afin de réduire ce délai.

Aujourd'hui, les recherches sont très actives concernant l'intégration monolithique de lien série, ce qui permet d'avoir une excellente base de concepts et de solutions. Dans la littérature, on note ainsi plusieurs solutions, la principale étant la transmission sans-fil intra-puces « wireless on-chip (WiNOC) », où des antennes intra-puces sont utilisées pour transmettre les données. On peut également noter l'utilisation de l'optoélectronique pour transmettre avec un délai minimal. Il en résulte un changement de processus.

Dans ce travail, on vise les liens de quelques mm de long, où aucune des solutions précédentes n'est optimale, soit à cause du temps de propagation soit à cause de la complexité de l'implémentation due au changement du procédé. Cette solution est complémentaire aux solutions existantes et nous pensons qu'elle permet de résoudre certains de leurs problèmes et prolonger la durée de vie des architectures réseau sur puces (NoC) existantes.

On investigue la transmission en bande millimétrique (à 60 GHz) où la vitesse de propagation du signal est autour de $1{,}5.10^{8}$ m/s, impliquant un délai minimal (7 ps/mm). Par ailleurs, différentes modulations seront investiguées pour augmenter le débit et exploiter efficacement les bandes passantes disponibles à ces fréquences. On a choisi la modulation duobinaire pour son avantage en termes de compression du spectre, ce qui nous a permis de doubler le débit utilisé pour une même bande passante, ainsi que pour sa simplicité de modulation/démodulation. Dans notre cas, on utilise 5 GHz de bande pour transmettre un signal de 10 Gbps.

Cette approche théorique a été modélisée pour ensuite la comparer aux différents systèmes à l'état de l'art ; un débit maximal de 14 Gbps a été atteint avec un taux d'erreur inférieur a 10^{-12} en simulation. Un démonstrateur sur silicium à 10 Gbps a été conçu sur la base de la technologie CMOS avancée 28 nm FD-SOI de STMicroelectronics. Le transmetteur, le récepteur ainsi que des lignes de propagation d'une longueur de 4.6 mm ont été implémentés, les résultats de mesures seront publiées dans de futurs travaux. Les simulations ont montré que nous avons atteint un

débit plus élevé (au moins le double) que l'état de l'art, pour une surface plus faible et une efficacité énergétique comparable.

Nous avons également proposé d'utiliser la même approche pour les canaux d'interposeurs afin de connecter des chiplets avec un délai minimal. Nous étudions son application pour un interposeur passif en silicium en technologie BiCMOS 130 nm, mais il peut également être utilisé pour les circuits actifs. Nous avons connecté deux puces en technologie 28 nm FD-SOI à une distance de 7 mm et obtenu un taux d'erreur binaire inférieur à 10^{-12} avec une latence de 7 ps / mm en simulation.

Acknowledgment

It is a pleasure to thank the many people that made this thesis possible.

Firstly, I would like to express my sincere gratitude to my advisors Philippe Galy, Philippe Ferrari, Jean-Marc Duchamp, and Stefano D'Amico for the continuous support of my Ph.D. research and their patience. Their guidance helped me in all the time of research and writing of this thesis. I am incredibly grateful for their confidence and the freedom they gave me to do this work.

I would like to thank my thesis committee members for their time, their invaluable feedback and intellectual contributions.

My special appreciation and thanks go to Philippe Cathelin, a great mentor, for his guidance and support.

I would also like to acknowledge with gratitude Andreia Cathelin, Stéphane Le Tual, Sylvain Clerc, Denis Pache, and Roberto Guizzetti for their help and insightful comments. I am also extending my thanks to the bat.6000 colleagues and friends: Olivier Jeantet, Raphael Gras, Tarun Chawla, Thomas Ahrens, Bruno Spy, Sebastien Peurichard, and Daniel Pierredon.

I am grateful to my friends and Ph.D. students from STMicroelectronics Crolles with whom I shared enjoyable moments. Thanks to Valérian Cinçon, Ioanna Kriekouki, Geoffrey Delahaye, David Gaidioz, Zoltan Nemes, Clément Beauquier, Antoine Le Ravallec, Guillaume Tochou, Alexis Rodrigo Iga Jadue, Romane Dumont, Robin Benarrouch, Angel De Dios Gonzales Santos, Thibault Despoisse, Sebastien Sadlo, Dayana Andrea Pino Monroy and Soufiane Mourrane.

Special thanks to Thomas Bédécarrats, Renan Lethiecq, Louise De Conti, Raphael Guillaume, Florian Voineau, Abdelhakim Mahjoub, Yassine Oussaïti and Ioana Andrada Burducea for the support and kindness, words cannot express how grateful I am.

Most importantly, none of this could have happened without my family. To my parents, I am extremely grateful for your encouragement, caring, and sacrifices. To my sisters, Narimane and Meryame, thank you for your support and encouragement. I will be grateful forever for your love.

Table of Contents

List of Figures

List of tables

Introduction

Context

The Internet, computers, telephones, and airplanes are all considered part of the man's greatest inventions. They made our life more comfortable and simpler; they made the world a small village. So what do they have in common? They were all enabled by electronics. Electronics is a stepping-stone towards a better future, a more civilized and cleaner future.

For the last century or so, engineers have innovated by putting more and more applications into smaller spaces. At the beginning of the mobile revolution, a mobile weighed 4.5kg and required a carrying handle; Nowadays, our smartphones weigh less than 200 grammes and can fit in our pockets; they also have a camera (or several), a GPS, and some powerful CPUs...

These fast developments are mainly due to the advancements in microelectronics, integrated circuits architectures, and design. In the past, circuits were integrated on a large board (PCB, for example); the pitches were large and required a sizeable surface. Nowadays, more and more components are integrated into the same integrated circuit and are placed as close as possible. This resulted in a significant increase in performance. It also opened the door to new architectures such as the system-on-chip (SoC) or chip-multiprocessors improvements, where several functionalities are placed into the same integrated circuit (a CPU and its memory, for example); and are connected using an efficient on-chip communication network. The on-chip communication substantially impacts the chip performances, including its maximum speed, cost, reliability, and power consumption.

Intel predicted that communication energy efficiency would become the main challenge in more advanced nodes ICs. The interconnect energy is scaling slower than the computation energy within integrated circuits. Hence, the data movement on-die energy will dominate, as shown in Figure 1.

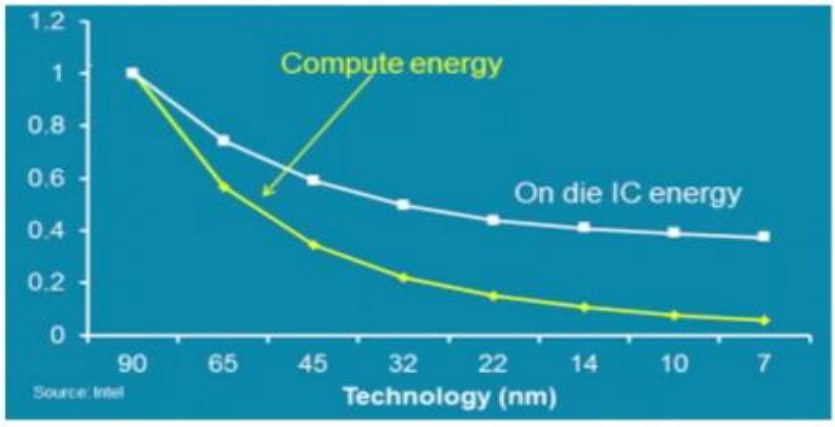


Figure 1 : Compute energy vs total on die IC energy for technology nodes.

Researchers have been working on several approaches to optimize the different metrics, either at the system level by proposing new architectures or at the devices

level by changing the used materials for wires, for example. Usually, trade-offs are required to meet different constraints.

One of the crucial steps toward optimized SoCs was developing Network-on-chips (NoC) solution; Before NoCs, designers relied on shared global buses to transmit the data; this approach was sufficient for up to ten cores single chip. However, as the chips kept growing and implementing extra functionalities, the shared global bus approach showed its limits and began to crumble. Mainly due to the wire's limitations on latency and overall speed of the chips.

NoCs, by design, allows for a larger number of cores and multiple communications at one time; they rely on routers to direct the signal from one point to the other, routers are connected through wires, similar to a traffic grid. In other words, the NoC connects different isolated IP blocks and defines the communication fabric/structure within the die; it also impacts its limits significantly. A poorly designed NoC will result in routing congestion and timing issues.

Fortunately, new NoC architectures or new topologies are popping up to handle different constraints and develop the quality-of-service with the increasing performance demands. In this work, we will propose a solution that may reduce the routing congestion within ICs, improve their performances, and help bring new architectures to light. One might imagine that the link developed in this work may be used for large ICs critical (timing) path or to place the memory further away from the CPU.

Motivation

Our work will not replace the standard wires solutions massively used nowadays. Instead, we propose it as a complementary solution to help solve some of the issues (especially timing issues) and extend the lifetime of the existing conventional NoC architectures for some time. We believe that the work proposed herein is most suitable for on-chip interconnects longer than two millimeters; because for distances lower than 2mm, the standard solutions are cost-efficient and more straightforward to implement.

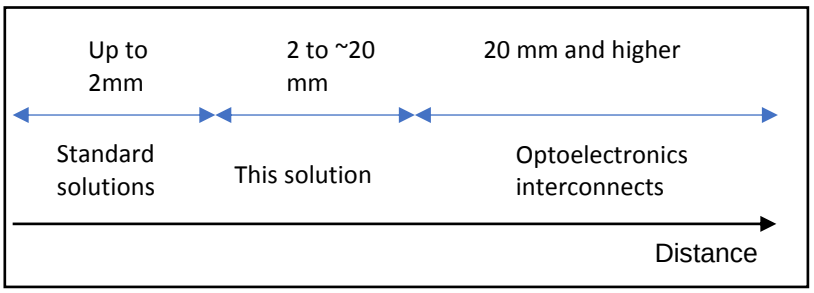


Figure 2 : The proposed solution versus common interconnects distance range.

Nowadays, large ICs require more and more long-range across the chip connections (up to several millimeters); for example, the 2cm × 2cm processor shown in Figure 3, requires interconnects as long as 15mm to connect the chips borders. If only NoCs are used, the transmission takes several clock cycles to arrive; furthermore, going through all the routers, buffers... uses more resources than it should. For such long distances, optical interconnects seem like a reasonable option; however, they usually require more complex processes than simple ones, i.e., increases the cost.

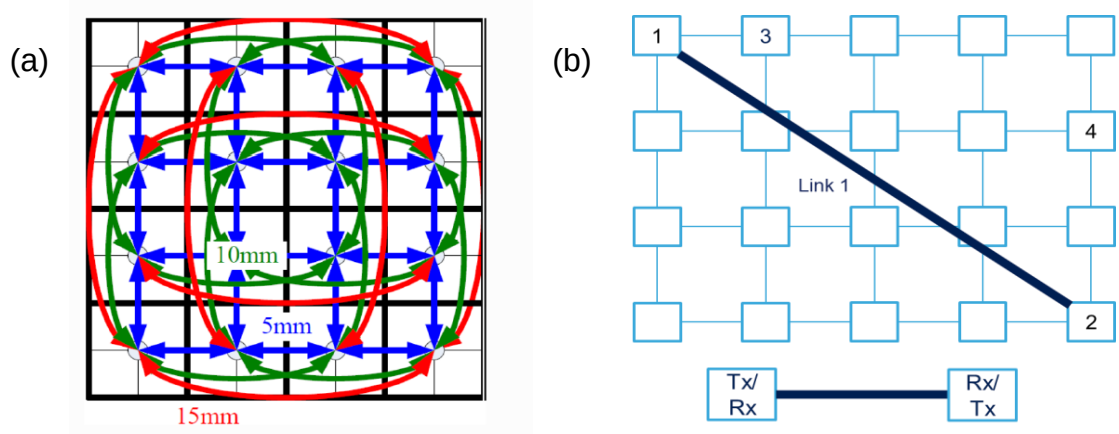


Figure 3: Possible usage applications of the proposed link. (a) Replace the long link in a NoC Torus topology with the proposed links.(b) Connect the furthest points directly.

The proposed approach's main benefit is to reduce the propagation latency (delay) of the wires. The propagation delay can be around 30 -40 ps/mm and higher in the standard solution, but it grows quadratically with the standard wires' length. In the used approach herein, the delay is linear and is around 7-8ps/mm. It is clear that the interest of this solution rises with distance. This link can be used to connect the edges of the chips creating high data rates fast links, as shown in red or green in Figure 3.a. This topology is similar to the standard Torus NoC topology; We propose to replace the long slow links with this fast links to improve the performances. If designed correctly, they can also connect the two furthest points on a die, as shown in Figure 3.b. Thus, the data sent through these links will not go through the NoC routers and switches; it will reduce their congestion and provide a back-up link to retransmit lost data, for example.

One of the main issues in communication systems is distortion due to the channel's physical properties such as the channel's dispersive nature, where the frequency components propagate at different velocities, or an increase in the channel's attenuation coefficient due to the skin effect, for example. This usually causes the pulse traveling through the channel to decrease in amplitude and spread in the time domain. Several methods have been carried out in order to solve these issues. Equalization was used to reduce the channel's frequency selectivity by reversing the distortion; Then reconstructing the initially transmitted signal.

However, the complexity and power consumption of the equalizers dramatically increased in the last years. The equalization approach was attractive with the transistor and voltage scaling, which is not optimal anymore since voltage scaling is slowing down.

In this work, we do not require equalization, since as it will be shown, the transmission lines do not suffer from such effects at higher frequencies; Thus, they provide a good medium for transmission with minimal distortion, and we will show that we achieved a BER lower than 10^{-12} up to 14Gbps data transmission.

Furthermore, we used bandpass duobinary modulation to take full advantage of the available bandwidth; Duobinary modulation is more efficient as it compresses the spectrum of the signal by a factor of two, and thus use a two time smaller bandwidth. This compression can also be translated into more relaxed design constraints on the RF components (linearity...) compared to an ASK modulation. In addition, non-

coherent demodulation is used relying on a simple power detector without any need for a carrier in the receiver. We designed a 10Gbps prototype chip, and we used the advanced 28nm FD-SOI technology from STMicroelectronics to design the transmitter, the channel, and the receiver. We will show that we doubled the state of the art data rate with comparable energy efficiency and signal integrity.

Thesis outline

This thesis is organized as follows:

-We start the first chapter by introducing the standard on-chip interconnect solutions and the emerging solutions such as the wireless solution (WiNoC), the optical solution (ONoC), and the radiofrequency interconnect (RF-I). Next, we discuss the different solutions, share their advantages and challenges; then, we dive more into the standard wires solution and its limitations for long-range links. We also explain the chosen RF interconnects approach and show its advantages (mainly delay) as well as its state of the art works.

-In the second chapter, we will introduce some common modulation schemes and the duobinary modulation. Next, we study the duobinary modulation, show its implementation and main issues. Next, we show the proposed system architecture and study the adaptive length approach where the transmitter and receiver can be used for different link lengths without redesign.

-In the third chapter, we will focus on the circuits implementations. Firstly, we start with the channel choice and its EM simulation results; Secondly, we present the 10Gbps transmitter and receiver simulation results; As well as its performances up to 14Gbps. Next, we show its simulation results for a more practical case with buffers at its output. Finally, we compare our results to the state of the art results.

Finally, we conclude and propose some perspectives to develop this approach and use it for different applications such as interposer interconnects.

Chapter I : Serial Links overview

Every day, researchers innovate to meet future performance requirements in different fields (medical, automotive, telecommunications...); In recent years, our life changed because of such innovations and will continue to change as long as we push the limits. Microelectronics is driving these developments as engineers continue to make smaller and more power-efficient devices to unblock the bottlenecks. In integrated circuits, interconnects are considered as a bottleneck that we are facing nowadays. In this chapter, we introduce the emerging on-chip interconnects. We discuss the existing on-chip interconnects, their issues, and their main parameters, but also the standard solution used for the last decades. Finally, we introduce the RF-interconnect state of the art, one of the emerging interconnects solutions.

CHAPTER I : SERIAL LINKS OVERVIEW

I. An introduction into emerging On-chip Interconnects

Over the past decades, the market demanded higher performances, lower power consumption along with lower integration cost. This was possible due to the CMOS technology scaling where digital gates size decreased at a fast rate, which lead to a reduction of gate delay and silicon areas [1]. Today's transistors are at least 20 times faster and occupy less than 1% of the area of those built 20 years ago.

Due to this progress, numerous functions are now integrated on the same chip. Each function typically requires one or several cores or devices. Thus, the number of integrated cores inside systems-on-chip (SoC) grew significantly, which lead to the adoption of Networks-on-chip (NoC) as a communication network in large integrated circuits [2]. An example is shown in Figure I.1, the message is relayed from one IP core to the other through a network of routers and network interfaces.

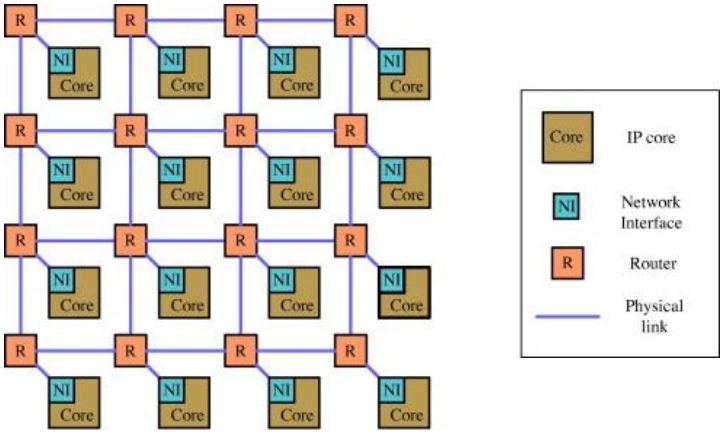


Figure I.1 : Architecture of a typical NoC.

Chips Multi-Processor (CMP) [3] was proposed by the research community to take advantage of this scaling. The area of each processor core is reduced then multiple cores are integrated on the same chip to take advantage of high parallelization.

It is clear that as the number of processor cores and memory increases, the rate of communication increases significantly. Thus, the number of required interconnects to connect different IPs raises too [4].

The research community has been working on different solutions to overcome on-chip communications challenges, such as 3D integration [5]. 3D integration adds an extra dimension for interconnection, and thus shorter distances are required, i.e., long interconnects used in 2D structures are replaced by vertical interconnects. For example, critical path gates can be placed very close to each other by stacking them and connecting them in the Z direction.

While 3D solutions are up-and-coming, 2D solutions are still prominent and still have a lot to offer without added complexity and cost. Besides improving the standard 2D wire signaling techniques, researchers proposed several solutions such as optical interconnects (ONoC) [6], wireless interconnects (WiNoC) [7], and radiofrequency interconnects (RF-I) [8]. Each solution comes with its advantages and challenges and might be optimal for specific applications.

The standard on-chip metal wire was used for a long time because it was cheap; moreover, its delay and power consumption were acceptable. These wires, also known as RC wires, are characterized by there dominant resistance and capacitance, the signal is transmitted by charging/discharging the wire [9]. RC wires delay is proportional to the inverse of the capacitance and resistance of the wire. Furthermore, it grows quadratically with length; thus, the delay will significantly increase for long-range distances. One solution to lower the latency is to use larger wires. Larger wires have smaller resistance values at the cost of a lower ratio of bit per area, i.e., reducing data throughput per surface. RC wires will be discussed in detail in section IV.2.1.

Researchers investigate optical interconnects for on-chip interconnects, it is a more radical approach to solve the expected long-range interconnects bottleneck. Adding an extra layer of optical interconnects offers a high bandwidth per channel; the transmitted signal can cross long distances with low latency due to the speed of light propagation; these interconnects are immune to electromagnetic noise. This approach is practical for clock distribution as it reduces the clock skew, as shown in [10]. One example of these links was shown in [11], where they designed an optical intra-chip link at 10Gbps per channel; they achieved a total data throughput of 100Gbps by multiplying the number of links 10x10Gbps.

Optical interconnects are very promising for long-range on-chip interconnects, but they face various vital challenges [6], [10]. Firstly, complexity because optical circuitry requires large areas and non-friendly CMOS processes. Power consumption and thermal regulation are an essential challenge. Even though the channel losses are negligible, the power consumption of current optical devices, e.g., light sources, modulators... required to generate the signal is significant. Thus, from a power consumption view, the optical solutions will only be interesting for long-range communications, the distance at which they become optimal has still to be defined. To summarize, optical solutions will be used more often once they become CMOS manufacturing and assembly processes friendly, and with a high yield. Furthermore, to become financially attractive, their cost should be reduced.

Another approach is to use wireless interconnects, also known as wireless NoC (WiNoC) [7]. This approach transmits the signal as an electromagnetic signal using on-chip antennas. Thus, this solution can connect all adjacent nodes with very low latency, taking advantage of the shared medium of transmission. The scaling of transistors enables the use of higher carrier frequencies and thus avoid transmitting in the currently occupied spectrum. Higher frequencies in the mmWave range require small integrated antennas [12], and researchers are striving to design efficient wideband antennas at these high frequencies [7], since the antenna bandwidth is the limiting factor in this approach. Implementation in the THz range is also interesting because of large available bandwidth and small antennas, but the CMOS transistors cut-off frequency becomes the limiting factor in this case. Furthermore, a multi-carrier transmission will require different antennas with different center frequencies resulting in significant area overhead.

Instead of free space signal propagation, waveguided propagation can be used to transmit the signal. This approach is called RF interconnects (RF-I) [3], [8] and uses low-loss transmission lines (TL) to propagate the high-frequency signal over long distances. This solution uses the standard CMOS processes, and do not require any changes. Moreover, it benefits from faster transistors due to CMOS scaling. It offers a

low-latency transmission since the signal travels along the TL at the effective speed of light in SiO_2.

Furthermore, when using TL, a large bandwidth is available, allowing for high data rates transmission with minimum distortion. To reduce the attenuation, the lines are implemented using the higher metals in the BEOL (Back End Of Line) stack because of their thickness. But also, the TLs require wide wires, which means that the number of these links is limited. Another challenge to consider is the crosstalk between these lines, i.e., the lines should not be placed too close to each other to avoid coupling. The dimensions of the TLs and the spacing between them limits the interconnect density. Thus the efficiency of each link should be improved to reach a higher total data throughput.

The RF-I approach is more power-efficient than the standard wire solution for long-distance links, due to the low attenuation of the transmission lines. The standard solution uses power-consuming repeaters to regenerate the signal periodically; longer distances require more repeaters.

In this work, we will discuss the standard wire solution briefly, and we will focus on the RF-I solution since we believe it is not efficiently used yet, and better performances can be reached.

The previously techniques mentioned are different, and none of them is ideal, each method has its advantages and drawbacks that might be suited for unique applications and uses. For example, wireless propagation (WiNoc) is suitable for broadcasting to all receivers. Or use optical solution and RF-I when the latency is critical. And lastly, use the standard wire solution for short distances where the other solutions will not be suitable. Table I summarizes the different approaches, advantages, and challenges, but a more detailed comparison can be found in [13].

	RC Wires	Optical interconnects	Wireless interconnects	RF interconnects
Advantages	-Cheap and CMOS process friendly. -Simple implementation for low data rates.	-Very large bandwidth. -Low latency due to propagation at speed of the light. -Low losses in optical waveguides.	- CMOS process friendly. -Broadcasts to different neighbor receivers. -Low latency due to high speed propagation.	-Cheap and CMOS process friendly. -Low latency due to high speed propagation. -Power consumption is acceptable.
Challenges	-Delay grows quadratically with length without repeaters. -Repeaters insertion adds to floorplan constraints. -Complex circuitry is required for high data rates.	-Some devices are not CMOS compatible. -High power consumption.	-Interference from adjacent antennas and environment. -Bandwidth limited by antenna or CMOS transistors cut-off frequency. -Multi-band transmission requires several antennas.	-Crosstalk between the TLs if not shielded. -Large area required to implement the TLs.

Table I.1 : Advantages and challenges of differents emerging on-chip interconnects.

II. High speed serial link architecture

A typical high-speed serial link includes three main elements, a transmitter (TX), a channel, and a receiver (RX). In a serial link, most of the signal integrity issues are due to the channel's low-pass response, making the transmitted signal sometimes unrecognizable to the receiver. Ideally, a transceiver should transmit and recover the data without errors, use a small area, and consume very little power.

The transceiver's complexity depends on several parameters, such as the channel properties, the expected data rates, and the allocated power budget. Usually, designers compromise on one of several parameters to achieve the specifications.

Figure I.2 shows the architecture of a typical high-speed serial link. The transmitter multiplexes low-speed parallel data into a high-speed serial stream and then transmits it through the channel. To increase the data quality and consequently the data rates, designers use equalizers and pre-emphasis techniques in the transmitter (or the receiver) to compensate for the channel's imperfection, such as frequency-dependent losses.

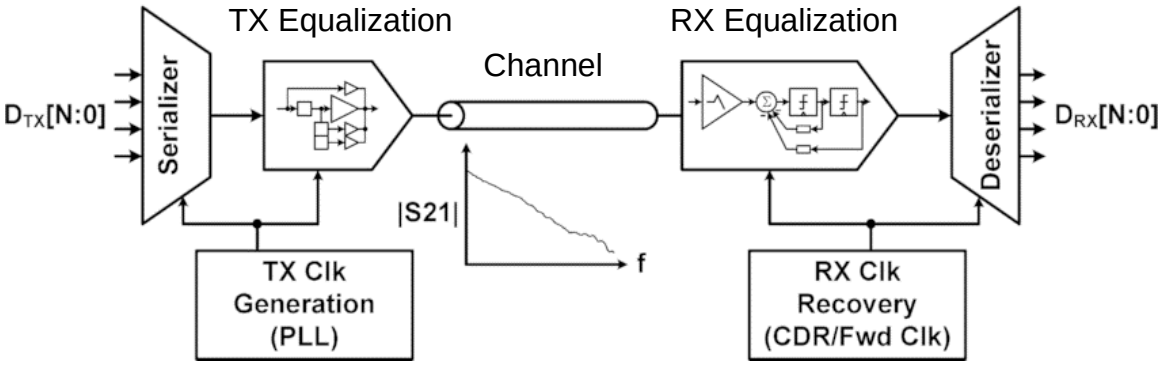


Figure I.2 : Architecture of a typical high-speed serail link.

Equalization is performed at both ends of the transceiver when the channel degrades the signal severely; for example in RC Wires cases. Hence, when the signal arrives at the receiver, it goes through an equalization stage before sampling. Sampling the received data at the proper rate is crucial for a reduced number of errors. However, in most high-speed serial links, the clock is not transmitted. Therefore, the receiver needs to recover the clock to sample the data accurately. In other words, detect the most suitable sampling location in a jittery signal to produce a correct signal.

High-speed serial links use a Clock Data Recovery (CDR) to regnerate the clock from received data, then uses the recovered clock as the reference to trigger a retiming flipflop to clean up the incoming data as shown in Figure I.3.

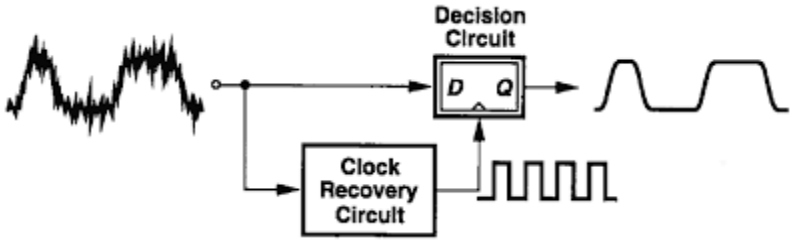


Figure I.3 : Clock recovery and data retiming for a CDR circuit.

The complexity of the CDR circuitry depends on the link type. We can distinguish two types of links: source-asynchronous and source-synchronous systems.

A serial link is called source-synchronous when both the TX and RX use the same clock source. For this type of link, the CDR only provides a finite phase capturing range and requires a lower complexity.

However, most of wireline serial links belong to the second category and are called source-asynchronous. For this link, the TX and RX use different clock sources, which results in a frequency offset between the received data and the RX local clock, and thus increases the complexity of the needed CDR. The CDR usually drives several transmitters and receivers for efficient use of the area and power. The CDR clock can also be used to drive the equalizers.

As we will see in the equalization techniques section, most of the equalizers are discrete-time equalizers that require a clock. Thus, equalizing at both TX and RX means that a clock should be available at both ends.

Researchers proposed a wide variety of CDR architectures based on Phase-locked loop (analog and digital) and delay-locked loop (DLL) amongst different designs. These architectures will not be discussed in this work. However, the authors in [14] provided an overview of these architectures.

II.1. Line coding

Line code is the code used to transmit digital data over a channel such as an RC wire or a TL. These codes are used to convert a sequence of bits into a digital signal, and efficiently use the channel's bandwidth to reduce noise and interference. Line codes cover unipolar, polar, bipolar signals, and multilevel signals. Each can be divided into subcategories, but the most common ones are the Non-Return to Zero (NRZ) and Pulse amplitude modulation (PAM). Figure I.4 shows an example of unipolar Return to Zero (RZ) and NRZ signals; the binary states are easily distinguishable in the transmitted sequences,i.e., the presence of a signal is equivalent to a binary '1'; its absence is a binary '0'. Square signals are preferred because their generation and detection processes are simple, especially that most of the logic digital circuits function in a two-states approach, i.e., on and off. However, analog signals do not follow the same rule, and they can take an endless variety of forms.

Choosing the appropriate code depends on several factors, such as the presence or absence of DC component, simple regeneration, and power efficiency [15].

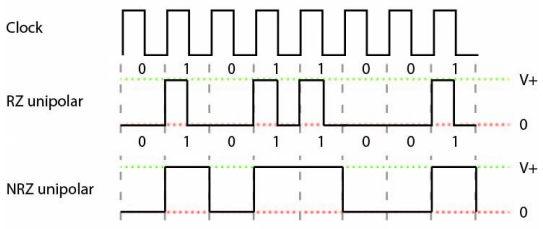


Figure I.4 : An example of unipolar signals (RZ and NRZ).

II.2. Bandpass modulation

Bandpass modulation [15] techniques modify the shape of a carrier wave (usually sinusoidal) to encode the transmitted data; In other words, shift the signal up in frequency. In this case, the spectral magnitude is nonzero in the band around the carrier frequency and is negligible elsewhere. This transmitted data is a baseband waveform (analog or digital) and is called the modulating signal. For example, it can be the NRZ or RZ signals shown in Figure I.4. This technique is useful for a channel that acts as a bandpass filter, for shared channels where only certain frequency bands are available. We can distinguish three categories of modulation: amplitude, frequency, and phase modulation. The carrier's amplitude is varied in proportion to the data signal in amplitude modulation; While its phase is varied in phase modulation. Newer modulations called Quadratic Amplitude Modulations (QAM) combine amplitude and phase carrier modulations to define more symbols in the constellation, increasing the data rate with the same bandwidth. In frequency modulation, the frequency of the signal is altered in proportion to the data.

The inverse operation is called demodulation, and it consists of recovering the data from the carrier at the far-end, usually the receiver. Each modulation has its advantages/disadvantages that fit specific applications.

III. Serial link performance metrics

In serial links design, compromising between the different parameters is essential, although some metrics are more critical than others. These include data rates; The data rate of a serial link is defined as a the number of bits transferred per second. To satisfy the need for high data rates, serial links advanced from several Gbps to several tens of Gbps, at the cost of complexity and power consumption. Hereafter, we introduce the main tools and metrics a designer uses to identify whether a transmission is reliable or not.

III.1. Eye Diagram

Useful tools such as eye diagram are used to visualize high-speed serial links signals in time domain. An eye diagram [16] indicates the quality of the signals transmitted, and is generated by overlapping successive waveforms corresponding to the received signal. These waveforms can be the bits of a long data stream.

It is a very powerful tool that allows the designer to intuitively interpret the quality of the signal, by looking at the area in the middle of the eye diagram named the "eye" or eye-opening. The eye closes vertically if the signal-to-noise-ratio (SNR) is degraded, mainly due to noise and Intersymbol-interference (ISI). Signal to Noise ratio (SNR) is simply the ratio of the power of a signal to the power of the noise and is generally expressed in decibels (dB).

The noise is identified at the top and bottom lines of the eye. Thicker lines are an indicator of the presence of more noise. Thus, if the eye closes vertically, the detection of different states becomes difficult.

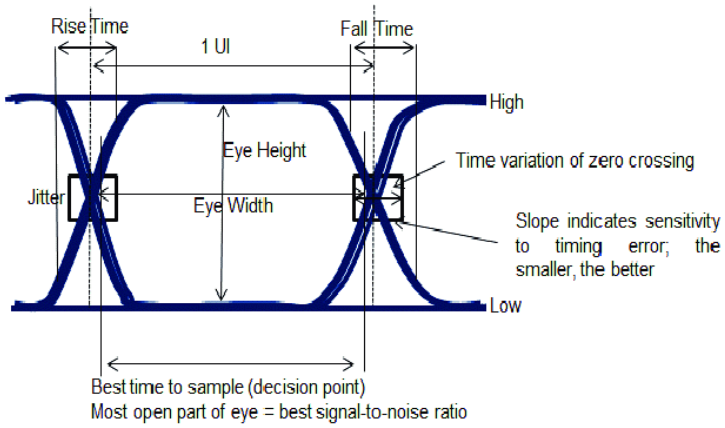


Figure I.5 : An Eye diagram.

III.2. Bit error rate

Another critical metric is the number of error bits per transmitted bits. Bit error rate (BER) is one of the popular metrics used to assess the performance of a communication system and prove the reliability of the data transmission. BER indicates the rate at which errors occur in a transmission system, it can be calculated by comparing the transmitted sequence of bits to the received bits and counting the number of errors. Or in other words, it is the ratio of the number of error bits over the total received bits as given by:

$$BER = \frac{Nbr\ of\ errors}{Total\ number\ of\ bits\ sent} = \frac{N_{err}}{N_{bits}} \tag{I-1}$$

If the link between the transmitter and receiver is good i.e. no noise or jitter, then this value will be very small. In serial links. In practice, the number of bits sent N_{bits} is chosen depending on the BER threshold and its confidence level (CL). CL is the percentage of tests that the system's true BER (if $N_{bits} = \infty$) is less than the BER threshold. The confidence level formula is given by:

$$CL = 1 - e^{-N_{bits} * BER} \tag{I-2}$$

In standard wireline communications, a BER of 10^{-12} and a *CL* of 95% is often required to avoid using error-correcting codes. We should expect no errors if 3×10^{-12} bits are sent, if errors are detected the mathematical expressions can be found in [17].

As mentioned previously, noise and jitter can degrade the communication channel and its BER. Figure I.6 shows the statistical relation between BER and the signal to noise ratio (gaussian noise) per bit E_b/N_0 for different types of modulations.

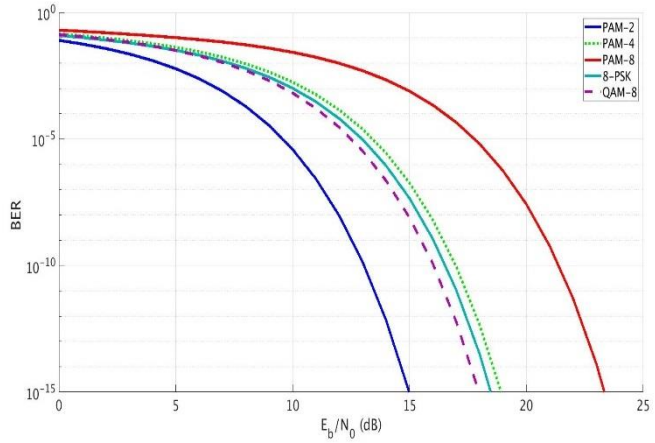


Figure I.6 : BER versus E_b/N_0 for different modulations.

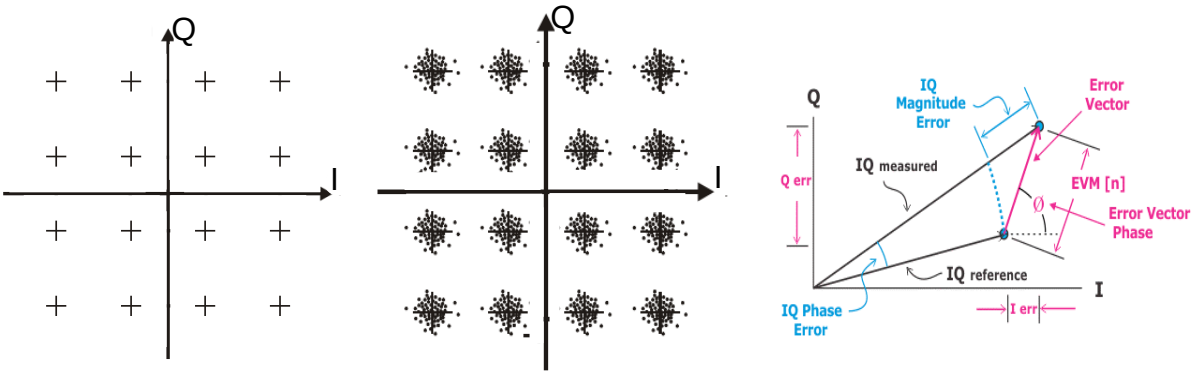
The normalized signal to noise ratio (E_b/N_0), also known as SNR per bit is usually used to compare digital modulations. It is the ratio of energy per bit (E_b) to the spectral noise density (N_0). The energy per bit E_b is the signal energy of a single bit; it can be computed by dividing the signal power by the bit rate; it is expressed in joules. N_0 is the noise spectral density (noise in $1Hz$ bandwidth); it is expressed in Joule per Hetz.

As mentioned previously, bit errors can be caused by amplitude noise degrading the SNR or by to timing uncertainty,i.e., jitter. In this case the demodulator or receiver can detect the wrong bit, which may be the previous or the next bit.

III.3. Error Vector Magnitude

Error Vector Magnitude (EVM) measures digitally modulated signal quality; it expresses the difference between the voltage of the expected symbol and the received one. EVM provides extra information compared to BER, where the BER reports the number of bits containing errors only. Constellation diagrams are used to visualize the demodulated signals and the EVM. A constellation diagram is a plot of symbols represented by their unique magnitude (and sometimes phase) voltage value.

Figure I.7.a represents a 16QAM constellation diagram, the constellation contains 16 unique symbols with different amplitudes and phases. Figure I.7.b represents measured EVM; the small dots on the diagram represent the errors in the measured symbols. Figure I.7.c shows a single error vector magnitude presentation; the error vector magnitude is the vector's length that connects the reference symbol position and the measured symbol position.



Ideal constellation Diagram.

Measured constellation Diagram.

Figure I.7 : (a) Ideal constellation diagram. (b) Measured constellation diagram. (c) Single error vector manitude calculation.

III.4. Inter-symbol interference

Inter-symbol interference (ISI) is a form of signal distortion that happens when one symbol interferes with previous or subsequent symbols, it is caused mainly by the dispersion of the channel and its finite bandwidth, the majority of interconnect channels have a low-pass frequency response. Passing a signal through such channel results in attenuation of the high-frequency components, furthermore, in such a medium the signal components will travel at different speeds and thus arrive at different moments. The distortion of the channel results in the pulse spreading out and interfering with its previous symbols, as shown in Figure I.8. Thus, depending on when the pulse is sampled, the receiver can make the wrong decision. The performances and reliability of these links depend on several figures of merit such as bit error rate.

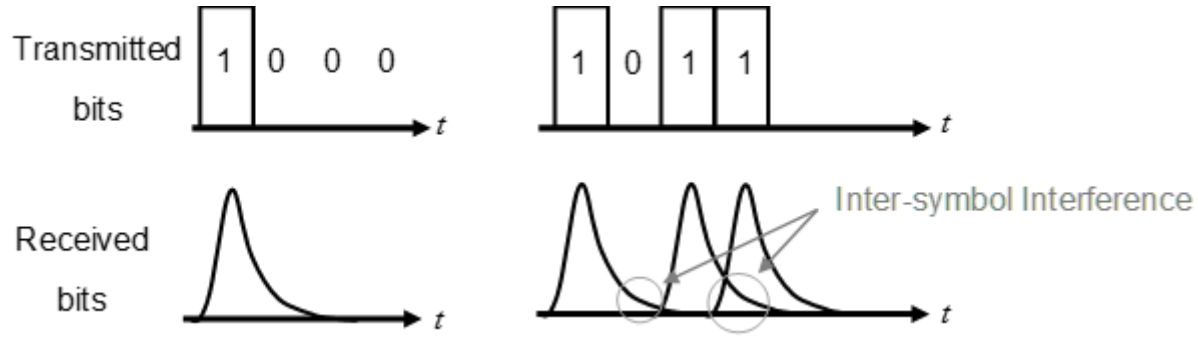


Figure I.8 : Inter-symbol interference.

III.5. Jitter

The eye diagram in Figure I.5 closes horizontally if the jitter increases. The eye diagram is rich in information and can be further utilized to extract the jitter histogram, and the Total jitter TJ.

Jitter is the temporal deviation of the signal from its ideal value at a certain point, as shown in Figure I.9.

In digital transmissions, the jitter is defined at the transition from $0/1$ and $1/0$, and it can be written as $t = T + \varphi$ where t is the time of the transition, T its ideal position and φ the time offset of the transition also called timing jitter.

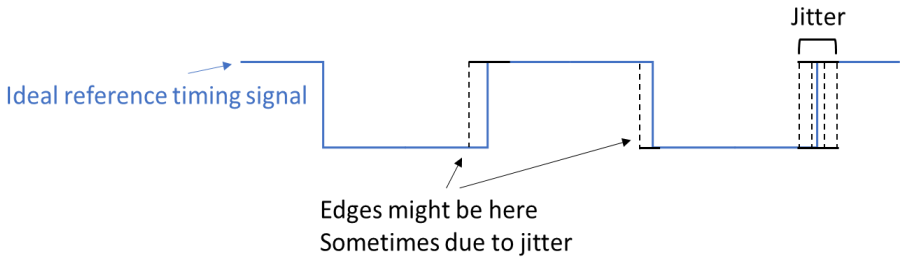


Figure I.9 : Waveform timing variations.

In analog transmissions, jitter is known as phase noise given by:

$$S(t) = P\big(t + \varphi(t)\big), \tag{I-3}$$

where $S(t)$ is the jittered signal, $P(t)$ the ideal signal, and $\varphi\ (t)$ the phase noise. The jitter is defined either using time units (ps and ns) or in unit interval [UI]. The UI is the proportion of total jitter TJ per bit and is given by:

$$Jitter_{(UI)} = \frac{TJ}{T_{bit}}, \tag{I-4}$$

The total jitter is decomposed into two main categories: Random jitter (*RJ*) and Deterministic jitter (*DJ*) as shown in Figure I.10.

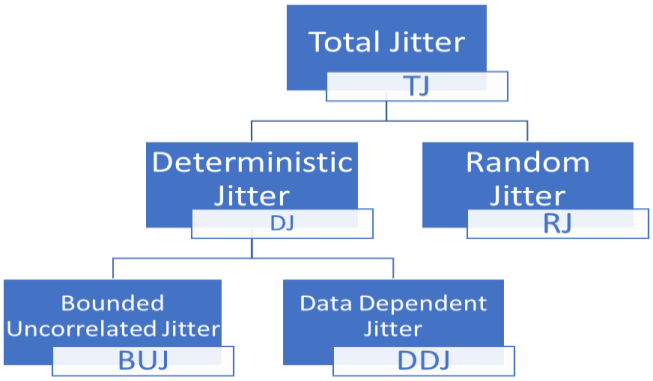


Figure I.10 : Jitter components [17].

Random jitter covers the unpredictable noise introduced randomly by thermal noise, shot noise, or "Pink" noise in the circuits. It is described probabilistically and usually follows a normal distribution. In theory, it is a constant value generally expressed as the root mean square (rms) value equivalent to its standard deviation σ.

The rms random jitter RJ_{rms} derived from the standard deviation can be converted to a peak-to-peak RJ_{pp} using the formula given by:

$$RJ_{pp}(BER) = N(BER) \times RJ_{rms}, \tag{I-5}$$

This formula relates the RJ_{pp} to RJ_{rms} through a BER function, $N(BER)$ depends on the required BER. for a BER of 10^{-12}, $N(BER)$ is equal to 14.069.

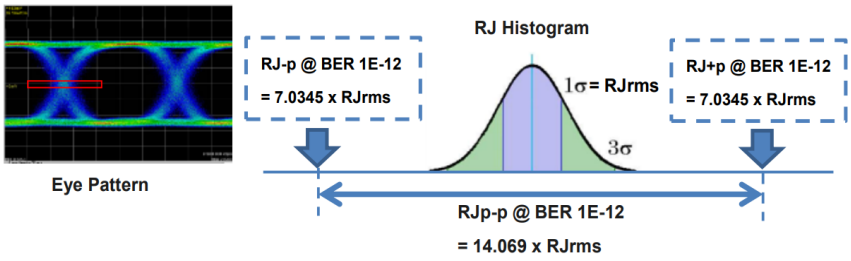


Figure I.11 : Typical waveform including RJ, and RJ Histogram[19].

Deterministic jitter represents the jitter with an identifiable and predictable source. Its peak-to-peak value is bounded with a well-defined minimum and maximum. As shown in Figure I.10, deterministic jitter covers data-dependent jitter (DDJ) and bounded uncorrelated jitter (BUJ). Each component can be decomposed to subcategories to diagnose the causes of the jitter more precisely.

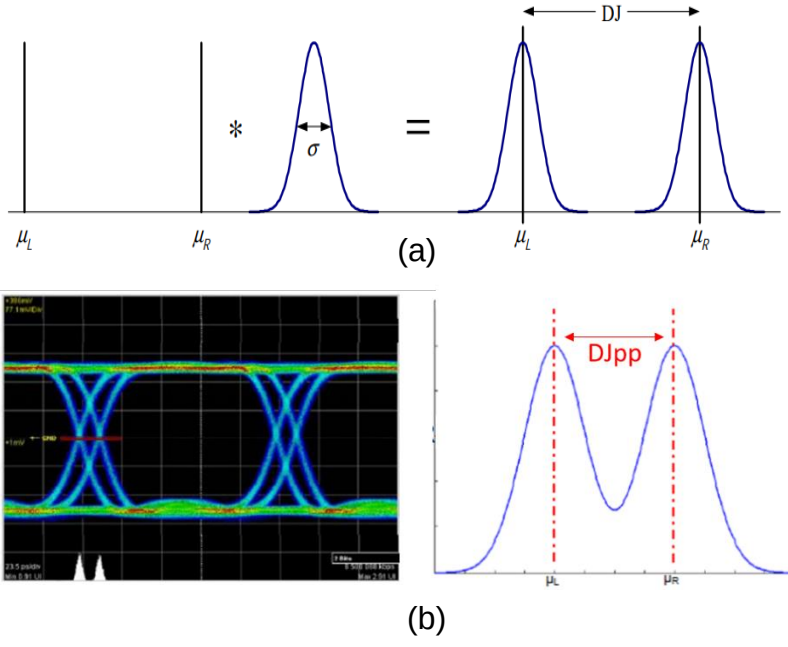


Figure I.12 : (a) Dual dirac method . (b) Typical DJ Waveform, and histogram[19].

To rapidly estimate TJ [20] the sum of DJ and RJ, dual dirac method is used.This method describes TJ as two Dirac delta functions convolved with a gaussian function, the distance between μ_L and μ_R is the deterministic jitter $DJ(\delta\delta)$, while the dashed lines shown in Figure I.12.b represent the random jitter.

The advantage of this model is that it does not require to calculate the different components of deterministic jitter individually. The formula for TJ as a function of RJ and DJ is given by:

$$TJ(BER) = N(BER) * RJ_{rms} + DJ(\delta\delta), \quad \text{(I-6)}$$

To evaluate the SNR and the jitter in a system quickly, engineers use eye diagram and bathtub curves.

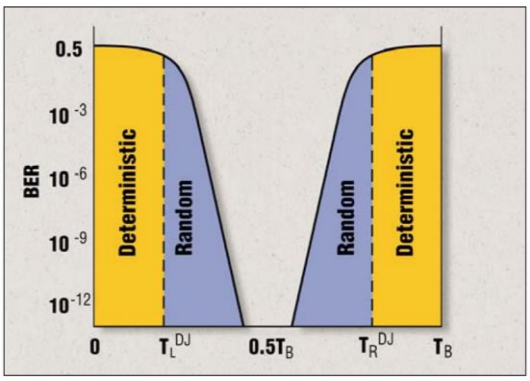


Figure I.13 : a bathtub plot showing the BER as a function of sampling point delay.

Another powerful tool to analyze jitter is the bathtub plot [21]; This plot can be used to determine the acceptable total jitter in the system for a specific BER.

The bathtub plot is shown in Figure I.13. It represents the BER versus sampling point within the unit interval. We can distinguish three regions, the deterministic jitter region, the random jitter region where the BER falls rapidly, and the third region is the eye-opening. To open the “eye” and improve the BER, designers use several techniques such as equalization to counter the channel distortion effect, or error correcting codes, these techniques are often combined.

III.6. Error correcting codes

Error correcting codes [15] (ECC) was introduced by Hamming [18] in the '40s, and has been crucial since then for telecommunications systems that cannot re-send the message, such as interplanetary communication; Or for when a 100% reliable system is not affordable i.e. a less reliable system combined with ECC costs less.

In this approach, a portion of the signal bandwidth can be exchanged for a better transmission fidelity, when the communication channel is noisy or unreliable, redundant information is added to the original data; these extra bits allow the detection and correction of the errors if they exist.

We can distinguish two types of ECC; the first type is called block codes where the correction bits are added as a block at the end of the original bitstream; for this type of ECC the original bitstream must have a fixed length. The second type is called convolutional codes, where the correction bits are added continuously into the bitstream. Furthermore, contrary to the block code, the bitstream does not have a beforehand fixed length. However, these flexibilities of the convolutional codes compared to the block code comes at a complexity cost.

The code-rate of an ECC is the ratio between the code correction bits and the total number of bits (bitstream plus the code correction codes), a high code-rate means that high reliability is achievable. Still, it also means that a large number of bits are added to the original data. Thus, a tradeoff between data rate and channel reliability is always necessary.

IV. On-chip Interconnects

In recent ICs, on-chip global signaling became a bottleneck for the required data rates. Firstly, due to the production of larger chips, which leads to longer links. But also due to the wire pitch decreasing in smaller nodes [1], which significantly increases the resistance and increases the parasitic capacitance due to the dense routing.

In the first stages of the design process, to connect the nodes and devices, we usually use ideal wires that show no attenuation or latency. This first approach is not wrong, since modeling all wires is too complicated, and depending on the case, some parameters might be negligible compared to more dominant parameters. For example, for a wire with small cross-section, the inductance can be neglected since the resistance will be more significant. Hence, a quick review of on-chip wires main properties is desirable; we will use first-order models to understand these effects.

IV.1. RLCG model

Wires are used to transmit energy from one point to another using two conductors or more; its source is usually TX driver and load RX. The term transmission line can be used for any structure that guides an electromagnetic wave; its parameters are usually optimized to avoid wasting energy during the transmission.

TL effects are not considered in low frequencies usually. However, at high frequencies, they affect the power transfer and are considered even for very short distances. In general, if a transmission line has a length higher than 1/16 of a wavelength, then the line length will significantly affect the circuit's impedance. Figure I.14.a shows a typical representation of a two-wire transmission line.

Figure I.14.b shows a uniform network distributed along the two-wire lines where each network represents an infinitesimal length Δz of the line.

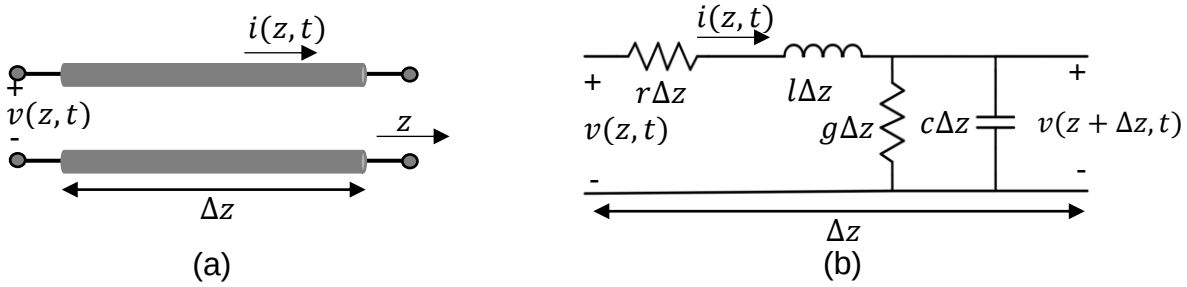


Figure I.14: (a) An incremental length of a transmission line. (b) Lumped-element model.

When a signal is injected into the transmission line, it creates electrical and magnetic fields around the line. The energy stored in the electrical field between the line and its ground is modeled as a shunt capacitor $c\Delta z$. The energy stored in the magnetic field around the line is called the series inductance $l\Delta z$.

The series resistance $r\Delta z$ represents the losses due to the finite conductivity of the transmission line. The shunt conductance $g\Delta z$ represents the losses due to the non-ideality of the dielectric surrounding the TL.

The main parameters to describe the network in Figure I.14 are resistance, inductance, capacitance, and conductance; these parameters are described per unit of length.

The voltage V and current I at a point z along the transmission can be described as [23] :

$$\frac{\partial V}{\partial z} = -rI - l\frac{\partial I}{\partial t}, \tag{I-7}$$

$$\frac{\partial I}{\partial z} = -gV - c\frac{\partial V}{\partial t}, \tag{I-8}$$

To describe the signal propagation along the transmission line: we differentiate equation (I-7) with respect to z, then insert equation (I-8) into its results, which gives the wave propagation equation (I-9):

$$\frac{\partial^2 V}{\partial z^2} = rgV + (rc + lg)\frac{\partial V}{\partial t} + lc\frac{\partial^2 V}{\partial t^2}, \tag{I-9}$$

For most on-chip interconnects, we can consider that the leakage conductance g is equal to 0. Thus the formula becomes :

$$\frac{\partial^2 V}{\partial z^2} = rc\frac{\partial V}{\partial t} + lc\frac{\partial^2 V}{\partial t^2}, \tag{I-10}$$

with r,l,c the resistance, inductance, and capacitance per unit length, respectively.

Two types of signal transmission can be considered depending on which term dominates in equation (I-10). At low frequency, i.e. up to few GHz, the rc term dominates since the inductance equivalent impedance $l\omega$ is much lower than the resistance r. At high frequency, i.e. above a few GHz, depending on the transmission line type, the lc term in equation (I-10) becomes the dominant one, and the propagation is mainly controlled by the linear inductance and capacitance of the transmission line.

Nowadays, designers use advanced parasitic extraction tools to extract the capacitances, inductances, and resistances from the layout. This step is necessary since these parameters highly depend on the environment, as we will see in the next section. For example, the parasitic capacitance of a wire depends on its distance from its substrate, environment, and distance to its immediate neighbors. These parasitics should be accurately identified beforehand to avoid unexpected crosstalk and signal degradation.

IV.1.1. Interconnect parameters

IV.1.1.1 Interconnect capacitance

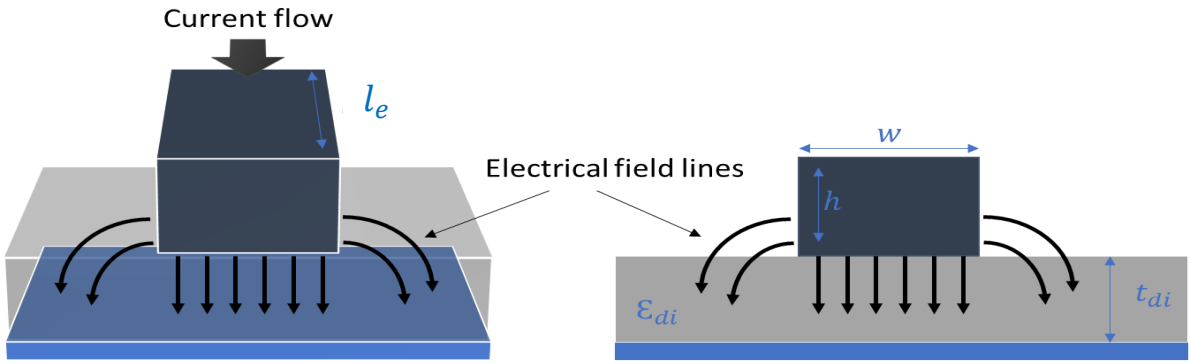


Figure I.15 : Microstrip line and its electrical field lines.

Figure I.15 shows a metal wire (or a strip) over a ground plane; this structure is called a microstrip [23], and it is one of the most popular types of planar transmission lines. w, h, l_e are the strip width, height, and length, while t_{di} is the thickness of the dielectric between the bottom of the strip and the ground plane, and ε_{di} its relative permittivity.

By injecting a voltage signal into the strip, we create an electrical field that is concentrated below the strip and is orthogonal to the ground plane. But also, the rising of a capacitance C_{tot}, which is the sum of the parallel-plate capacitor C_{pp} , and the fringe capacitance C_{fringe} due to the electrical field at the edges of the wire [24].This total capacitance of the isolated wire can be approximated by the formula (I-11):

$$C_{tot} = C_{pp} + C_{fringe} = \frac{w\varepsilon_{di}}{t_{di}} l_e + \frac{2\pi\varepsilon_{di}}{\ln(2 + 4\,t_{di}/h)} l_e, \qquad \text{(I-11)}$$

As mentioned earlier, this is a simple approximation that isolates the wire from its environment. Figure I.16 shows a wire in a more realistic environment with wires at different metal levels and different distances from each other.

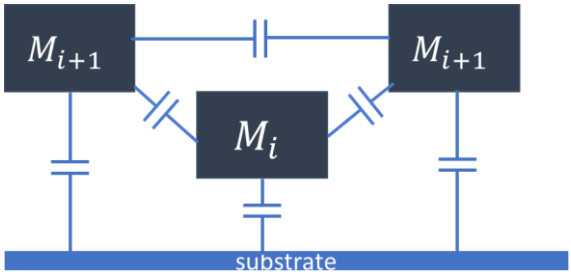


Figure I.16 : Capacitive coupling between wires.

Contrarily to the previous case where the wires parasitic capacitances were grounded, in Figure I.16 the wires are coupled to neighbor wires. This phenomenon is called crosstalk and will be introduced in section IV.2.1.5. Crosstalk, along with the non-ideality of the channel, e.g., its limited bandwidth and delay, reduce the quality of the signal significantly.

IV.1.1.2 Interconnect resistance

The DC-resistance of the microstrip shown in Figure I.17 is given by:

$$R_{DC} = \frac{\rho}{A} l_e, \tag{I-12}$$

where ρ is the resistivity of the material $[\Omega.m]$. The resistance of the wire is proportional to its length l_e and inversely proportional to its cross-section $A = h.w$. This equation can be generalized to any rectangular conductor.

More often, we compute the wire resistance from the sheet resistance formula $R_{sq} = \rho/h$. The sheet resistance of the material is always stated in the design manuals, it defines the resistance of a square conductor of a dimension h, h is well-defined in each technology, and its unit is $\rho/square$.

Nowadays, Copper and Aluminum are the most commonly used metals; this is due to their compatibility with the process. However, Copper is preferred due to its lower resistivity.

Figure I.17 : Skin effect flow primarily on the surface of the conductor.

At higher frequencies, the resistance is not linear. This non-linearity is due to a phenomenon called the skin effect that increases the resistance with the frequency; thus, it becomes frequency-dependent. Most of the high-frequency current flows through the surface or the 'skin' of the “standalone” conductor, as shown in Figure I.17, which decreases the effective cross-section area carrying the current. The current density falls off exponentially with depth into the conductor. This parameter is called the skin depth, and it is the depth at which the current decreases to e^{-1} (about 0.37) of its nominal value, i.e., 63% of the current flow in one skin depth (in m) given by:

$$\delta = \sqrt{\frac{\rho}{\pi f \mu}}, \tag{I-13}$$

with f the frequency of the signal, μ the permeability of the surrounding dielectric.

The skin-effect phenomenon starts to show above a specific frequency called the skin frequency f_s; this frequency is defined as the frequency where the skin depth is equal to half of the largest conductor dimension $D = max(w, h)$ and is given by :

$$f_s = \frac{4\rho}{\pi \mu D^2}, \tag{I-14}$$

To determine the resistance of on-chip interconnects for different frequencies, the formula given in [4] can be used :

$$R(f) = f(x) = \begin{cases} R_{DC}, & f < f_s \\ R_{DC}\sqrt{\frac{f}{f_s}}, & f \geq f_s \end{cases}, \quad \text{(I-15)}$$

The conductor losses are highly affected by its current density distribution. Figure I.18 shows the current density in a microstrip at high frequencies; it is concentrated in the bottom conductor surface. We notice that it is maximal at the surface edges due to charge density increasing significantly near sharp edges [9]. In fact, the majority of losses occur at this surface of the conductor. More details about microstrip losses can be found in [9], [25].

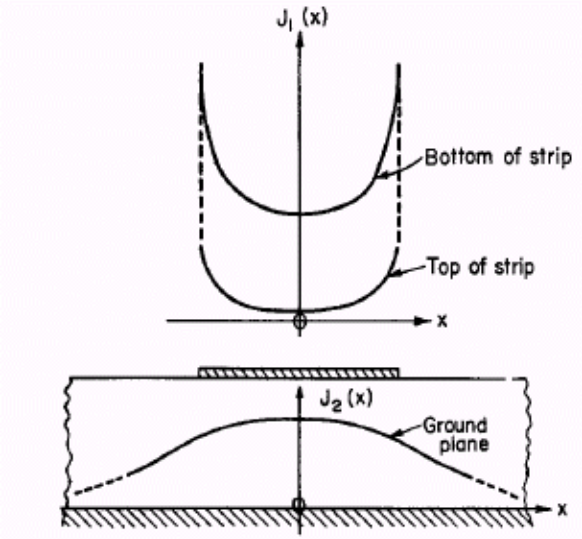


Figure I.18 : Sketch of current distribution on a microstrip conductor [25].

IV.1.1.3 Interconnect Inductance

At very low frequencies, inductance is usually dismissed in integrated circuits due to its negligible impact. However, at higher frequencies, inductance can cause severe signal integrity issues [9], [22] such as overshoot effects, ringing, reflections of signal, and switching noise due to the $L\frac{di}{dt}$ voltage drops.

An inductance is induced when a current flows through a conductor, thereby creating a magnetic field around it [23]. Inductances follow a loop property, i.e., the inductance of a wire is determined if both the current path and its return path self-inductance are known.

The current favors the return path with of least impedance [22], thus at low frequency, the current will flow back through the least resistive path, while at higher frequencies, the current will flow back through the closest path with least inductance.

In integrated circuits, the wires are placed close to each other, as shown in Figure I.19. and the exact return path is not precisely defined as it might be spread over a broad zone [26]. Thus, the return paths might be the adjacent wires and the power distribution network.

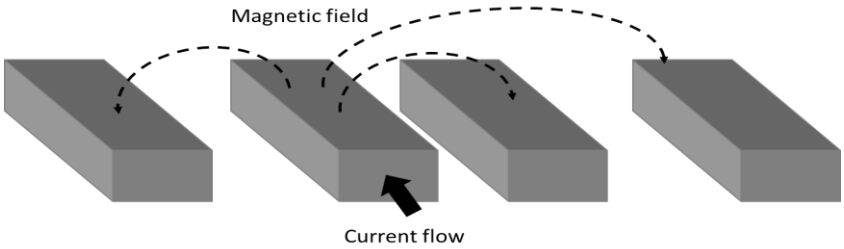


Figure I.19 : Mutual inductance between wires

Therefore, extraction software is used to find the total inductance of the wire by decomposing it into internal and external inductances. The external inductance is due to the magnetic flux surrounding the wire, i.e., taking into consideration the environment and the nearby signal paths. The internal inductance is due to the magnetic flux inside the wire [26], [27].

The inductance of a wire's segment can be calculated from the fundamental relation between the inductance fundamental definition and Lenz law and is given by:

$$\Delta V = L\frac{di}{dt} \tag{I-16}$$

This formula describes the voltage drops by ΔV when a changing current flow through the inductance.

IV.2. RC wire

Different electrical models were developed to introduce the effects of the resistance, capacitance, and inductances of wires on the circuits. These models try to approximate the real behavior of the wires. However, depending on the required accuracy and resources, different orders of approximation can be used.

IV.2.1. RC wire models

IV.2.1.1 Lumped model

Firstly, the lumped model; this approach can be used when only one parasitic component is dominant. The parasitic elements distributed along the wire can be lumped into one component for simplicity. As shown in Figure I.20, the capacitances previously distributed along the wire are combined into one capacitance; This is possible when the wire's resistive component is negligible, and the circuit's characteristic length is vastly lower than its operating wavelength.

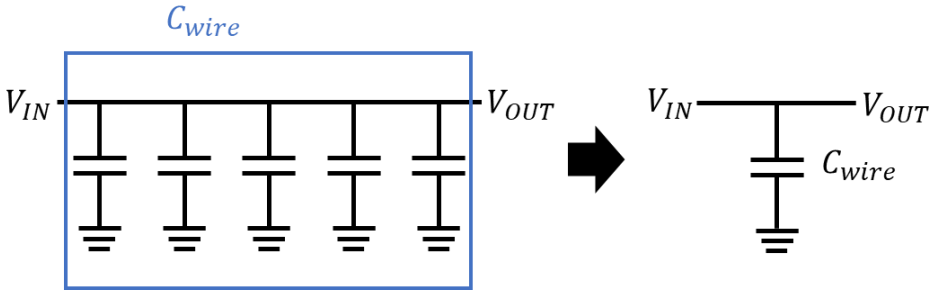


Figure I.20 : Distributed and lumped model for a capacitance dominant wire

Consequently, ordinary differential equations can be used instead of the partial differential equations required for the distributed elements.

The previous assumption that all the capacitances of the wire can be lumped into one capacitance does not hold when the wire has a significant resistance. Consequently, a more appropriate approach will be to consider both R and C of the wire. Thus, the wire cannot be considered as an equipotential segment anymore. This model is called the RC lumped model. In this model, each wire has its resistance and capacitance. This model is only appropriate for very short wires, as its accuracy falls for long wires.

IV.2.1.2 Distributed RC line

For RC dominant wires, it was shown [28] that using a distributed model improves the accuracy of the model.

A wire can be decomposed into several segments, as shown in Figure I.21. D represents the total length, $r\Delta D$ and $c\Delta D$ are the resistance and capacitance per unit length, respectively. It was shown in [28] that the spatial derivative of the voltage could be related to its time derivative by the diffusion equation which is equal to equation(I-10) when the inductance is neglected and is given by :

$$rc\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial z^2} \tag{I-17}$$

V is the voltage at a particular point in the wire, z the distance between the point and the source. When the diffusion equation closed-form solution is not obtainable, these wires can be approximated by a lumped RC chain [29], [30].

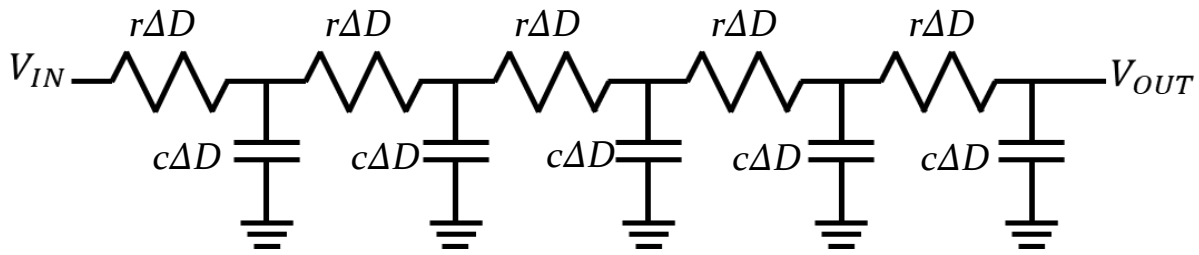


Figure I.21 : RC Distributed line.

IV.2.1.3 Elmore delay

Elmore Delay formula [31] can be used to extract the delay easily. The formula is given by:

$$\tau_N = \sum_{i=1}^{N} (i.r.\Delta D) \times c\Delta D = rc(\Delta D)^2(1 + 2 + \cdots + N) \tag{I-18}$$

The wire in Figure I.21 is composed of N equal segments, with $r\frac{D}{N}$ and $c\frac{D}{N}$ the resistance and capacitance of each segment.

Thus, the delay can be computed using the formula :

$$\tau_N = (rcD^2)\frac{N(N+1)}{2N^2} \quad \text{(I-19)}$$

With $\sum_{i=1}^{N} i = \frac{N\,(N+1)}{2}$; We notice that if N is large, the delay equation can be approximated to :

$$\tau_{DN} = \frac{rcD^2}{2} \quad \text{(I-20)}$$

Equation (I-20) shows that the delay of a wire is a quadratic function of its length D. To reduce this effect, designers use wider lines to reduce the resistance, at the cost of higher capacitance. Furthermore, increasing the width of wires reduces the density of the interconnects.

IV.2.1.4 Repeaters insertion

Another solution to minimize the total propagation delay of a D long wire is to insert repeaters [32]. Where the D long wire can be divided into several segments [33] with smaller delay. Usually, the wire k sections are identical, and a repeater is used to drive each segment.

Thus, the delay of each segment becomes:

$$\tau_{DK} = \frac{rcD^2}{2k^2} \quad \text{(I-21)}$$

However, we notice that the required number of repeaters increases for longer interconnects significantly. This technique results in a larger area and more routing constraints but also increased power dissipation in on-chip interconnects.

The repeaters increase the delay too, and the total delay is the sum of the delay of k section, where each section is composed of an interconnect and a repeater. Bakoglu suggested in [29] a method to evaluate the optimal size of the buffers and the optimal number of repeaters k for the shortest total delay, the formulas were given:

$$n = \sqrt{\frac{R_0 C_W}{R_W C_0}} \quad \text{(I-22)}$$

$$k = \sqrt{\frac{R_W C_W}{2.3 R_0 C_0}} \quad \text{(I-23)}$$

With C_W and R_W the wire's capacitance and resistance. C_0 and R_0 the input capacitance and output resistance of the unity buffer. The unity buffer is usually scaled for better delay performances. When the unity buffer is scaled n times, i.e., the buffer's transistors width to length ratio is increased n times; its output resistance can be approximated to $\frac{R_0}{n}$, and their input capacitance as $C_0 n$.

We notice from equation (I-23) that if the interconnect delay grows large compared to the gate delay, then more buffers should be inserted since the gate delay effect is not significant. The buffers are sized as not to have a significant delay, and their output resistance should be kept close to the resistance of each wire segment. If the output resistance of the buffer is much smaller than the wire section resistance, the

wire section delay will dominate over the gate delay [29]. Several repeater insertion techniques [34] were introduced to reduce timing [35], area, or the number of repeaters used [36].

IV.2.1.5 Crosstalk

Another factor affecting RC wires delay is crosstalk. Crosstalk describes the coupling of the signal from a conductor carrying a signal to its neighbor conductors. This signal appears as unwanted noise in the victim wire, and it can cause signal integrity problems or system failure in some instances.

In the standard terminology, the conductor carrying the signal and causing the crosstalk is called the aggressor, while the quiet net is known as the victim.

It is worth mentioning that crosstalk happens between connectors, packages, and antennas too [22], i.e., any victim conductor can pick up the aggressor signal. Still, we will only discuss crosstalk between wires in this case.

Crosstalk can happen indirectly through the shared return paths, or directly through mutual inductance L_m(magnetic field) and capacitance C_m (electrical field) as shown in Figure I.22.a.

Figure I.22 shows an aggressor line A terminated by its characteristic impedance Z_0 at the far end, and a victim line B terminated by its characteristic impedance at both ends. When a voltage step is injected into line A, the signal moves from its near-end to its far-end. Along the lines, parts of the signal couple into the victim line at different points, this noise signal moves to both ends of the victim line. Crosstalk measured at the near-end of the wire and is called near-end crosstalk, also known as NEXT. Far-end crosstalk, also known as FEXT, is measured at the far end of the wire.

Crosstalk affects the amplitude of the signal traveling along the line and its adjoining signals but also their timing. As shown in Figure I.22.b, the delay at the output of the driver is well defined; however, the delay through the wire depends on the Coupling capacitor C_m, which leads to uncertainty at the reception. The value of the capacitance and inductance depends on the signal and are considered data-dependent,i.e., it depends on the bits “0” or “1” traveling through their neighbor.

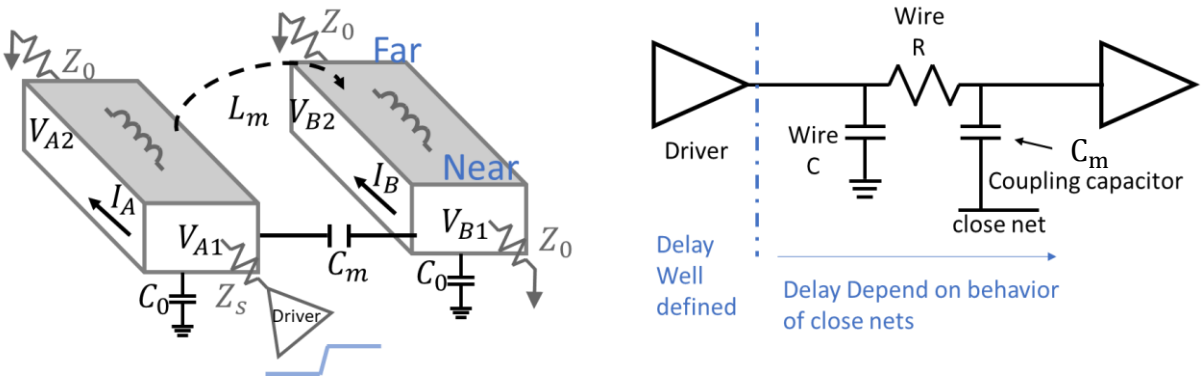


Figure I.22 : (a) Crosstalk mechanisms. (b) timing uncertainty due to crosstalk.

When the signal injected into line A is in the opposite direction of the line B signal, the propagation is called "odd-mode" propagation, with $I_A = -I_B$ and $V_{A2} = -V_{B2}$.

From Kirchhoff's voltage and current laws, we can express the voltage and currents V_{A1} and I_A as:

$$V_{A1} = L\frac{dI_A}{dt} + L_m\frac{dI_B}{dt} \quad \text{(I-24)}$$

$$I_A = C_0\frac{dV_{A2}}{dt} + C_m\frac{d(V_{A2} - V_{B2})}{dt} \quad \text{(I-25)}$$

Thus, for odd-mode propagation, the voltage and current are given by:

$$V_{A1,B1} = (L - L_m)\frac{dI_{A,B}}{dt} \quad \text{(I-26)}$$

$$I_{A,B} = (C_0 + 2C_m)\frac{dV_{A2,B2}}{dt} \quad \text{(I-27)}$$

Where $L - L_m$ is the effective odd-mode inductance L_{odd}, while $C_0 + 2C_m$ is the effective odd-mode inductance C_{odd}.

When the signals injected into the lines are propagating in the same direction, the propagation is called even-mode propagation with $I_A = I_B$ and $V_{A2} = V_{B2}$. The voltage and current are then expressed by:

$$V_{A1,B1} = (L + L_m)\frac{dI_{A,B}}{dt} \quad \text{(I-28)}$$

$$I_{A,B} = C_0\frac{dV_{A2,B2}}{dt} \quad \text{(I-29)}$$

With $L + L_m$ is the even-mode effective inductance L_{even}, and C_0 its even mode capacitance C_{even}.

As mentioned early, wires delay highly depend on the capacitor C. Thus any variation in the C value translates to a variation in the propagation delay and thus leads to timing noise. We notice that $C_{even} < C_{odd}$ which means that the capacitive crosstalk effect is more notable when the wires carry different bits.

Capacitive coupling should be reduced to reduce crosstalk; one technique commonly used is to shield the sensitive signals by inserting grounded wires below or around them, as shown in Figure I.23; these wires ground the floating coupling capacitances. Another technique consists of using differential signaling, i.e., take advantage of the low coupling capacitance in odd-mode.

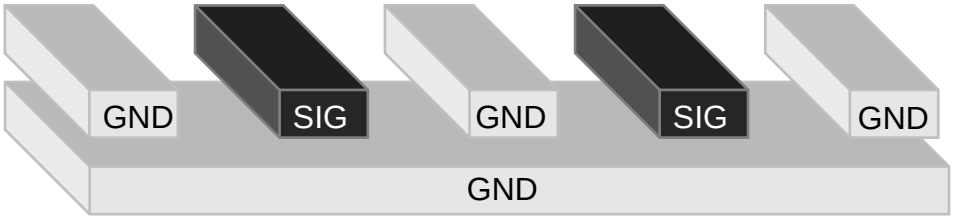


Figure I.23 : shielding of sensitive signals.

IV.2.1.6 Differential signaling

In single-ended signaling, a transmission line carries the signal with the ground as its reference usually, i.e. we measure the voltage difference between the transmission line and its return path.

While in differential signaling [22], two transmission lines carry two complementary signals. The two transmission lines are driven 180° out of phase, and the signal is extracted from the voltage difference between the two lines. Figure I.24 shows an example of two lines driven by complementary signals in differential mode. The signals are called balanced signals when the two signals have equal amplitude but opposite polarity relative to a reference voltage, also known as the common-mode voltage. Ideally, no current flows through the ground connection when balanced signals are used, because the return currents of balanced signals cancel each other. Thus, the transmitter and receiver do not require a shared ground connection. But an extra line for ground-connection used in some cases such as USB and CAN, where DC-Coupling is necessary to maintain the common-mode voltage at a certain level.

Differential signaling is beneficial for canceling common-mode noise i.e. noise present in both transmission lines. For example, it is very effective against electromagnetic interference (EMI) and crosstalk, since it tends to affect both lines identically if the lines are very close to each other. Thus, when looking at the difference between the lines, EMI noise is significantly reduced.

Furthermore, when lines are close to each other, outgoing EMI and crosstalk are greatly reduced, it is because the fields are confined between the lines and are less likely to fringe out to other transmission lines.

Moreover, differential signaling enables lower power consumption and lower voltage operations than the single-ended transmission while keeping an adequate signal-to-noise ratio. The differential signaling has a higher SNR compared to the single-ended one because the differential voltage range is double the voltage range of each line separately.

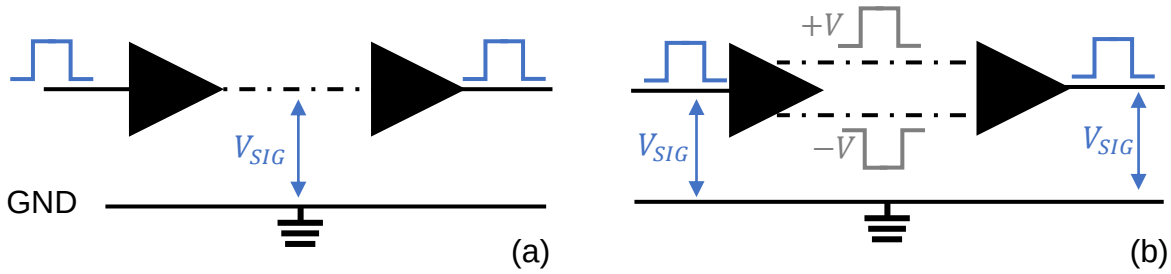


Figure I.24 : (a) single-ended signaling. (b) Differential signaling.

IV.2.2. Equalization techniques

As mentioned previously, ISI is a result of the channel dispersion i.e. unequal attenuation and delay of the different frequency components of the signal, which might be due to various factors such as frequency-dependent skin effect or dielectric absorption .... Figure I.25 shows an example of a single pulse sent over a channel. The pulse spreads over multiple symbol periods causing ISI. In Figure I.25, a_{-1} represent

the first pre-cursor while a_1 and a_2 represent the post-cursors, a_0 is called the primary cursor.

Typically, the signal suffers from the low-pass nature of the transmission channel, where the channel attenuates the high-frequency components more than the low-frequency ones, which leads to a degradation of the signal, especially the rise/fall time information contained in the high-frequency components [9], [22].

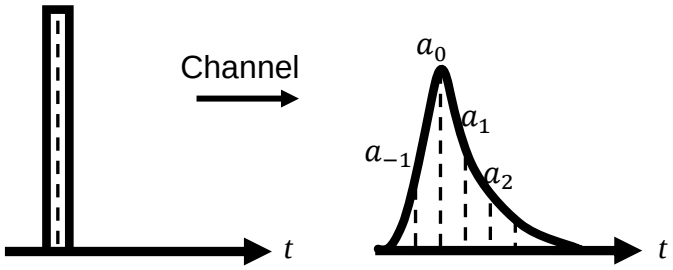


Figure I.25 : Dispersion causes a pulse to spread out over time.

Thus, for better signal integrity, all frequency components should be attenuated equally i.e. attenuate the low-frequency components using a high-pass filter at the receiver or compensate for the attenuation in high frequencies components by boosting them. This process of frequency enhancing used to eliminate dispersion is called equalization [37].

The latter approach extends the effective bandwidth of the channel [38], Figure I.26 shows an example of a low pass channel along with a high-pass frequency response receiver; their combination extends the bandwidth.

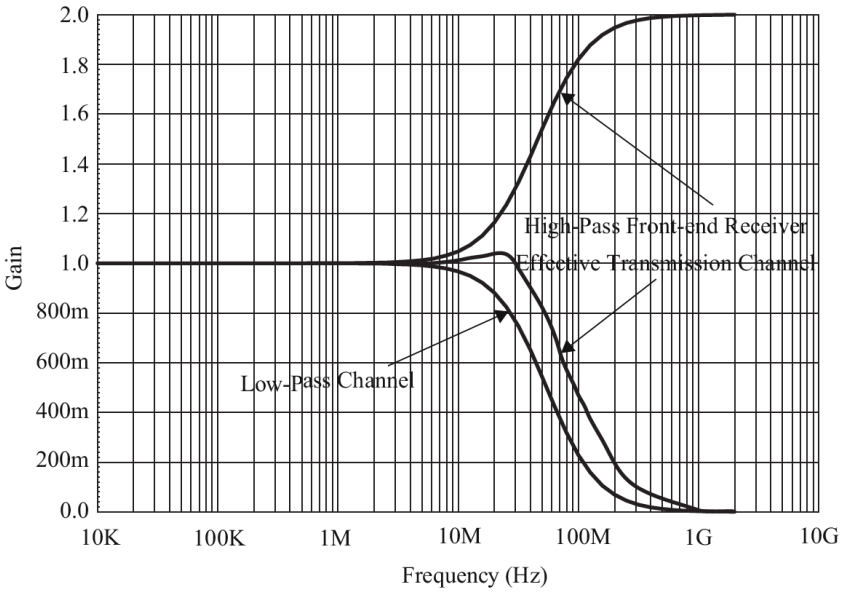


Figure I.26: Enhancing the high frequency components effect on the effective bandwidth [38].

Fixed equalizers [9] are used for pre-identified and fixed channels. In this case, the equalizer is optimized for a specific channel, if used for a different one, it will generate ISI. However, this knowledge is not always available as its the case for wireless transmissions... In these cases, fixed equalizers cannot be used. Instead, adaptive equalizers [20],[21] are used, their frequency response adapts automatically

to the transmission channel. Nowadays, different equalizer architectures are implemented in the digital or analog domain. They can be used at the transmitter or receiver side, and more often, they are combined for enhanced performances.

IV.2.2.1 Continuous-time linear equalizers CTLE

Continuous-time linear equalizers (CTLEs) [37] are analog equalizers that may be passive (use of capacitors, resistors, and inductors) or active based mainly on amplifiers. Their implementation does not require digital devices.

By boosting the high-frequencies, they sharpen the rising and falling edges of the signal. They can cancel both pre-cursor and post-cursor taps.

IV.2.2.1.a. Passive CTLEs

This type attenuates the low-frequency components to achieve a larger effective bandwidth, Figure I.27 shows an example of RC implementation [37], where the resistor attenuates the low-frequency components while the high-frequency components pass through the capacitor. The Laplace transfer function is given by:

$$H(s) = \frac{R_2}{R_2 + R_1} \frac{1 + R_1 C_1 s}{1 + \frac{R_1 R_2}{R_2 + R_1}(C_2 + C_1)s} \tag{I-30}$$

with its zero and pole frequency are given by:

$$\omega_Z = \frac{1}{R_1 C_1} \tag{I-31}$$

$$\omega_p = \frac{1}{\frac{R_1 R_2}{R_2 + R_1}(C_2 + C_1)} \tag{I-32}$$

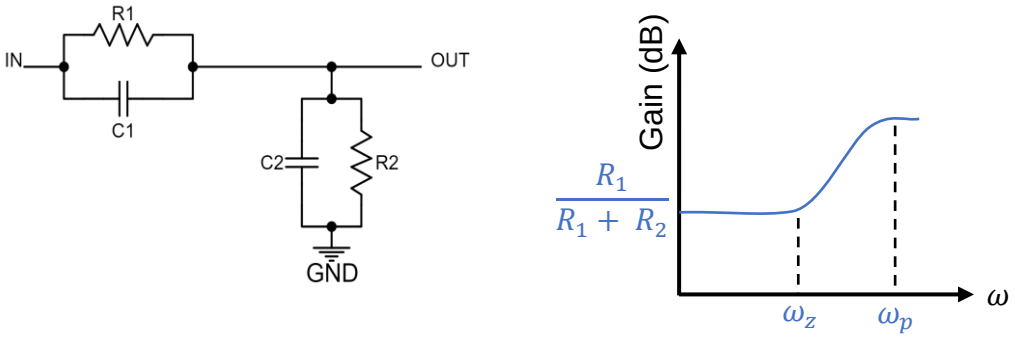


Figure I.27 : Passive implementation of a CTLE and its frequency response

The peaking frequency is defined by the zero and pole frequencies, as shown in Figure I.27, while the boosting factor is approximately the ratio of high frequency gain $HF\_Gain = \frac{C_1}{C_2 + C_1}$ and the DC Gain $DC\_Gain = \frac{R_2}{R_2 + R_1}$. These formulas make the design of passive CTLE a straightforward process. However, The *RC* network introduces impedance discontinuity at the interface of the equalizer and the channel which might lead to reflection [41].

IV.2.2.1.b. Active CTLE

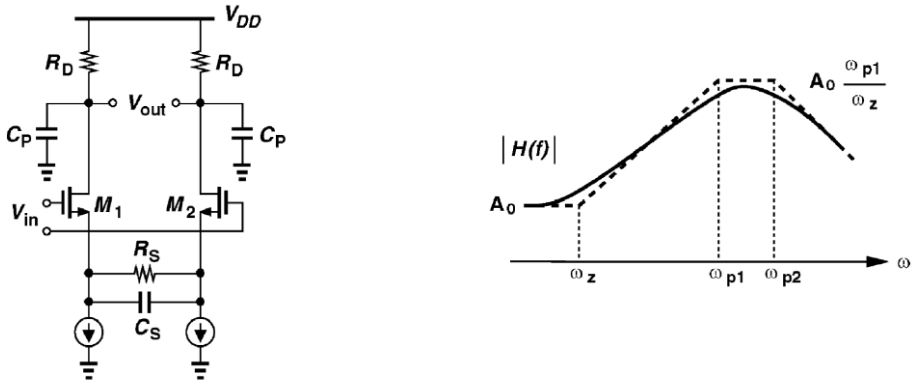


Figure I.28 : Implementation of an active CTLE and its frequency response [42].

Amplifiers are used to provide some gain, Figure I.28 shows an example of one of the most common CTLE implementation [42]. The circuits consists of a differential pair degenerated with a pair of resistor and capacitor. The pair of resistor and capacitor inserts a zero ω_Z at low-frequencies, leading to high-pass frequency response. The frequency response of the circuits along with its zero and poles are given by :

$$H(s) = \frac{g_m}{C_p} \frac{s + \frac{1}{R_s C_s}}{\left(s + \frac{1 + g_m R_s/2}{R_s C_s}\right)\left(s + \frac{1}{R_D C_p}\right)} \tag{I-33}$$

$$\omega_z = \frac{1}{R_s C_s} \tag{I-34}$$

$$\omega_{p1} = \frac{1 + g_m R_s/2}{R_s C_s} \tag{I-35}$$

$$\omega_{p2} = \frac{1}{R_D C_p} \tag{I-36}$$

The DC-gain is given by :

$$DC\_Gain = \frac{g_m R_D}{1 + g_m R_s/2} \tag{I-37}$$

The RC degeneration pair is often tunable to adjust the zero frequency, the first pole and the DC Gain, it can also be used to compensate for the process, voltage and temperature (PVT) variations of the circuit. Generally, the zero and the first pole are placed closely for good CTLE performance. The zero provides the high-frequency boost that allows for faster rise/fall times which consequently opens the eye diagram. Meanwhile, the poles limit the bandwidth in order not to over-equalize by boosting high frequencies more than required, and thus not degrade the noise performance. This CTLE is usually limited to channels with a behavior similar to a 1st order Low Pass Filter (LPF) i.e. -20 dB/dec attenuation at high frequency. The circuit in Figure I.28 can achieve a signal gain over 0dB at the cost of power dissipation.

IV.2.2.2 Feed forward equalizer FFE

Unlike CTLE where the processing uses analog components, Feed forward equalizer (FFE) usually use digital finite impulse response (FIR) in discrete-time. FFE [38] can be implemented either in the transmitter or the receiver, but it is more often used in the transmitter.

In the time domain, FFE [21][24] pre-distorts the signal by combining delayed versions of the data signal, where each delayed version is multiplied by a different weight coefficient. Choosing the appropriate weight coefficient for each version eliminates the *ISI*. Figure I.29 shows an example of FFE structure, the total number of coefficients is called the number of taps e.g. an FFE with 3 coefficients is referred to as a 3-tap FFE. Furthermore, FFE can eliminate both pre-cursor and post-cursor *ISI*.

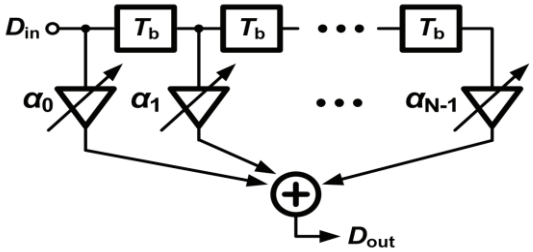


Figure I.29 : A conventional FFE structure.

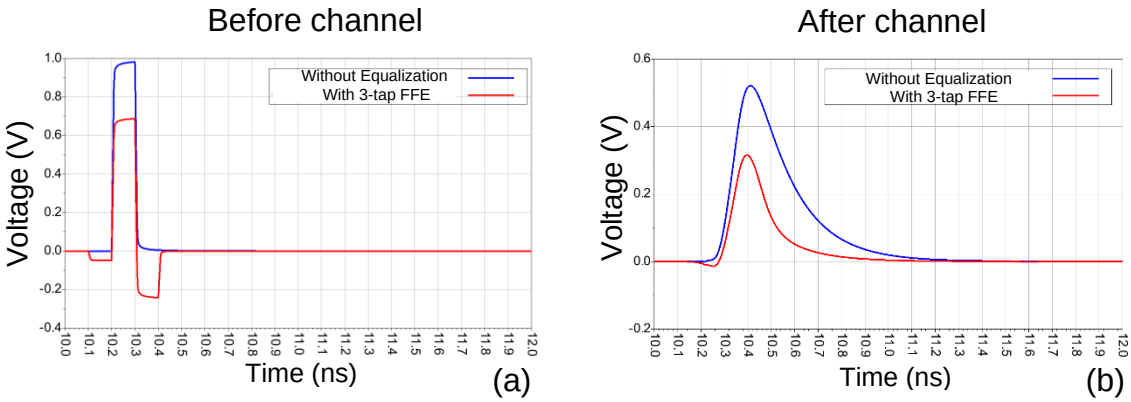


Figure I.30 : (a) Transmitted data with and without FFE equalization. (b) Received Data with and without FFE Equalization.

Figure I.30 shows the pulse response of a channel without FFE; we notice the resulting pre-cursor and post-cursor. Figure I.30 shows the pulse after passing through a 3-tap FFE (pre and post cursor taps are used), by delaying the pulse and multiplying it by the appropriate coefficients, two negative pulses are created before and after the main pulse. As the positive main pulse spreads when traveling in the channel, the negative surrounding pulses cancel out this spreading. As a result, the signal after the channel has a much reduced *ISI*. However, we notice that the amplitude of the main pulse before the channel is reduced compared to the non-equalized one. One of FFE drawbacks is that it reduces the signal amplitude; this is due to its linearity property, where the sum of all tap values must be equal to 1.

Figure I.31 shows the high-pass filter behavior of a 3-tap FFE. We notice attenuation and amplification at different harmonics of the Nyquist frequency due to the nature of FIR filters.

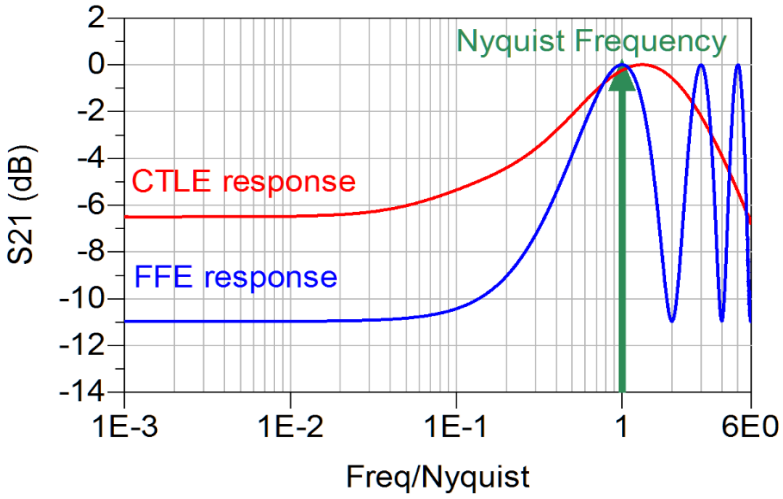


Figure I.31 : Comparison between CTLE and FFE spectrum, FFE amplifies odd harmonics of the Nyquist frequency [44].

IV.2.2.3 Decision Feedback equalizer DFE

As mentioned earlier, ISI is due to the spreading of one bit over more than its symbol period. The DFE [9], [41] removes ISI by undoing the spread at the receiver part. Knowing the preceding bits, the DFE can estimate their contribution to the currently received bit, then subtracts it. Figure I.32 shows the DFE architecture, and that it subtracts a weighted version of the previous decisions. DFE feedback path can be implemented using D flip-flops for the delay, and variable gain amplifiers controlled by high-resolution DACs. The feedback signals are summed then sampled by a decision circuit.

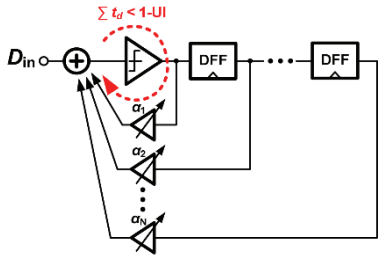


Figure I.32 : Conventional DFE architecture.

This is based on the assumption that all the previous decisions are correct. When errors are present, the error-propagation in the loop will worsen the ISI, which is one of the drawbacks of the DFE. Moreover, because the DFE requires the knowledge of the previously received bits, it can only remove post-cursor ISI and is usually combined with CTLE or FFE at the transmitter to remove both pre and post-cursor. DFE suffers from a harsh timing issue since it needs to finish its calculation and subtraction in less than 1UI, especially the feedback loop of the first tap where the sum of the ck-to-q delay and setup time of the slicer and the feedback path delay needs to be less than 1UI. This timing issue limits the data rate that DFE can handle, but

techniques such as loop unrolling [45] have been investigated to alleviate this problem. In 65nm CMOS process, the data rate of a DFE stage exceeds 21Gb/s [46], [47] But the number of taps increases in higher data rates devices.

IV.2.2.4 Adaptive equalization

In practice, removing ISI requires a certain flexibility in the equalization process. For example, PVT variability can affect the required taps weights from one implementation to another. Equalizers that can adapt their weight taps to the channel characteristics are called adaptive equalizers. This principle was introduced by Lucky [48] in 1964 to improve data transmission from 1200b/s to 9600b/s over telephone lines. However, this flexibility complicates the system design significantly by adding control circuitry, mainly power-hungry ADCs. Different adaptive equalization algorithms can be used to reduce the PVT variabilities and the channel real-time variations effect on the signal. Different adaptive equalization algorithms can be used, such as Least-mean square (LMS) [40], [49], or Zero Forcing (ZF)[50], to optimize the weights of the taps periodically or continuously when the channel properties change.

An adaptive equalizer has two operating modes, training mode and decision-directed mode. In the training mode, a well-known bit sequence is transmitted; this training sequence is used by the equalizer to adapt its weights even in the worst possible case.

For example, LMS algorithm tries to reduce the well-known mean-squared error function $e_n^2 = (y_n - x_n)^2$, where y_n is the resulting equalized output at each iteration, and x_n is the transmitted training sequence. The equalizer starts with small weights then increases at each iteration to converge to the correct weights with minimum error function.

IV.3. Transmission lines

Figure I.33 shows the most common transmission line structures: microstrip line (MSL), differential lines also known as coplanar strips (CPS), and coplanar waveguide (CPW). Each has its advantages and drawbacks. For example, the microstrip structure is simple and does not require a large area overhead, while CPS requires more surface but is more robust against crosstalk.

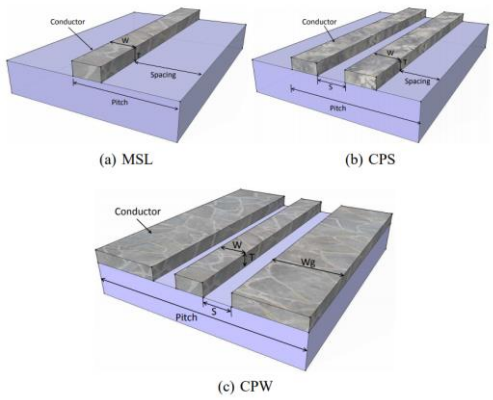


Figure I.33 : Structure of the main three types of the transmissions lines: (a) microstrip line (MSL), (b) differential line or coplanar strips (CPS), and (c) coplanar waveguide (CPW) [13].

IV.3.1. Interconnect parameters

IV.3.1.1 Characteristic impedance

The characteristic impedance Z_c of a transmission line is the relation between voltage and current at any point along the line. For a uniform line, we will see the same characteristic impedance as we propagate down the line anywhere we choose to look. Therefore, Z_c is the same, looking into an infinitesimal section of the transmission line, and can be described as :

$$Z_c = \sqrt{\frac{r + j\omega l}{g + j\omega c}} \tag{I-38}$$

We notice that for higher frequencies, the term l/c is more dominant, and the characteristic impedance approaches $\sqrt{l/c}$.

For a lossless transmission line ($r = g = 0$), the characteristic impedance Z_c becomes $Z_c = \sqrt{l/c}$.

IV.3.1.2 Propagation constant

Let us consider the equations (I-7) and (I-8), for a sinusoidal steady-state condition, the equations become :

$$\frac{\partial V(z)}{\partial z} = -(r + j\omega l)I(z) \tag{I-39}$$

$$\frac{\partial I(z)}{\partial z} = -(g + j\omega c)V(z) \tag{I-40}$$

Solving these equations gives the wave equations $V(z)$ and $I(z)$:

$$\frac{\partial^2 V(z)}{\partial z^2} - \gamma^2 V(z) = 0 \tag{I-41}$$

$$\frac{\partial^2 I(z)}{\partial z^2} - \gamma^2 I(z) = 0 \tag{I-42}$$

where :

$$\gamma = \alpha + j\beta = \sqrt{(r + j\omega l)(g + j\omega c)} \tag{I-43}$$

The complex propagation constant γ relates $V(z, \omega)$ of a traveling wave at any point of the transmission line to its initial point.

The real part α in equation (I-43) is the attenuation constant; it represents the wave's attenuation as it travels the transmission line. The losses are due to non-idealities of the medium such as the limited conductivity of the transmission line, or the dielectrics non-perfect insulation. The greater α is, the higher the losses along the line. When $\alpha = 0$, the line is said lossless.

The imaginary part β in equation (I-43) is known as the phase constant; it represents the phase shift of the voltage at a point z along the transmission line with

respect to the initial voltage. In other words, it represents the rate at which the electromagnetic wave travels the medium.

For a low-loss line with r and g small, the propagation constant can be reduced into [23] :

$$\gamma \approx \frac{1}{2}\left(r\sqrt{\frac{c}{l}} + g\sqrt{\frac{l}{c}}\right) + j\omega\sqrt{lc} = \frac{1}{2}\left(\frac{r}{Z_c} + gZ_c\right) + j\omega\sqrt{lc} \quad \text{(I-44)}$$

with $\alpha_c = r/2Z_c$ and $\alpha_d = gZ_c/2$ the attenuation factor due to resistive and dielectric losses, respectively. For on-chip interconnects, we can assume that the leakage conductance is equal to 0. Thus, the dielectric losses are neglected, and the frequency-dependent resistive losses are dominant.

In this case, the phase constant is reduced to $\beta = j\omega\sqrt{lc}$.

IV.3.1.3 Lossless transmission lines delay

Furthermore, let us assume that the resistance r is negligible. In this case, a sinuisodal wave will propagate without attenuation or distortion. Furhermore, the signal travels like a wave with an effective phase velocity of :

$$v = \frac{1}{\sqrt{lc}} = \frac{c_0}{\sqrt{\varepsilon_r \mu_r}} \quad \text{(I-45)}$$

c_0 is the velocity of light in vacuum, ε_r is the relative dielectric constant, μ_r is the relative permeability. In SiO_2, the maximum effective propagation velocity is $0{,}5.\, c_0 = 1{,}5.\, 10^8 \frac{m}{s}$, because ε_r in SiO_2 is around 3.8, this velocity decreases when the losses are included.

From equation (I-46) we see that the delay of a D length transmission line can be expressed as :

$$\tau_d = \frac{D}{v} = \frac{D\sqrt{\varepsilon_r \mu_r}}{c_0} \quad \text{(I-46)}$$

We notice that the delay is a linear function of the length; the delay is in the region of 7 ps/mm for the SiO_2 case.

IV.3.1.4 Termination

Waves propagating along a transmission line suffer from reflection when the impedance changes, these locations are known as impedance discontinuities. In other words, part of the signal does not arrive at its destination (the load), but instead, it travels back to its source due to impedance mismatch. Reflections happen whenever the instantaneous impedance changes abruptly; for instance, vias, connectors, or packages.

The ratio of the reflected signal to the transmitted signal is known as the reflection coefficient :

$$\Gamma = \frac{Z_L - Z_c}{Z_L + Z_c} \quad \text{(I-47)}$$

We notice from equation (I-47), that the greater the impedance difference, the higher the amount of reflected signal, and consequently, more significant distortion, attenuation, and in general lower signal integrity.

For example, for $Z_L = 25\ \Omega$ and $Z_c = 50\ \Omega$,one third of the incident wave will be reflected at the load. The reflected wave will have a phase shift of $180°$ with respect to the incident wave.

When the transmission line is shorted at the far-end ($Z_L = 0$, $\Gamma \to -1$), all of the energy will be reflected but with a $180°$ phase shift. Consequently, the signals cancel each other at the far-end of the line.

On the other hand, if the load is an open circuit with $Z_L = \infty$, $\Gamma \to 1$, all the energy is reflected to the source. In other words, the forward incident wave and the reflected wave will have the same phase and amplitude.

IV.3.2. Distortionless (Low dispersion) transmission line

The fact that phase velocity and attenuation constant are frequency-dependent in a lossy line means that the line is dispersive. In other words, the different frequency components of a pulse will travel at different speeds, which leads to the signal spreading in time, also known as dispersion.

Ideally, the attenuation constant α and the phase velocity v should be frequency-independent for a pulse or a modulated signal to travel without distortion, which is not the case for low loss lines. Achieving a frequency-independent phase velocity v is possible when $\beta = a\omega$ is a linear function of frequency. In this case, the phase velocity becomes $v = \omega/\beta = 1/a$.

A lossy transmission line can be designed to be distortionless by choosing the parameters r, l, c, g rightly, to do so we choose: $r/l = c/g$.

Hence, the propagation constant becomes :

$$\gamma = \alpha + j\beta = r\sqrt{\frac{c}{l}} + j\omega\sqrt{lc} \qquad \text{(I-48)}$$

We notice that the attenuation constant α is frequency-independent when r is frequency-independent, that is to say, the signal components are attenuated equally. Furthremore, $\beta = \omega\sqrt{lc}$ is a linear function of frequency, accordingly, the phase velocity for this particular case is given by $\beta = 1/\sqrt{lc}$. Hence, this transmission line will only attenuate the signal but do not distort it.

V. RF-Interconnect state of the art

To summarize, depending on which term dominates in equation (I-10). we can distinguish two types of transmission.

At low frequencies and up to a few GHz, the rc is much higher than the inductance equivalent impedance $l\omega$. In this rc region, the values of r and c vary rapidly, which leads to high dispersion resulting in distortion of the signal.

At higher frequencies and depending on the transmission line type, the lc term dominates [51]. In this case, the linear inductance and capacitance of the transmission line control the propagation of the signal. It is beneficial to have a constant group delay

across the transmission bandwidth to avoid distortion. The group delay, defined as the time delay of the signal as a function of frequency by:

$$\tau_g(\omega) = -\frac{d\varphi}{d\omega} \tag{I-49}$$

where φ is the phase delay, and ω the angular frequency.

Ideally, the group delay would be constant for a distortionless transmission line. However, in practice, the signal frequency components take different amounts of time to pass through the transmission line.

This latter case is valid for lc region transmission. The group delay varies very little at high frequencies, as shown in Figure I.34 for a microstrip, for example. Thus, it is evident that at these frequencies, the distortion and ISI are reduced significantly.

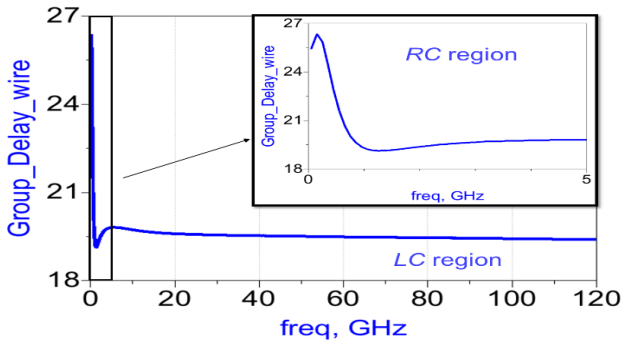


Figure I.34 : Group delay for a wire.

Furthermore, the propagation velocity in the lc region of the transmission line is much higher than the rc region. Hence, the delay is much lower in the lc region of the microstrip. For example, a 20mm bus is crossed in 800ps using the rc region of the wire [8], while in lc region, it can be crossed in 140ps only.

LC transmission offers several advantages compared to the traditional RC-limited interconnects, including low and linear latency. In other words, the delay in LC region is a linear function of the distance, while in RC region, it is a quadratic function of the distance. Another advantage is that LC transmission does not require repeaters and thus have fewer constraints on the floorplan.

This approach was introduced by the authors of [51] in 2003, where they exploited the LC region of the transmission line to transmit a 1Gb/s data signal using a modulated 7.5GHz carrier over a 20mm differential transmission lines. They achieved a maximum velocity propagation of half the speed of light. They also stated that the crossover frequency between RC and LC regimes could be reduced to the single-gigahertz range by emphasizing the parasitic inductance and reducing the transmission line's resistance. Therefore, they intentionally used explicit return ground to predict and control the parasitic inductance of the line.

The RF modulation can be combined with a baseband modulation, as suggested in [52] for off-chip applications. Where they proposed to extend the baseband data rate by transmitting in RF-modulated frequency bands, in other words, divide the channel into several bands with sufficient bandwidth, as is the case for Frequency-division multiple access (FDMA). In theory, this approach should reduce the interference between the transmission lines. However, this approach also requires

generating several carriers. They achieved a maximum data rate of 4Gb/s by combining a 2Gb/s baseband transmission and a 2Gb/s RF transmission over a 10cm microstrip on a FR4 PCB board.

The previous approach was applied for on-chip applications in [53], where they achieved a total data rate of 10Gb/s. They used common mode to transmit a 2Gb/s in baseband, and 4Gb/s per RF channel (28.8GHz and 49.5GHz) across a 5mm on-chip differential transmission lines with a latency of 6ps/mm for both RF bands. The transmission lines are shown in Figure I.35.b. The transmission lines are 1um thick, 3um wide with a 3um spacing. They used a simple ASK modulation for Base Band (BB) and RF modulation. For each RF band in the transmitter, the carriers (28.8GHz and 49.5GHz) generated by the VCO are inductively coupled to the ASK modulator using a transformer, as shown in Figure I.35.a. Next, the ASK modulated signal is once again inductively coupled to the transmission lines using a transformer. The second transformer reduces Intercarrier Interference (ICI) caused by signal leaking into adjacent RF bands.

At the receiver side, the signal goes through a transformer that acts as a bandpass filter; each RF demodulator comprises a self-mixer and a baseband amplifier.

They used a 90nm technology from IBM and achieved a BER lower than 10^{-9}, and an energy efficiency of 0.45pJ/bit and 0.625pJ/bit, for RF & baseband, respectively (VCOs' power consumption is not included).

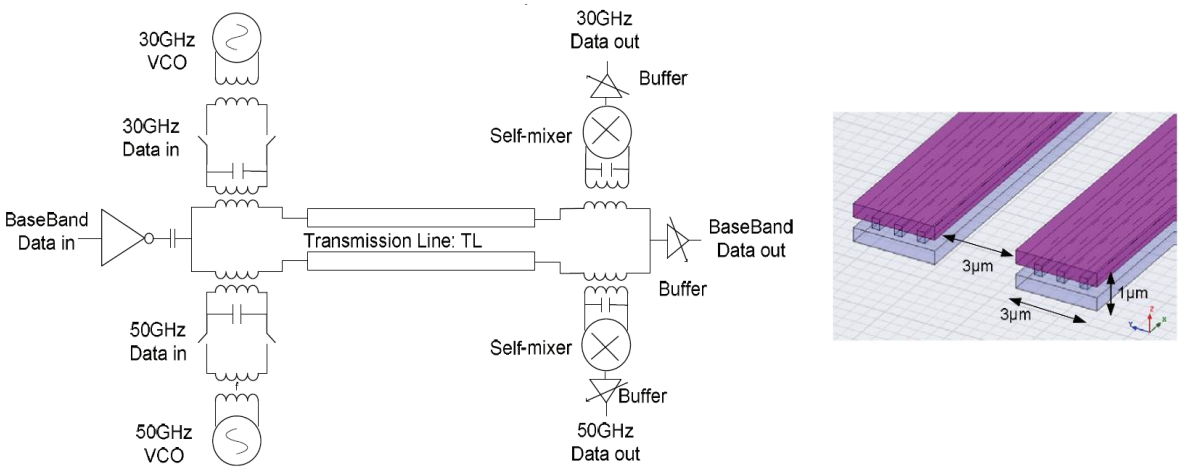


Figure I.35 : (a) Schematic of the tri-band RF-interconnect [53]. (b) On-chip differential TL [53].

The authors of [54] realized a simultaneous bidirectional link to exchange data in both directions simultaneously between two transceivers. They used a dual-band approach by combining a 4.4Gb/s RF band transmission and a 10Gb/s baseband modulation to reach a total data rate of 14.4Gb/s. Bidirectional frequency division consists of separating the transmit and received data in the frequency domain, in its simplest form consists of dividing the physical channel into a low-pass and high-pass band. The main challenge is to isolate the incoming and outcoming signal bands properly. Furthermore, using several bands increases the complexity significantly. The authors used a 4-level pulse-amplitude modulation (PAM4) for baseband communication to increase data throughput.

Figure I.36 describes the system architecture, the left side of the chip is comprised of a BB receiver named multi-level receiver (MLRX), and an RF transceiver. The RF transceiver is composed of a phase-locked loop (PLL), an ASK modulator, and an on-chip transformer to split the BB and RF bands. The authors used a VCO to generate an output frequency for a frequency range between 0.6GHz~2.4GHz, then used an injection-locked frequency multiplier (ILFM) to multiply the VCO frequency by 10, thus generating a carrier at 20GHz. The ASK modulator modulates the 20GHz carrier with the 4.4Gb/s data signal. Next, the RF signal is inductively coupled into the single-ended transmission line using a transformer to minimize inter-channel interference. On the same side, the MLRX is composed of voltage amplifiers, three comparators to demodulate the PAM4 signal, sampling DFF and a decoder to recover the data stream.

The chip's right side is composed of an RF receiver (RFRX), and a baseband transmitter, also known as a multi-level transmitter (MLTX). Firstly, the signal is split into BB and RF signals using a transformer. The received RF signal is converted into a differential signal, then a mutual-mixer down-converts the signal into baseband, the same approach was employed in [55] for off-chip links. The MLTX employs a 4-bit encoder to encode two data streams, a DFF to synchronize the signal, and a voltage mode driver to generate the 4-level voltage signal.

The chip was designed using a 65nm CMOS process and achieves a BER lower than 10^{-15} for a 5mm single-ended transmission. The BB and RF band's energy efficiency is 1.2pJ/b and 1.04pJ/b, respectively (the 12.4mW PLL power consumption is not included).

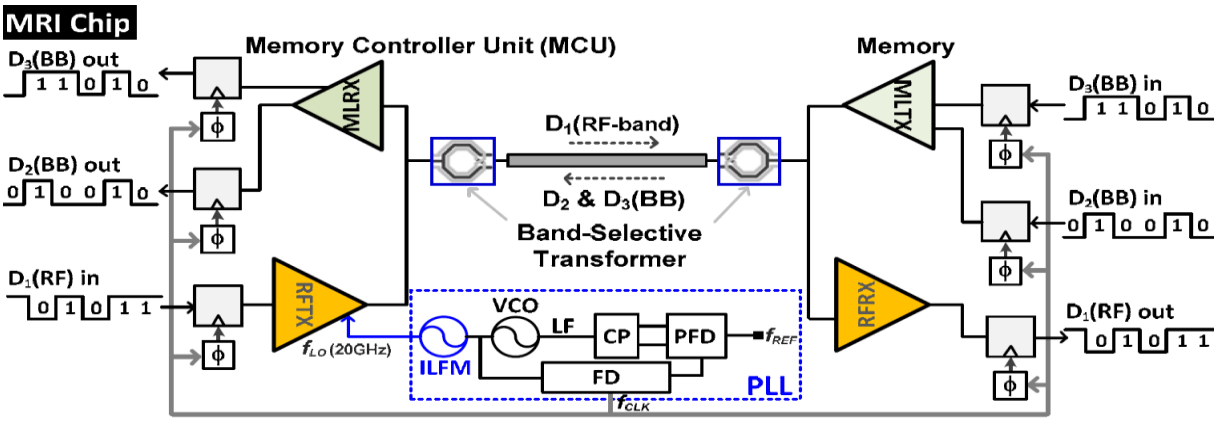


Figure I.36 : Architecture of the proposed system [54].

The authors in [56] realized a 5Gb/s bidirectional RF interconnect. This link employs RF transmission only, and thus provides a low-latency communication, estimated at 9ps/mm, contrarily to the dual-band links (BB + RF) that offer low-latency for the RF transmission only.

Contrarily to the previous works where the links were point-to-point, this work proposes a link with multi-drop capability, where the channel connects four nodes, each node includes a transmitter and receiver. Multi-drop topology allows each node of the topology to communicate either with a specific node or all nodes. The latter is also known as broadcasting. The four nodes share the same communication channel with fairness and without collision using an arbitration ability.

Figure I.37 shows the four nodes of the multi-drop RF-interconnect, node A multi-casts to the other nodes. Signal flow is highlighted in green. The inactive nodes, i.e., node A receiver and transmitter of nodes B, C, and D, are turned off to save power.

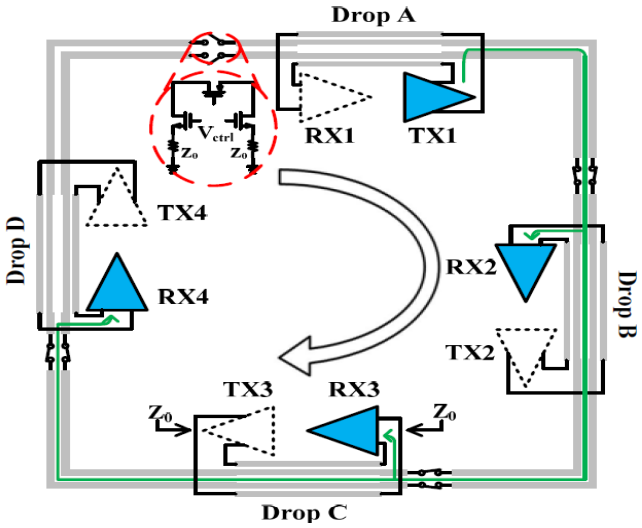


Figure I.37 : 4-drop RF interconnect (Drop A multicasts) [56].

Each node includes an ASK transceiver able to both transmit and receive, and a $\lambda/4$ directional coupler. The coupler helps to suppress reflections from multiple drops along the line with its isolation and matching capability. Figure I.37 shows digitally controlled switches used between adjacent nodes. The switches can terminate the signal with a proper impedance $Z_0(100\Omega$ differentially), pass it to the following nodes, or divide the physical link into two separate sections (node A communicates with node B while node C communicates with node D). This flexibility is a result of the switches high isolation.

The couplers were implemented on top of the differential transmission channel with no extra silicon area, as shown in Figure I.38. A 5.5mm differential transmission line is used to connect the farthest points in the link, while the distance between adjacent nodes is 1.5mm. The attenuation at 60GHz is estimated at 1.2 dB/mm, thus the maximum attenuation, between the farthest points, is 6.6dB/mm.

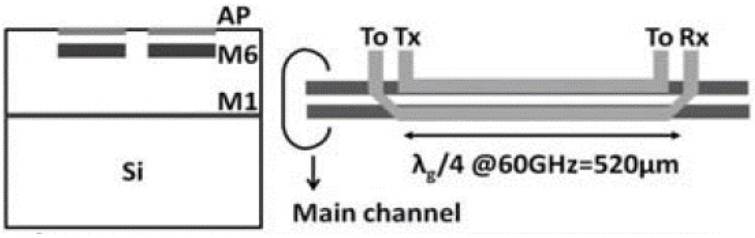


Figure I.38 : On-chip directional coupler [56].

The transmitter includes a free-running cross-coupled differential pair with a resonant tank 60GHz VCO and an ASK modulator that modulates the 60GHz carrier. The modulated data is then inductively coupled into the directional coupler using a transformer.

The receiver comprises a differential low noise amplifier (LNA) with inductive degeneration. The LNA has a bandwidth of 9GHz, a noise figure of 6dB, and a gain of

18dB at 60GHz. After the LNA, the signal is fed into a mutual differential mixer operating in sub-threshold. The mixer acts as an enveloppe detector with a simulated bandwidth of 6Gb/s at 60GHz, and an input sensitivity of $10mV_{pk-pk}$. The receiver consumes 5mW.

The chip was realized in a TSMC 65nm CMOS process, the core area of each transmitter and receiver was estimated at $0.0048mm^2$ and $0.034mm^2$, respectively. A BER lower than 10^{-12} was achieved for all drops in different configurations ,e.g., point-to-point and broadcast communications. The energy efficiency for a 5.5mm link is estimated at 1.32pJ/bit (VCO power consumption not included).

Figure I.39 shows the measured BER versus data rate at different drops when node A is broadcasting to nodes B, C, and D.

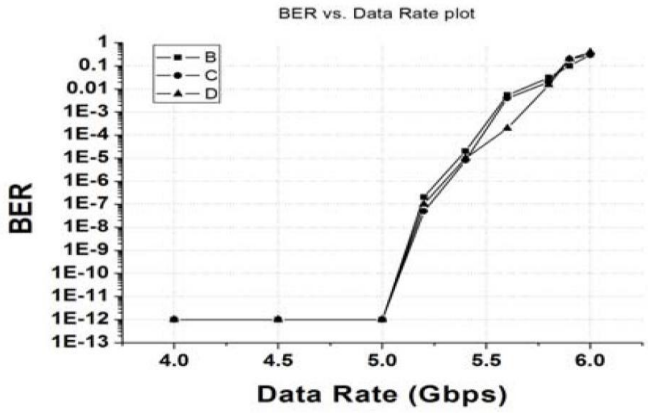


Figure I.39 : Measured BER vs data rate at different drops [56].

We notice that the BER degrades significantly for higher data rates, according to the authors, this is due to the circuitry speed limitations.

VI. Conclusion

This chapter discussed the standard RC wire solution and its bottlenecks; it is clear that latency is one of these bottlenecks; since it grows quadratically with the length of the wire. Using repeaters reduces the delay, but it results in more floorplan constraints. Furthermore, the distortion due to the limited RC available bandwidth requires considerable efforts to keep the signal integrity and have a reliable transmission at high data rates.

We also introduced some of the emerging on-chip interconnects such as wireless, optical, and RF-Interconnects. They all offer certain advantages compared to standard RC solutions, such as larger bandwidth and lower latency. However, in this work, we will focus on the prospective RF-interconnects since 1) It is a CMOS compatible solution and can be used in digital circuits. 2) it has a large bandwidth available (i.e., high data rates) without any equalization. 3) It can cover long distances without floorplan constraints. 4) can be used to multicast to different nodes.

The next chapter will introduce the different common modulations and less common duobinary modulation used in this work; we will also discuss the system's architecture and its main components.

VII. References

[1] J. Warnock, “Circuit design challenges at the 14nm technology node,” in *Proceedings of the 48th Design Automation Conference*, 2011, pp. 464–467.

[2] L. Benini and G. De Micheli, “Networks on chips: A new SoC paradigm,” *computer*, vol. 35, no. 1, pp. 70–78, 2002.

[3] M.-C. F. Chang *et al.*, “Power Reduction of CMP Communication Networks via RF-Interconnects,” 41st IEEE/ACM International Symposium on Microarchitecture, Lake Como, 2008, pp. 376-387.

[4] E. E. Nigussie, *Variation Tolerant On-Chip Interconnects*. Springer Science & Business Media, 2011.

[5] L. P. Carloni, P. Pande, and Y. Xie, “Networks-on-chip in emerging interconnect paradigms: Advantages and challenges,” in *2009 3rd ACM/IEEE International Symposium on Networks-on-Chip*, 2009, pp. 93–102.

[6] J. Bashir, E. Peter, and S. R. Sarangi, “A Survey of On-Chip Optical Interconnects,” *ACM Comput Surv*, vol. 51, no. 6, Jan. 2019.

[7] S. Deb, A. Ganguly, P. P. Pande, B. Belzer, and D. Heo, “Wireless NoC as interconnection backbone for multicore chips: Promises and challenges,” *IEEE J. Emerg. Sel. Top. Circuits Syst.*, vol. 2, no. 2, pp. 228–239, Jun. 2012.

[8] M. F. Chang *et al.*, “CMP network-on-chip overlaid with multi-band RF-interconnect,” in *2008 IEEE 14th International Symposium on High Performance Computer Architecture*, 2008, pp. 191–202.

[9] S. H. Hall and H. Heck, *H. Advanced Signal Integrity for High-Speed Digital Designs. 2009*. Wiley & Sons.

[10] K. Ohashi *et al.*, “On-chip optical interconnect,” *Proc. IEEE*, vol. 97, no. 7, pp. 1186–1198, 2009.

[11] P. Dong, Y.-K. Chen, T. Gu, L. L. Buhl, D. T. Neilson, and J. H. Sinsky, “Reconfigurable 100 Gb/s silicon photonic network-on-chip,” *IEEEOSA J. Opt. Commun. Netw.*, vol. 7, no. 1, pp. A37–A43, 2015.

[12] S. H. Gade, S. S. Rout, and S. Deb, “On-Chip Wireless Channel Propagation: Impact of Antenna Directionality and Placement on Channel Performance,” in *2018 Twelfth IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*, Oct. 2018, pp. 1–8, doi: 10.1109/NOCS.2018.8512173.

[13] A. Karkar, T. Mak, K.-F. Tong, and A. Yakovlev, “A survey of emerging interconnects for on-chip efficient multicast and broadcast in many-cores,” *IEEE Circuits Syst. Mag.*, vol. 16, no. 1, pp. 58–72, 2016.

[14] M. Hsieh and G. E. Sobelman, “Architectures for multi-gigabit wire-linked clock and data recovery,” *IEEE Circuits Syst. Mag.*, vol. 8, no. 4, pp. 45–57, 2008.

[15] B. Sklar and others, *Digital communications: fundamentals and applications*. 2001.

[16] “Enabling Precision EYE Pattern Analysis Extinction Ratio, Jitter, Mask Margin.” [Online]. Available: https://dl.cdn-anritsu.com/en-au/test-measurement/files/Technical-Notes/Technical-Note/MP2100A_EE1101.pdf.

[17] M. Müller, M. Russ, and M. Russ, “Total Jitter Measurement at Low Probability Levels, Using Optimized BERT Scan Method.” Agilent Technologies, 2005.

[18] R. W. Hamming, “Error detecting and error correcting codes,” *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, 1950.

[19] Anritsu Corporation, “Jitter Analysis : Basic Classification of Jitter Components using Sampling Scope.” [Online]. Available: https://dl.cdn-anritsu.com/en-au/test-measurement/files/Application-Notes/Application-Note/MP2100A_EF3100.pdf.

[20] Agilent Technologies, Inc., “Jitter analysis: The dual- Dirac model, RJ/DJ, and Q-scale.” [Online]. Available: https://www.keysight.com/upload/cmc_upload/All/dualdirac1.pdf.

[21] “Agilent 71501D Eye-Diagram Analysis User’s Guide.” [Online]. Available: http://literature.cdn.keysight.com/litweb/pdf/70874-90023.pdf.

[22] E. Bogatin, *Signal and power integrity–simplified*. Pearson Education, 2010.

[23] D. M. Pozar, “Microwave Engineering,” *Fourth Ed. Univ. Mass. Amherst John Wiley Sons Inc*, pp. 26–30, 2012.
[24] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolić, *Digital integrated circuits: a design perspective*, vol. 7. Pearson Education Upper Saddle River, NJ, 2003.
[25] R. A. Pucel, D. J. Masse, and C. P. Hartwig, “Losses in Microstrip,” *IEEE Trans. Microw. Theory Tech.*, vol. 16, no. 6, pp. 342–350, 1968.
[26] M. W. Beattie and L. T. Pileggi, “Inductance 101: Modeling and extraction,” in *Proceedings of the 38th Design Automation Conference (IEEE Cat. No. 01CH37232)*, 2001, pp. 323–328.
[27] Y. I. Ismail and E. G. Friedman, *On-chip inductance in high speed integrated circuits*. Springer Science & Business Media, 2001.
[28] J. A. Davis and J. D. Meindl, *Interconnect technology and design for gigascale integration*. Springer Science & Business Media, 2012.
[29] H. Bakoglu and J. D. Meindl, “Optimal interconnection circuits for VLSI,” *IEEE Trans. Electron Devices*, vol. 32, no. 5, pp. 903–909, 1985.
[30] J. Rubinstein, P. Penfield, and M. A. Horowitz, “Signal delay in RC tree networks,” *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 2, no. 3, pp. 202–211, 1983.
[31] W. C. Elmore, “The transient response of damped linear networks with particular regard to wideband amplifiers,” *J. Appl. Phys.*, vol. 19, no. 1, pp. 55–63, 1948.
[32] D. Pamunuwa and H. Tenhunen, “Repeater insertion to minimise delay in coupled interconnects,” in *VLSI Design 2001. Fourteenth International Conference on VLSI Design*, 2001, pp. 513–517.
[33] D. Pamunuwa, L.-R. Zheng, and H. Tenhunen, “Maximizing Throughput Over Parallel Wire Structures in the Deep Submicrometer Regime,” vol. 11, no. 2, p. 20, 2003.
[34] M. A. El-Moursy and E. G. Friedman, “Optimum wire sizing of RLC interconnect with repeaters,” *Integration*, vol. 38, no. 2, pp. 205–225, 2004.
[35] V. Adler and E. G. Friedman, “Repeater design to reduce delay and power in resistive interconnect,” *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 45, no. 5, pp. 607–616, 1998.
[36] M. Nekili and Y. Savaria, “Parallel regeneration of interconnections in VLSI & ULSI circuits,” in *1993 IEEE International Symposium on Circuits and Systems*, 1993, pp. 2023–2026.
[37] P. K. Hanumolu, G.-Y. Wei, and U.-K. Moon, “Equalizers for high-speed serial links,” *Int. J. High Speed Electron. Syst.*, vol. 15, no. 02, pp. 429–458, 2005.
[38] A. Fayed and M. Ismail, *Adaptive techniques for mixed signal system on chip*, vol. 872. Springer Science & Business Media, 2006.
[39] J. Han, Y. Lu, N. Sutardja, and E. Alon, “6.2 A 60Gb/s 288mW NRZ transceiver with adaptive equalization and baud-rate clock and data recovery in 65nm CMOS technology,” in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 112–113.
[40] J. Jaussi *et al.*, “26.2 A 205mW 32Gb/s 3-Tap FFE/6-tap DFE bidirectional serial link in 22nm CMOS,” in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 440–441.
[41] S. Palermo, “High-speed serial I/O design for channel-limited and power-constrained systems,” *CMOS Nanoelectron. Analog RF VLSI Circuits K Iniewski N. Y. NY McGraw-Hill*, pp. 289–336, 2011.
[42] S. Gondi and B. Razavi, “Equalization and clock and data recovery techniques for 10-Gb/s CMOS serial-link receivers,” *IEEE J. Solid-State Circuits*, vol. 42, no. 9, pp. 1999–2011, 2007.
[43] A. Agrawal, J. F. Bulzacchelli, T. O. Dickson, Y. Liu, J. A. Tierno, and D. J. Friedman, “A 19-Gb/s serial link receiver with both 4-tap FFE and 5-tap DFE functions in 45-nm SOI CMOS,” *IEEE J. Solid-State Circuits*, vol. 47, no. 12, pp. 3220–3231, 2012.
[44] Tim Wang Lee, “Eye-opening Experience with FFE,” *Keysight*, Jan. 12, 2018. https://community.keysight.com/community/keysight-blogs/eesof-eda/blog/2018/01/12/tim-s-blackboard-9-eye-opening-experience-with-ffe?_sm_au_=iVVP04kSFkkrrkTHcLpsvK618Vf61.

[45] S. Kasturia and J. H. Winters, “Techniques for high-speed implementation of nonlinear cancellation,” *IEEE J. Sel. Areas Commun.*, vol. 9, no. 5, pp. 711–717, 1991.
[46] H. Wang and J. Lee, “A 21-Gb/s 87-mW transceiver with FFE/DFE/analog equalizer in 65-nm CMOS technology,” *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 909–920, 2010.
[47] J. Savoj *et al.*, “Design of high-speed wireline transceivers for backplane communications in 28nm CMOS,” in *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*, 2012, pp. 1–4.
[48] R. W. Lucky, “The adaptive equalizer,” *IEEE Signal Process. Mag.*, vol. 23, no. 3, pp. 104–107, 2006.
[49] V. Stojanovic *et al.*, “Autonomous dual-mode (PAM2/4) serial link transceiver with adaptive equalization and data recovery,” *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 1012–1026, 2005.
[50] Y. Hidaka, W. Gai, T. Horie, J. H. Jiang, Y. Koyanagi, and H. Osone, “A 4-channel 1.25–10.3 Gb/s backplane transceiver macro with 35 dB equalizer and sign-based zero-forcing adaptive control,” *IEEE J. Solid-State Circuits*, vol. 44, no. 12, pp. 3547–3559, 2009.
[51] R. T. Chang, N. Talwalkar, C. P. Yue, and S. S. Wong, “Near speed-of-light signaling over on-chip electrical interconnects,” *IEEE J. Solid-State Circuits*, vol. 38, no. 5, pp. 834–838, 2003.
[52] M.-C. Chang *et al.*, “Advanced RF/baseband interconnect schemes for inter-and intra-ULSI communications,” *IEEE Trans. Electron Devices*, vol. 52, no. 7, pp. 1271–1285, 2005.
[53] S.-W. Tam, E. Socher, A. Wong, and M.-C. F. Chang, “A simultaneous tri-band on-chip RF-interconnect for future network-on-chip,” in *2009 symposium on VLSI circuits*, 2009, pp. 90–91.
[54] M. Jalalifar and G. Byun, “A 14.4Gb/s/pin 230fJ/b/pin/mm multi-level RF-interconnect for global network-on-chip communication,” in *2016 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2016, pp. 97–100.
[55] Y. Kim *et al.*, “Analysis of noncoherent ASK modulation-based RF-interconnect for memory interface,” *IEEE J. Emerg. Sel. Top. Circuits Syst.*, vol. 2, no. 2, pp. 200–209, 2012.
[56] H. Wu *et al.*, “A 60GHz on-chip RF-interconnect with λ/4 coupler for 5Gbps bi-directional communication and multi-drop arbitration,” in *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*, 2012, pp. 1–4.

Chapter II : System overview

In the first chapter, we introduced serial links and their main metrics as well as the different types of interconnects, namely RC and LC interconnects. This chapter discusses the second part of this work, the adopted innovative duobinary bandpass modulation.

Firstly, we introduce some of the main (line code) modulations; then we discuss the correlative techniques known as polybinary modulation for both baseband and passband modulation. In this work, we used passband modulations to take advantage of the high available bandwidth, amongst other benefits. Secondly, we present the architecture that makes this transmission possible; Then, we show that this technique can be adapted for serial links of different lengths in the length adaptive serial link section to make this technique more attractive for digital circuits implementation.

Finally, we also introduce some of the components and their main parameters.

CHAPTER II : SYSTEM OVERVIEW

I. Overview of data transmission main functions

The purpose of any data transmission is to keep its signal integrity, e.g., zero error, even in the presence of noise. In the real world, however, we do not always get what we want. So, is it possible to have a zero-error transmission?

Shannon [1] proved in the 1940s that designing a system with a probability of error approaching zero is possible, under certain assumptions. He proposed a theorem that defines the channel's theoretical limit to carry information, also known as the channel capacity. The channel capacity C (bits/s) is calculated in the presence of white Gaussian noise on the signal and depends on the channel bandwidth B in Hertz (Hz) and the signal-to-noise ratio $\frac{S}{N}$. It is given by:

$$C = B \log_2 \left(1 + \frac{S}{N}\right) \tag{II-1}$$

Theoretically, if the rate of information R (bits/s) in the transmission system is less than the channel capacity C (bits/s), the error probability approaches zero. To maximize the channel capacity, we can either improve the signal to noise ratio or the bandwidth.

I.1. The bandwidth of a signal

Each signal has a unique composition and is formed of different frequencies. For example, the ideal square wave shown in Figure II.1 can be written as a combination of a sine wave at a frequency f and a series of its odd harmonic frequencies ($3f, 5f, 7f$ ...). The frequency-domain representation of such periodic ideal square signal, also known as spectrum, shows that the amplitude of the harmonics is inversely proportional to f [2], and is null for all even harmonics. The frequency-domain representation of a signal is vital and provides valuable insights. The power spectral density (PSD) of a signal is of interest; it describes its power content as a function of frequency [3].

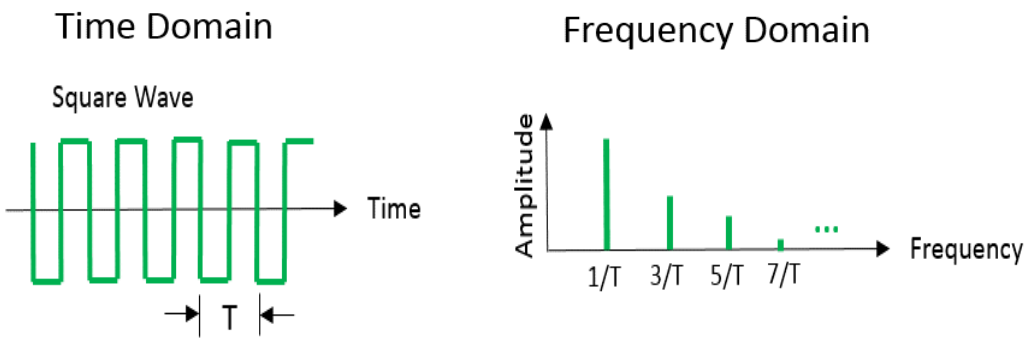


Figure II.1 : Time and frequency domain views of an ideal square wave.

The bandwidth of a signal is very important; for example, most of the RF bands are licensed, and international organisms regulate their use, i.e., users are assigned a particular band, and they are allowed to use this band only. Thus, the equipment and modulation should be designed to take advantage of the available bandwidth efficiently. The bandwidth of a signal depends on its modulation.

The term bandwidth has several definitions [4], as shown in Figure II.2. The appropriate definition depends on the application. Firstly, the absolute bandwidth $(f_2 - f_1)$ describes where the spectrum is zero outside the interval $f_1 < f < f_2$.

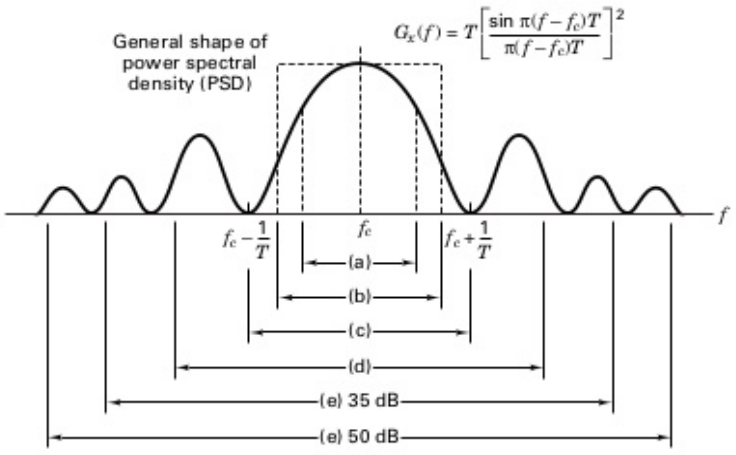


Figure II.2 : Bandwidth of digital data. (a) Half-power. (b) Noise equivalent. (c) Null to null. (d) 99% of power. (e) Bounded PSD (defines attenuation outside bandwidth) at 35 and 50 dB [3].

Secondly, the 3-dB bandwidth is the bandwidth $(f_1 < f < f_2)$ containing half of the power where $|H(f)| > \left|\frac{\max(H(f))}{\sqrt{2}}\right|$.

Null-to-null bandwidth describes the width of the main spectral lobe of the power spectral density (PSD)[4]. For most baseband digital modulations, this is where the majority of the signal power is contained.

I.2. Classical baseband modulations (NRZ, RZ, MLS)

Line code is the technique used to convert a sequence of bits into a digital signal. Hereafter, we describe the most common ones. Non-return-to-zero (NRZ) is the most popular line coding in serial transmission, mainly due to its simplicity: “1” is high level and “0” is low level. Different variations of NRZ exist. The NRZ family is not inherently self-clocking, i.e., the clock is not a part of the signal.

On the other hand, return-to-zero (RZ) is a self-clocking line code, and as its name suggests, returns to 0 levels for each bit, and thus RZ does not require a separate clock to be sent. However, for the same data rate, RZ requires double the NRZ bandwidth.

Another approach is to use multi-level signaling (MLS) instead of binary signaling. One can argue that binary signaling is the simplest form of multi-level signaling, where the two levels represent a single-bit symbol. Increasing the number of bits represented by a single symbol increases the channel's capacity. For example, four-level pulse amplitude modulations (PAM4) uses four voltage levels to represent four combinations of two bits (“00”, “01”, “10”, ”11”); these combinations are also known as symbols.

Figure II.3.a. shows the coding of the word "001001101110" in NRZ (or PAM2) and PAM4. The time it takes to send the word using PAM4 is half that of the time required for the NRZ.

This modulation has many advantages, especially in terms of bandwidth. Figure II.3.b. shows the power spectrum density of both modulations. We notice that for the same bit rate (BR), the null-to-null bandwidth is halved using PAM4 modulation compared to NRZ modulation. This means that for available bandwidth, by using PAM4 instead of NRZ, we can double the density of data.

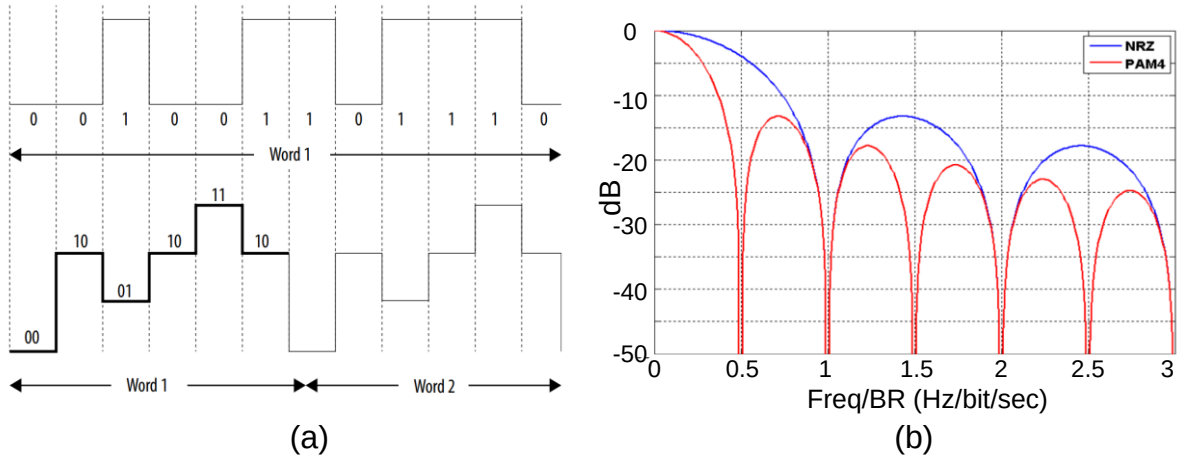


Figure II.3 : (a). Time-domain NRZ and PAM4 coding. (b). Comparison of an NRZ and PAM4 power spectrum [5].

Nevertheless, there are some disadvantages, such as SNR degradation, due to reduced spacing between the voltage levels. In fact, the PAM4 signal has a third of the amplitude of a similar NRZ signal, as shown in Figure II.4. The SNR loss of PAM4 modulation compared to NRZ modulation is given by:

$$SNR\ loss\ in\ dB = 20\log_{10}\left(\frac{1}{3}\right) = -9.5\ dB \qquad \text{(II-2)}$$

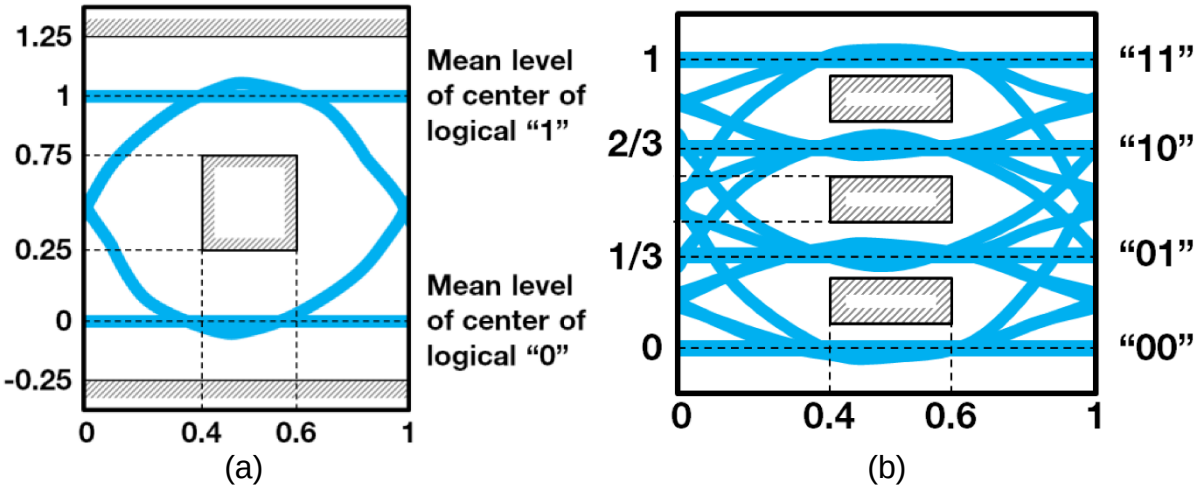


Figure II.4 :(a) NRZ eye diagram. (b) PAM4 eye diagram.

I.3. Correlative baseband modulations

Correlative techniques [6-7] allow the distribution of the spectral energy. Generating a time wave consisting of correlated signaling levels permits overall spectrum shaping. For example, it is possible to eliminate any power at low frequencies

or concentrate most of the energy at low frequencies. The latter is possible using polybinary modulation [6].

Polybinary modulation was developed by Lender [6], [8] in the 1960s as a more straightforward replacement for quaternary modulation. In its simplest form, known as duobinary [8], this modulation can double the data rate transmitted through a channel bandwidth compared to NRZ signaling. This modulation has the same spectral efficiency as PAM4 modulation for the same data rate but is less complicated.

Polybinary modulation, in its general form, is a two-step transformation of the binary data at the transmitter input into an M-level signal at the receiver input. M can be either even or odd; however, it was shown in [9] that this modulation is only beneficial for an odd number of levels.

The first step of polybinary modulation consists of precoding the binary data sequence a_n into a second binary signal d_n; this transformation is also known as differential encoding. This first transformation consists of a modulo-2 sum operation, also known as an exclusive-or logic operation, of the present bit of a_n and the preceding $(M-2)$ bits of d_n as follows:

$$d_i = a_i \oplus d_{i-1} \oplus d_{i-2} \oplus \dots \oplus d_{i-M+2} \tag{II-3}$$

Next, the two-level signal is transformed into an M-level polybinary signal p_n. This sequence is formed by adding the present digit algebraically to the previous $(M-2)$ digits of the sequence p_n; this operation can be expressed as:

$$p_i = A \sum_{k=0}^{M-2} d_{i-k} \tag{II-4}$$

With A a constant.

As a result of these two steps, the binary "0" (or "1") of the initial sequence is mapped into even (or odd) number levels of p_n. Consequently, each digit in p_n is independently identified despite the correlation characteristics; this significantly simplifies the detection process. Furthermore, two sequential elements of p_n can vary in value by only one level; thus, this encoding possesses an innate potential for error detection.

For the rest of our reasoning, we will focus on duobinary modulation, i.e., using a 3-level polybinary modulation. To better understand this principle's advantages, we compare the PSD of a binary and a duobinary pulse train signal.

I.3.1. Comparison of PAM2, PAM4, and Duobinary modulations

The PSD of a binary signal is given by [4]:

$$W_1(f) = \frac{1}{T_b}|G(f)|^2 pq \tag{II-5}$$

where p is the probability of transmitting "0", and $q = (p-1)$ is the probability of transmitting "1", T_b is the time symbol, and $G(f)$ is the Fourier transform of the pulse shape. $G(f)$ depends on the pulse shape, e.g., a rectangular or a raised cosine shaping. For a rectangular pulse shaping, $G_{rect}(f)$ is defined as:

$$G_{rect}(f) = T_b \frac{\sin \pi f T_b}{\pi f T_b} \quad \text{(II-6)}$$

Thus, for rectangular shaped NRZ pulses, and considering that the probability of "0" and "1" data bits is equal ($p = q = 0.5$), the PSD of the binary signal is given by:

$$W_b(f) = \frac{T_b}{4} \left(\frac{\sin \pi f T_b}{\pi f T_b} \right)^2 \quad \text{(II-7)}$$

The PSD of the duobinary signal is given by:

$$W_{duo}(f) = \frac{T_b}{4} \left(\frac{\sin 2\pi f T_b}{2\pi f T_b} \right)^2 \quad \text{(II-8)}$$

$W_b(f)$ and $W_{duo}(f)$ are shown in Figure II.5, we notice that the duobinary energy is more concentrated near-zero frequency than the binary one [10]. It is worthwhile to mention that almost 90% of the signal power is located in the main lobe.

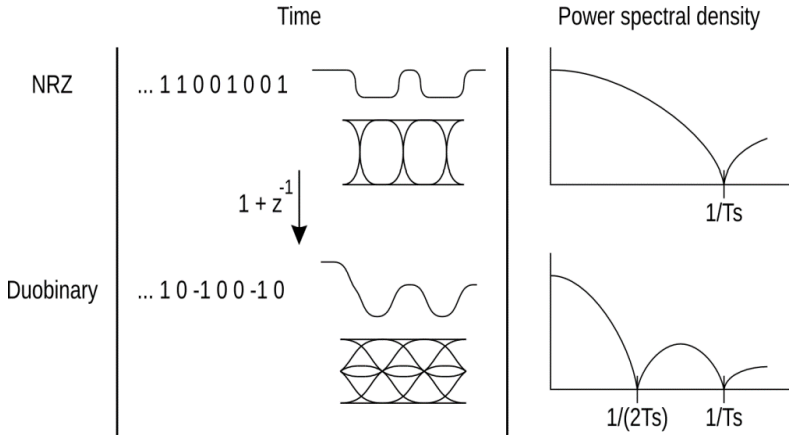


Figure II.5 : NRZ and duobinary: waveform, eye diagram, and spectrum [10].

The authors in [11] showed that by using a "brickwall" filter at $1/2\,T_b$, and for signals with the same data rate, an NRZ signal loses 51,4% of its power while a duobinary loses only 10%. Consequently, the duobinary suffers from less ISI and jitter issues, as shown in the eye diagram in Figure II.6.

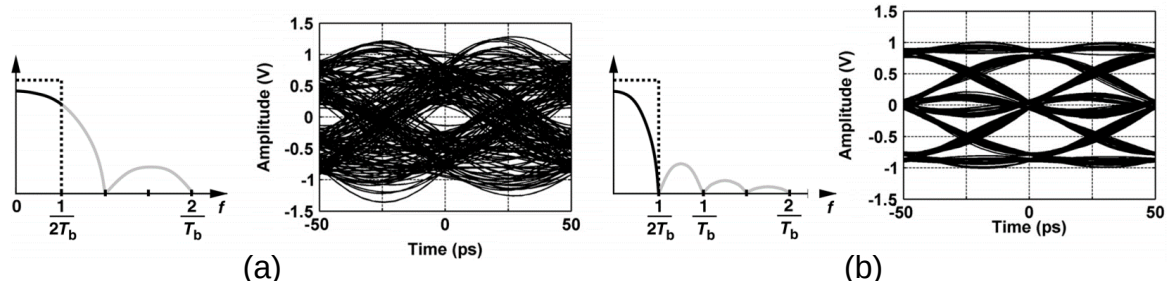


Figure II.6 : Spectrum and waveforms (data rate 20Gb/s) for different data formats passing through an ideal brickwall filter. (a) NRZ. (b) Duobinary [11].

In this section, we discussed PAM2 (NRZ) signaling, PAM4, and duobinary. We established that duobinary modulation spectral efficiency is double that of NRZ signaling, as shown in Figure II.7, but is equally efficient as PAM4. In addition, for a fixed transmitted peak-power, PAM4 suffers from a reduction of voltage margins due

to the four voltage levels. In comparison, duobinary has only three levels, as shown in Figure II.7.

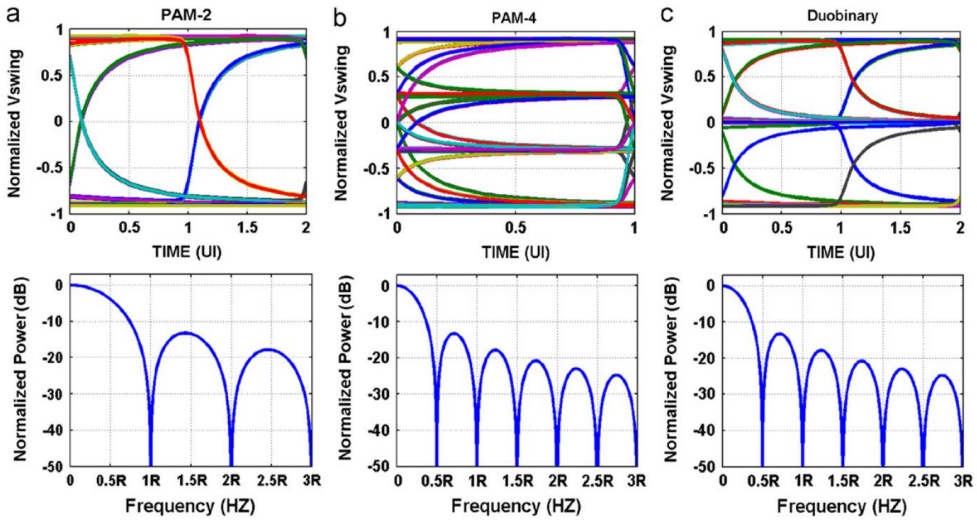


Figure II.7 : Eye diagram and Normalized power density spectrum: (a) PAM2. (b) PAM4. (c) Duobinary.[12].

Duobinary transmission systems are based on three parts: a differential encoder, a reshaping filter (this function can be accomplished using the transmission channel or a dedicated filter), and a decoder (in the receiver). These parts are described in the next sections.

I.3.2. Duobinary differential encoder

The first encoding is vital for the coding's right functioning as it protects the system from error propagation; it is also used prior to modulation in several communication systems. Differential precoder encodes the changes of bits; in other words, states' changes define the resulting data stream after differential encoding. Where a change “0” to “1” (or “1” to “0”) in the initial binary stream produces a “1” in the newly encoded stream. Whereas two successive similar bits,i.e., a “0” (“1”) followed by a “0” (“1”) result in an encoded binary “0”. This operation is called modulo-2 addition or exclusive-or logic (XOR) operation; its implementation is described in Figure II.8.a.

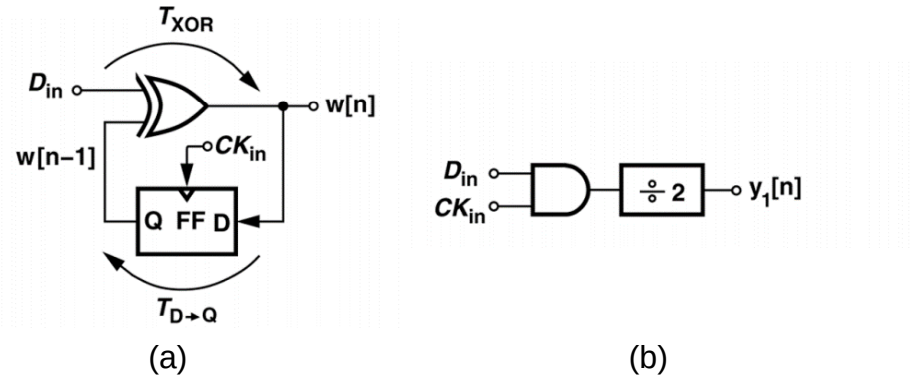


Figure II.8 : Duobinary precoder. (a) conventional. (b) Proposed [11].

The delay in the differential encoder loop is the bottleneck for such a circuit. It can be implemented using a clock-driven flip-flop, for example. However, this implementation is problematic due to the feedback loop, especially for high data rates, where timing constraints are significant [11]. To work properly, the sum of both the XOR gate and the flip-flop should be equal to the bit period T_b:

$$T_b = T_{XOR} + T_{FF} \tag{II-9}$$

This means that the flip flop (FF) clock signal should be tightly controlled. Hence the complexity of such controlling circuits increases significantly at high data rates. To avoid these constraints, the author in [11] proposed a more practical feed-forward scheme without the delay issue; its implementation is shown in Figure II.8.b.

This implementation requires a Divide-by-2 circuit and an AND gate. When the $d_k = 1$, the clock is transmitted through the AND gate. Consequently, the output state of the circuit changes. When the $d_k = 0$, the clock is not transmitted through the AND gate, and the circuit output does not change. This implementation results in the same operation as the modulo-2 adder, without the critical delay issue. This implementation is limited by the bandwidth of the logic gates at high data rates, where the signal at the output of the AND gate duration is half of the original data duration.

I.3.3. Duobinary reshaping filter

The architecture of the duobinary reshaping filter, in its original implementation [6], [8], is shown in Figure II.9. This architecture precodes the signal, as explained in the differential precoder section. Next, to transform the NRZ signal into a 3-level signal, the coded binary output goes through a filter whose Z-transform is $H(z) = (1 + z^{-1})$. Its frequency response is shown in Figure II.10.

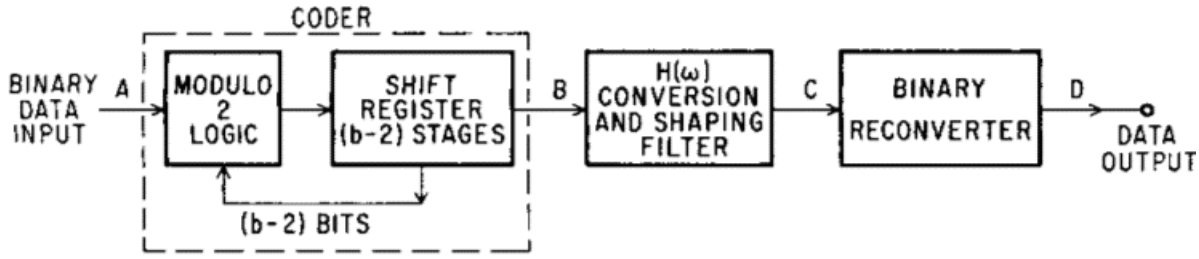


Figure II.9 : Original polybinary architecture [6].

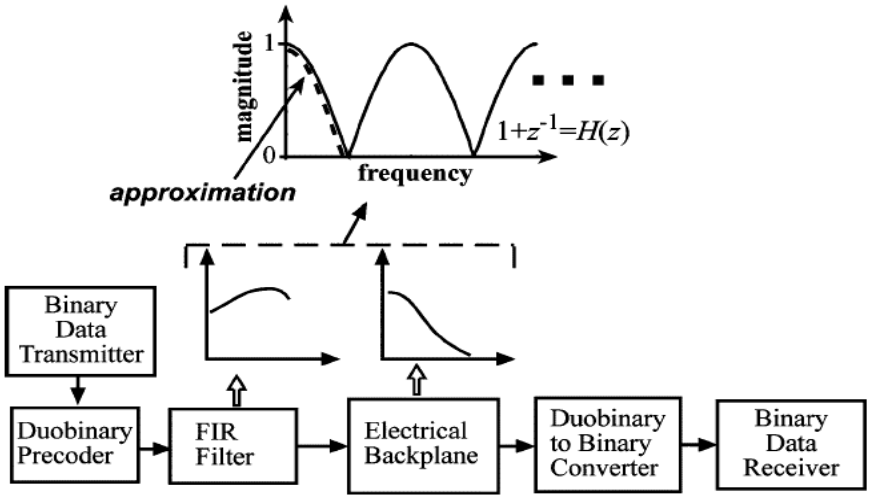


Figure II.10 : System architecture and duobinary required frequency response [13].

The duobinary required frequency response is illustrated in Figure II.10. We notice that this "ideal" filter has a low-pass filter response at low frequency and a passband response at higher frequencies. In practice, we are unable to duplicate this filter, but we can get reasonably close to this response at low frequency. The duobinary technique generates the three levels signal from two-level one using the low-pass behavior. It takes advantage of the channel to generate this low-pass behavior since most wireline channels behave as low-pass filters with different roll-off values. Thus, the channel can be used as a base to generate the required $H(z) = (1 + z^{-1})$ by cascading additional filters (or equalizers).

For example, in [14], the authors showed that a typical low-cost backplane has a frequency roll-off steeper than the one required for duobinary modulation. Thus, they emphasized the high-frequency components of the signal and flattened the group delay over the used bandwidth, using the two-tap pre-emphasis filter shown in Figure II.11.

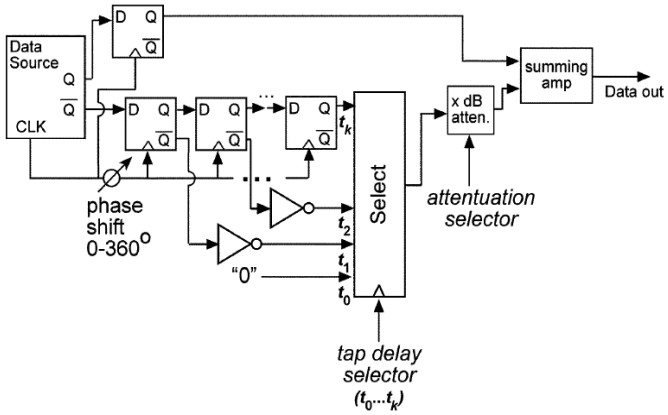
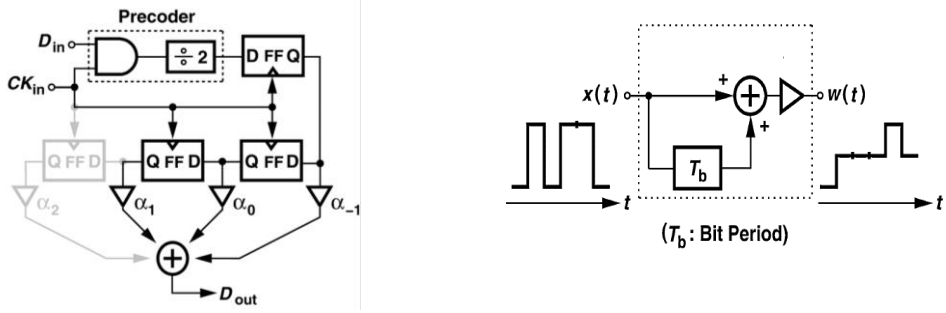


Figure II.11 : Proposed implementation of the FIR pre-emphasis filter [14].

To transmit a 20Gb/s duobinary signal [11], the authors choose the duobinary transmitter shown in Figure II.12.a. The blocks were implemented in current-mode logic (CML) for faster operation. This filter shaping operation results in a 3-levels signal, as shown in Figure II.12.b.



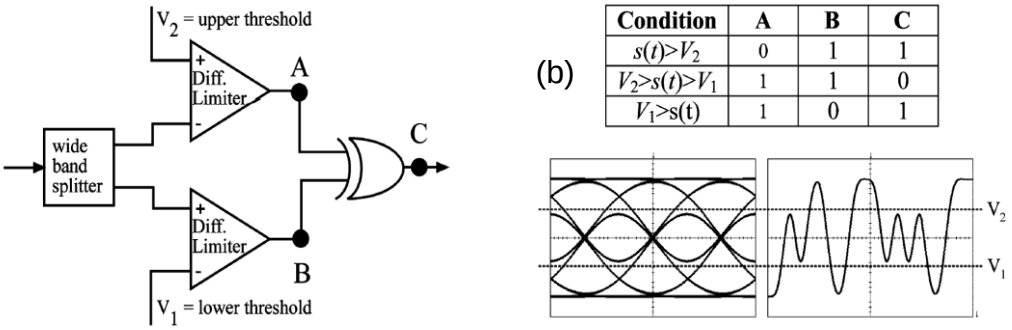
(a) (b)

Figure II.12 :(a) Proposed duobinary transmitter[11]. (b) Duobinary signaling model.

I.3.4. Duobinary decoder

When Lender [8] invented the duobinary modulation, he suggested using a full-wave rectifier to reconvert the received 3-levels signal, shown in Figure II.12.b., into a 2-level one and double the data bandwidth. Next, the binary signal goes through a slicer; the slicer threshold is halfway between high and low voltage levels. The output of the slicer should correspond to the transmitted data if the differential precoder is used. In this case, each bit can be recovered independently of the preceding bits. In other words, if an error occurs in one of the bits, the following bits will not be affected.

A second approach was suggested in [14] for high data rates because it is challenging to design a full-wave rectifier with a flat response over such a broad bandwidth without complex matching circuitry for tens of Gbps. Thus, they proposed the architecture presented in Figure II.13.a., where the signal is compared to V_1 (low threshold) and V_2 (high threshold) using two comparators simultaneously, the outputs A and B go through an exclusive-or (XOR) gate to recover the binary transmitted data. This receiver distinguishes the middle level (binary "0" for example) from the side levels ("1"), as shown in Figure II.13.b.



(a) (c)

Figure II.13 : (a) Proposed duobinary-to-binary converter [14]. (b) Circuit truth table (c) Duobinary signal as an eye diagram and a time waveform.

This solution requires accurate threshold voltages; however, depending on PVT variations, the threshold values drifting might lead to jitter. To resolve this, the authors in [11] proposed to use the architecture presented in Figure II.14. They included a negative feedback loop to compensate for the threshold levels drifting due to mismatch or PVT variations. This loop allows for threshold control and optimization. According to the authors, the jitter can be further reduced by implementing a CDR.

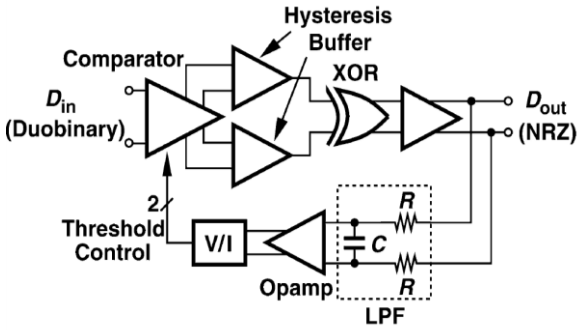


Figure II.14 : Proposed duobinary-to-binary converter [11].

To correctly demodulate the baseband duobinary signal, all voltage states must be distinguished accurately. However, this condition is relaxed if bandpass modulation is used.

I.4. Intermediate frequency transposition

In this work, we will use an intermediate frequency "IF" approach-based technique. This technique is widely used in nowadays communication systems, where part of the circuitry is used to transpose (or up-convert) a low-frequency signal into a higher frequency transmitted through the channel. Different parts (or frequency bands) of the channel can be used simultaneously using the IF technique, if the channel allows it, as used for radio channels or mobile communications. In this case, the transmitter and/or receiver are compatible with a broad range of possible carrier frequencies, usually by tuning the circuitry to transmit/receive the signal in a specific frequency band and ignore the rest. For wireless channels, the signal propagates through the air; the available frequency bands are very limited due to the governments' strict standardization. Furthermore, the transmission power in the considered frequency band (and neighbor bands) is regulated.

In this work, we will not use wireless channels, and thus we will not have the same specifications/constraints, but instead, we use transmission lines. Furthermore, we will use a high-frequency carrier to transmit the signal over the channel. In the frequency domain, it translates into moving the spectrum of our data to a higher frequency. Figure II.15 shows a practical example of basic amplitude modulation at the transmitter side. The top row in Figure II.15 describes the time-domain signals, while the bottom row shows its frequency representation. This up-conversion operation consists of multiplication between the carrier and the baseband data, which translates into a frequency-domain convolution. Of course, complicated modulations require complicated circuitry.

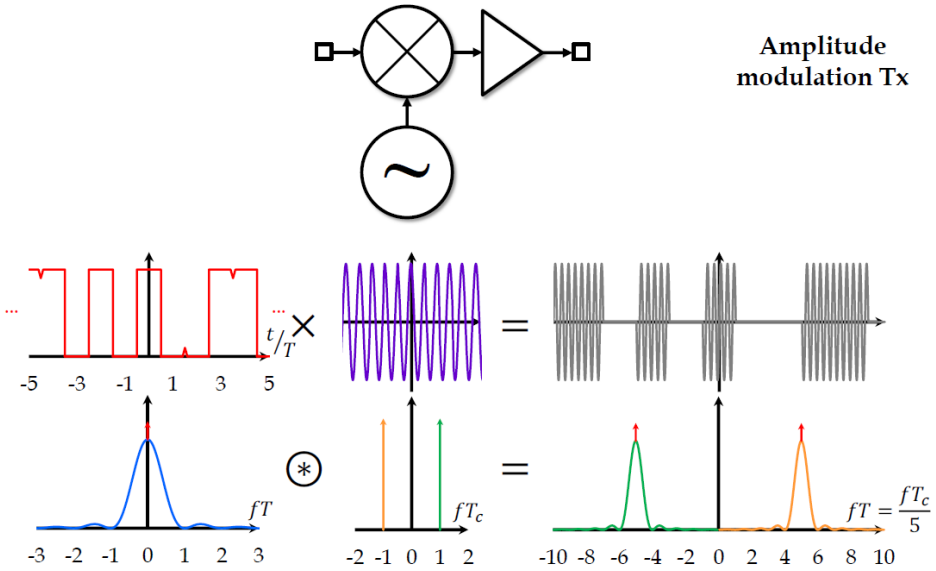


Figure II.15 : Basic amplitude modulation [15].

The basic architecture of an RF receiver used to demodulate the signal is described in Figure II.16, where the incoming signal is multiplied by the same carrier to recover the data baseband signal. As shown in Figure II.16, this approach's challenge, also known as coherent demodulation, is that any drift in the carrier frequency used for the receiver multiplication leads to severe signal degradation.

Usually, a low-pass filter is inserted at the end of this chain to block higher frequency signals.

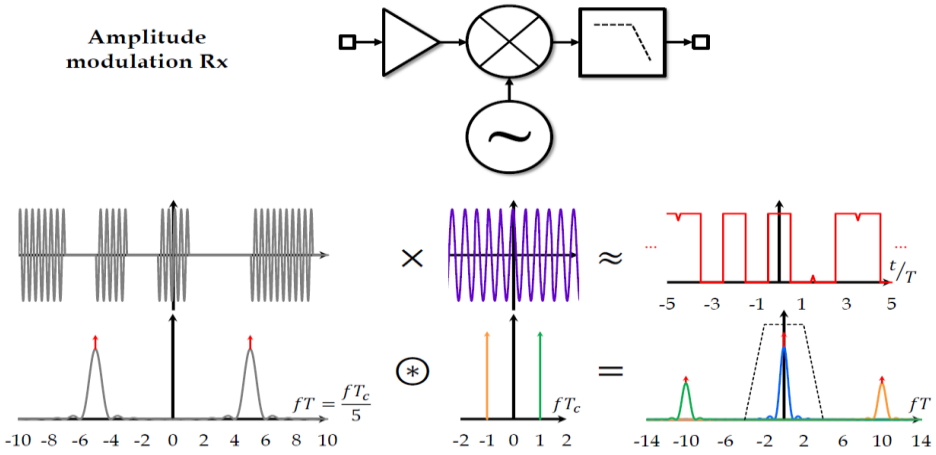


Figure II.16 : Basic amplitude demodulation[15].

I.4.1. Polybinary bandpass modulation

We studied Duobinary modulation for baseband, but it is also suitable for carrier modulation, or as Lender [6] called it Polybinary AM-PSK. Polybinary AM-PSK results are essential for this work. Before developing it, it is worth mentioning that this approach can only be applied for an odd number of levels.

For a duobinary signal, i.e., 3-levels signal, we showed that the highest and lowest voltage levels correspond to the same binary state "0", while the middle voltage level corresponds to a binary "1". The same approach can be applied for a carrier transmission, where the carrier's absence means a binary state "1", while the presence of carrier represents the binary "0". In this case, the difference between high and low voltage levels is the reversal of the carrier signal by $180°$, as shown in Figure II.17.b.

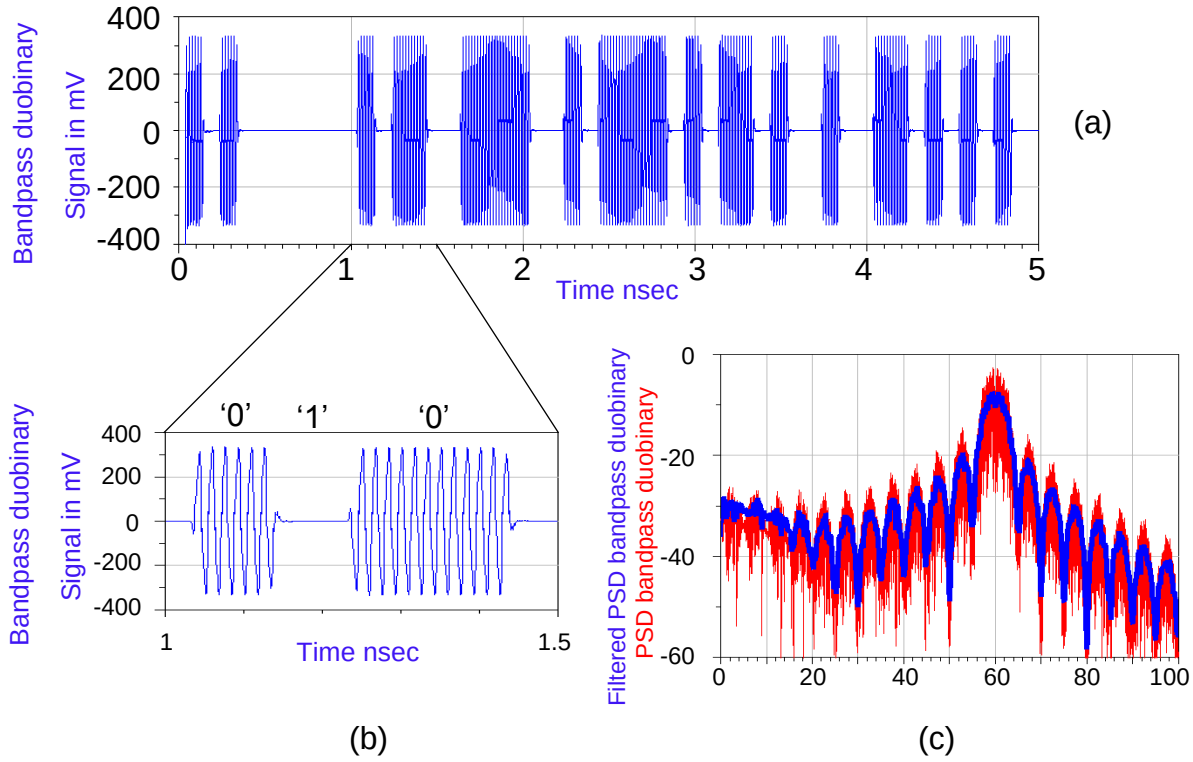


Figure II.17 : (a) Pass-band duobinary time-domain signal. (b) phase reversal in time-domain passband duobinary signal. (c) PSD and filtered PSD of a duobinary passband signal.

In this case, the detection is simple, and the demodulation only requires the envelope of the carrier; and the phase reversal is ignored. This way, using a carrier, this modulation keeps the same spectral efficiency as duobinary modulation, i.e., compression of the spectrum by a factor of two compared to the conventional binary ASK modulation. In other words, for the same available bandwidth, this modulation allows us to double the data rate compared to ASK modulation.

Figure II.17.b shows the PSD of the chosen duobinary passband signal. In this case, we used a 60GHz carrier, and we notice that for a 10Gb/s data, the first null is only 5GHz away from the 60GHz in both directions (55GHz and 65GHz). The compression ratio is equal to two, as shown for standard baseband duobinary modulation.

I.4.2. Mixers parameters and terminology

The modulator, also known as a mixer [16]–[18], is a three-ports device (active or passive); one port is used for the radio frequency (RF) input, the second is used for the local oscillator (LO) input, while the third is called the intermediate frequency (IF) output port. The mixer takes an input signal at a frequency f_{RF}, usually low-frequency signal, and mixes it with the LO signal, with the carrier frequency f_{LO}, and produces

the IF output signal that consists of the sum and difference of the RF frequency f_{RF} and f_{LO}, $(f_{RF} \pm f_{LO})$, as shown in equation (II-13). In other words, it converts the signal from one frequency to another. We suppose the input signals are defined as follows:

$$v_{RF} = A(t) \cdot \cos\left(\omega_{RF} \cdot t + \emptyset(t)\right) \tag{II-10}$$

$$v_{LO} = A_{LO} \cdot \cos(\omega_{LO} \cdot t) \tag{II-11}$$

Multiplying the signals:

$$v_{out} = v_{RF} \cdot v_{LO} \tag{II-12}$$

Then simplifying the results gives the output of an ideal multiplier:

$$v_{out} = \frac{A(t) \cdot A_{LO}}{2} \cdot \begin{bmatrix} \cos\left((\omega_{RF} + \omega_{LO})t + \emptyset(t)\right) \\ + \cos\left((\omega_{RF} - \omega_{LO})t + \emptyset(t)\right) \end{bmatrix} \tag{II-13}$$

Usually, designers place filters following the mixer to choose only one of the frequency bands $(f_{RF} + f_{LO})$ or $(f_{RF} - f_{LO})$, also known as sidebands. However, there are techniques called single sideband (SSB) down/up-conversion to suppress one of the sidebands in the mixing process.

When the sum of frequencies $(f_{RF} + f_{LO})$ is used, we say that the mixer is upconverting, i.e., the signal is transposed from one RF frequency to a higher frequency IF. When $(f_{RF} - f_{LO})$ is chosen, we say that the mixer is downconverting; in this latter case, the IF and RF ports are inverted.

In mixer terminology, we also speak about low-side and high-side injection. In a receiver (downconversion), If f_{LO} is lower than f_{RF}, It is called the low-side downconversion. When f_{LO} is higher than f_{RF} is it called high-side downconversion.

For an ideal multiplication, equation (II-13) produced two resulting frequency components only. However, mixers are not ideal nor linear devices. Thus, after the mixing process, several frequency components appear at the output; they can be classified as desired signals $(f_{RF} + f_{LO})$ or $(f_{RF} - f_{LO})$, and undesired signals such as harmonics of the signals, leakage such as LO signal, or results of the mixer non-linearity, as shown in Figure II.18.

Two main categories of mixers can be distinguished, passive mixers and actives ones. Usually, passive mixers use passive non-linear devices such as diodes or switches, while active mixers require transistors. Each type has its advantages and disadvantages; for example, passive mixers do not provide conversion gain as it is the case of active mixers, but active mixers lead to power consumption.

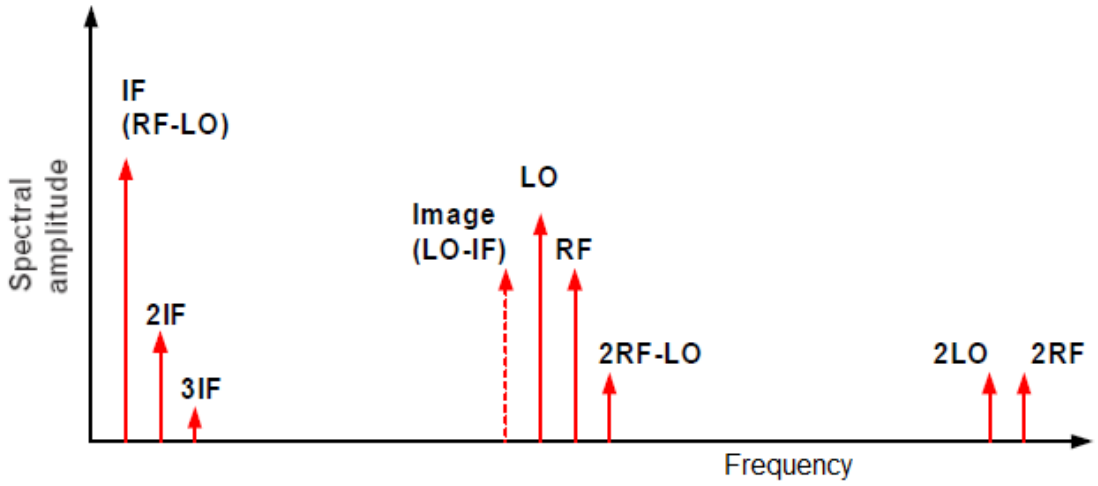


Figure II.18 : RF Mixer spectral components.

I.4.2.1 Conversion gain

Conversion gain CG is usually given in dB, it is defined as the ratio of the IF power and the RF power, given by:

$$CG = 10log\left(\frac{P_{IF}}{P_{RF}}\right) \quad \text{(II-14)}$$

The conversion gain can also be defined in terms of voltages as the ratio of the output IF voltage and the RF input voltage, given by:

$$CG = 20log\left(\frac{V_{IF}}{V_{RF}}\right) \quad \text{(II-15)}$$

I.4.2.2 Port to Port Isolation

As mentioned above, we usually find some leakage signals (LO or RF) in the output IF signal. The level of these signals is referred to as the mixer isolation. When isolation is high, the leakage is minimal.

In practice, part of the signal can leak from one port to another in the mixer; we distinguish three main isolation parameters:

1) LO to RF: when the LO signal appears at the RF port, the signal can potentially flow to the antenna and radiate, if the LNAs isolation $|S_{12}|$ is poor.
2) LO to IF: when the LO signal appears at the IF port, if the signal is large enough or not sufficiently filtered, it may saturate the succeeding circuits.
3) RF to IF: when the RF signal appears at the IF port. In this case, filtering is usually sufficient to suppress the small RF signal.

I.4.2.3 Noise Figure

Noise is generated by different mixer components (switches, diodes, and transistors). Noise Figure (NF) is defined as the ratio of the input signal to noise ratio to the output SNR; it is given by:

$$NF = \frac{SNR_{IN}}{SNR_{OUT}} \quad \text{(II-16)}$$

For mixers, we can distinguish two types of NF:

1) Double sideband (DSB) NF includes the signal and its noise contribution at both the RF and its image frequencies.
2) Single sideband (SSB) NF does not include the image signal contribution, although its noise is included.

If the signal and its image experience equal gains at the RF port, the DSB and SSB NF can be related by [18]:

$$NF_{DSB} = NF_{SSB} - 3\ dB \quad \text{(II-17)}$$

I.4.2.4 Linearity

Several intermodulation frequencies are generated due to the non-linearity of a mixer. Usually, the second, third, and higher harmonics are outside the band of interest and are easily filtered. However, non-linear devices may produce signals very close to the band of interest when two RF signals of equal amplitude arrive at the device's input.

Mixers' linearity is usually characterized by the third-order intercept point (IP3). It is measured by applying two closely spaced tones f_1 and f_2 at the input of the mixer, as shown in Figure II.19. Figure II.19 also shows the IF downconverted output of such input signals. We notice the presence of several components $(2f_1 - f_2)$ and $(2f_2 - f_1)$ close to the desired IF signal. The distance between intermodulation products and the adjacent carrier is the same as between the carrier signals. Thus, if the carrier signals are very close to each other, it is hard to filter the unwanted intermodulation products.

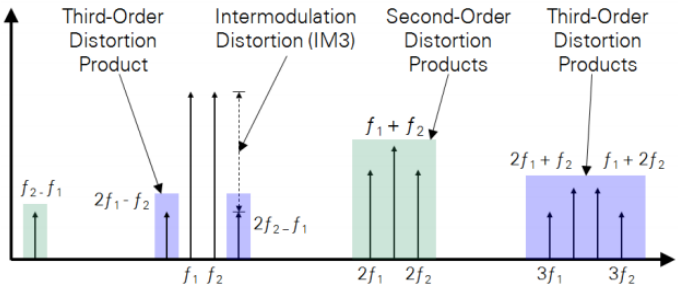


Figure II.19: Downconversion of third-order IM products [19].

The figure of merit IP3 is a theoretical point at which the third-order signal amplitude is equal to the fundamental signal amplitude, as shown in Figure II.20.

To get this figure, we plot on a logarithmic scale, the input and output power values of the fundamental signal power, and the third-order signal power, then extend the linear portions of the two gain curves. The two lines meet at the IP3 point. This point can be referenced to the input (IIP3) or output (OIP3) power.

To summarize, the IP3 indicates how large a signal the device can process before intermodulation distortion happens. Thus, for better linearity and lower intermodulation distortion, a higher output at the intercept point is required.

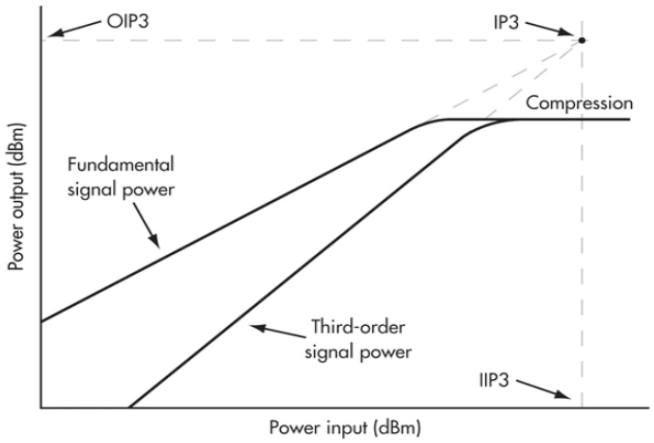


Figure II.20: Third-order intercept point and 1-dB compression points [20].

Another important parameter to characterize the linearity of the mixer is the 1-dB compression point. It is also shown in Figure II.20. It is defined as the input power (dBm) that causes the gain to degrade 1 dB from the theoretical linear first-order gain plot. Moreover, the approximative expression that relates IIP3 to 1-dB compression point [16] is given by:

$$IIP3\ (dBm) \approx P_{1-dB}\ (dBm) - 9.6 \tag{II-18}$$

I.4.2.5 Active and passive mixers

In this section, we briefly introduce the two types of mixers, passive and active ones. Detailed studies can be found in [16]–[18].

Active mixers are attractive because they can provide conversion gain, whereas passive ones usually have insertion loss instead. Active mixers are usually composed of two stages, a transconductance stage that converts the voltage to current, and a switching stage. Thus, this type of mixer consumes DC power.

The active mixing cell, in its simplest form, is shown in Figure II.21.a. The transistors in the switching pair, shown in Figure II.21.a, alternate between the saturation region (non-linear region) and cut-off state for a chosen LO voltage. Thus, the multiplication of LO and RF signals is accomplished by redirecting the current $I_{DC} + i_{RF}$ through one of the two branches. This scheme suffers from the feedthrough of LO harmonics to the output IF port; this is due to the addition of the I_{DC} term in the multiplication. These harmonics usually fall in the band of interest and removing them is difficult, as for intermodulation products.

The double-balanced scheme shown in Figure II.21.b. is used to avoid this issue. Basically, it consists of two single-band mixers joined together. This scheme permits to reject the LO harmonics and reduces common-mode noise by summing in-phase and out-of-phase signals ($0°$ and $180°$) of the RF or the LO.

The most common active mixer scheme is shown in Figure II.21.b, and was introduced by Gilbert in 1968. Several improvements were proposed afterward by Gilbert himself (in 1997) or other researchers. It is a fully differential and symmetrical topology. The differential output is preferred because it allows higher conversion gain, and better immunity against common mode noise, and it has better port-to-port isolation. The input differential pair composed of M1 and M2 converts the input RF

signal to an output current. This stage is responsible for the conversion gain. Next, the current flows to the IF output through the transistors M3, M4, M5, and M6 provides the frequency translation through its LO commutation.

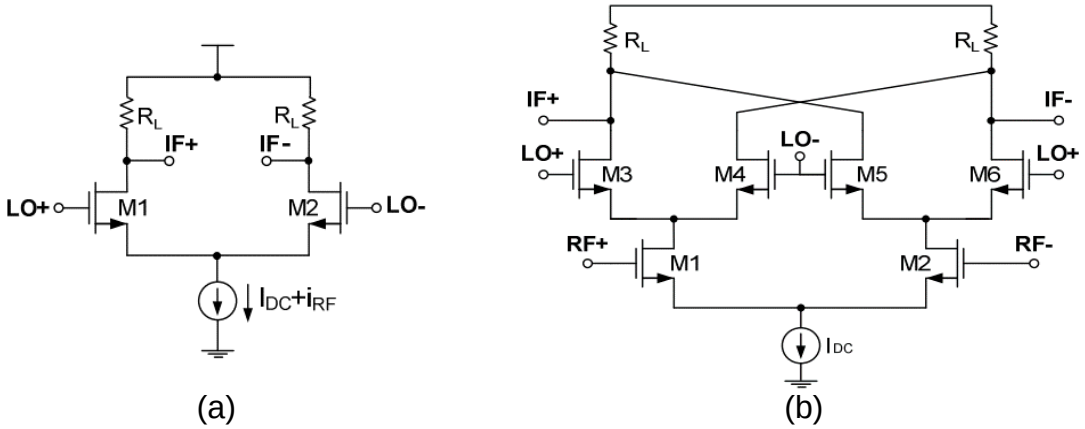


Figure II.21 : (a) Basic active mixing cell. (b) Basic Gilbert cell.

Passive mixers work in the voltage domain without any current conversion. Thus, they do not require DC power. However, they also do not provide any conversion gain, but instead, they cause losses that have to be compensated by different blocks in the system. The passive switches should have minimal ON resistance (R_{ON}) to reduce their conversion losses. However, for good isolation, a high off resistance (R_{OFF}) is required.

The classic discrete passive mixer scheme is shown in Figure II.22.a, and is called the diode double-balanced mixer. It offers high linearity over a broad range of frequencies.

The diodes in this circuit switch their state as the LO and RF signals swing between positive and negative peaks. The RF signal is routed to the IF port according to the alternating phases of the LO signal.

Thus, the RF signal is multiplied by the LO signal and appears in the resulting IF signal. This scheme usually requires a significant LO driving signal due to the high diode forward voltage drop.

Passive mixers can be implemented using MOS transistors, as shown in Figure II.22.b, and they are considered as an evolution of the diode ring mixer. Transistors, usually operating in the active region, used in passive mixers can be thought instead as resistors controlled by voltage. In active mixers, the transconductance stage limits linearity; its absence in passive mixers significantly improves linearity. The main disadvantage of passive MOS mixers is that they require a larger LO signal to turn the MOS switches ON and OFF. In this work, we choose to use the circuit shown in Figure II.22.b to upconvert the signal at the transmitter side, mainly due to its high linearity and no DC power consumption. Indeed, this scheme suffers from conversion losses; it will be studied in detail in the next chapter.

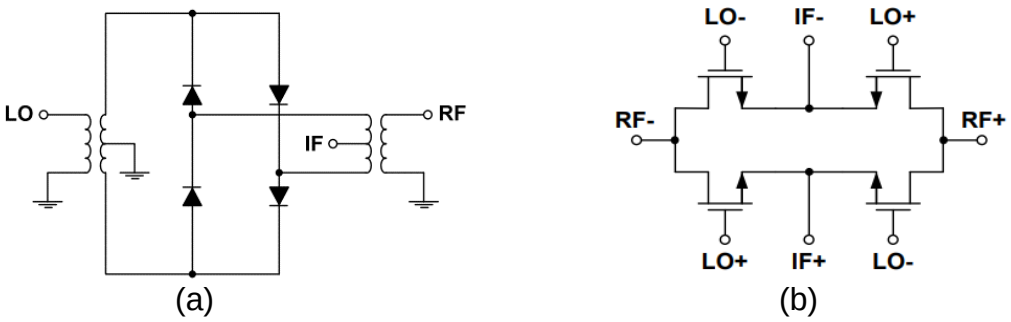


Figure II.22 : (a) double-balanced diode ring. (b) Transistor ring passive mixer.

I.4.3. The local oscillator (LO)

The signal is upconverted to a higher frequency (60GHz in our case) at the transmitter side; this is realized using a mixer. As mentioned above, the mixer requires two inputs, the low-frequency (RF) signal and the high-frequency (LO) signal. This latter is generated using a local oscillator.

Oscillators are a crucial part of any RF communication chain; they are non-linear circuits that convert DC power into an RF periodic waveform. There are two types of oscillators:

1- Sinusoidal oscillators, also known as harmonic oscillators, generate a theoretically pure sinusoidal output with a constant amplitude.
2- Non-Sinusoidal oscillators, also known as relaxation oscillators, generate other types of waveforms such as square-waves, triangular or saw-toothed waveforms.

The basic principle behind any oscillating circuit can be described using Figure II.23. It shows an amplifier with a voltage gain A and an output V_{out}. V_{out} goes through a frequency-dependent feedback network with a gain of less than one. Next, it is summed with the input V_{in}.

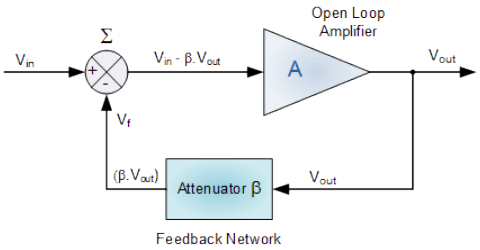


Figure II.23: Basic oscillator feedback circuit.

In this case, the output voltage is given by:

$$V_{out} = \frac{A}{1 + A\beta} V_{in} \qquad \text{(II-19)}$$

It is essential to notice that the denominator of (II-19) can be equal to zero for specific frequencies; This means that it is possible to have a non-zero output voltage for a zero input voltage. In other words, for specific frequencies, this circuit will generate a periodic output without necessarily requiring an input. For these frequencies, this circuit can be considered as an oscillator if it satisfies the Barkhausen criterias [18].

Barkhausen criterias defines the necessary conditions to ensure a sustained oscillation. They state that the loop gain must be slightly more or equal to 1, and the input and output signals should be in-phase, i.e., the overall phase shift must be 0.

Voltage-controlled oscillators (VCO) are used when the output frequency must be adjustable; in that case, a voltage can vary its output frequency. They are usually part of a phase-locked loop (PLL) made off a phase detector and a low-pass loop filter, as shown in Figure II.24. A PLL is usually used to produce a precise frequency f_{out} from an unstable frequency signal f_{in}. Using its feedback path, the PLL detects the phase difference between two signals, its reference, and its output. This information, i.e., the difference between the two signals, is converted to a voltage, also known as the error voltage signal. Next, it is used to control the VCO and match the output frequency to the desired frequency.

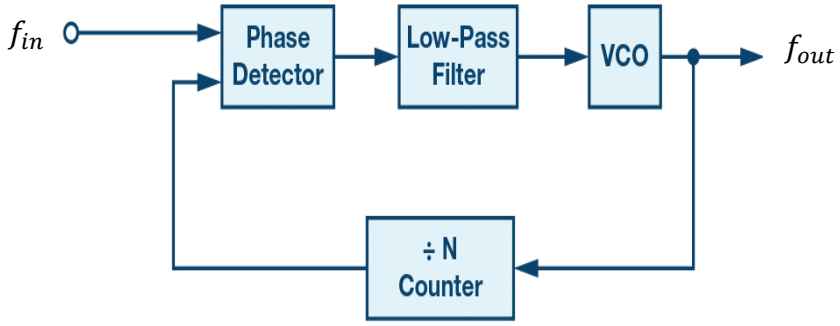


Figure II.24 : Phase-Locked loop architecture.

A VCO is characterized by several parameters, such as its power consumption, oscillation frequency, phase noise [21], and tuning range. Usually, a resonant circuit is used to precisely specify the oscillation frequency; this circuit more often consists of an LC tank (inductor and a capacitor).

Noise introduced into an oscillator may alter its output frequency and amplitude. In most cases, the effect on frequency is more important than the amplitude. Phase noise is the most crucial VCO parameter; it describes the random deviation of the output signal from its ideal position along the time axis, i.e., the jitter; however, it is usually defined in the frequency domain.

To understand the phase noise importance, we compare the output of an ideal and a "real" oscillator in the frequency domain.

In the time domain, the output of the ideal and "real" oscillator can be described as follows:

$$V_{ideal\,out} = V\cos(\omega_0 t) \tag{II-20}$$

$$V_{real\,out} = V[1 + A(t)]\cos\big(\omega_0 t + \theta(t)\big) \tag{II-21}$$

Where $\theta(t)$ and $A(t)$ represent the phase and amplitude variations of the output, respectively.

The spectral signature of phase noise is shown in Figure II.25. An ideal oscillator has a component at the center frequency, i.e., all the power is focused on this frequency. However, a "real" oscillator has sidebands instead of a single component. This shape is due to the output signal variations around the ideal positions in the time domain.

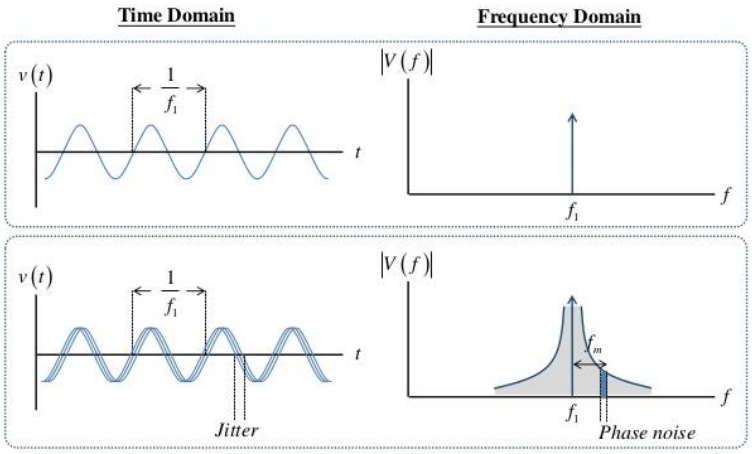


Figure II.25 : Oscillator phase noise.

When an ideal LO is convoluted with an RF signal using a mixer, it results in a non-distorted IF signal. In practice, the RF signal may be close to an adjacent channel; the signal in this adjacent channel is considered as an undesired signal or an interferer, as shown in Figure II.26. In this case, if the LO has high phase noise, as shown in Figure II.26, the downconverted signal resulting from mixing the RF signal with the LO signal results in the overlapping spectra shown in Figure II.26. The desired signal can be masked by the phase noise of the strong interferer, also known as reciprocal mixing.

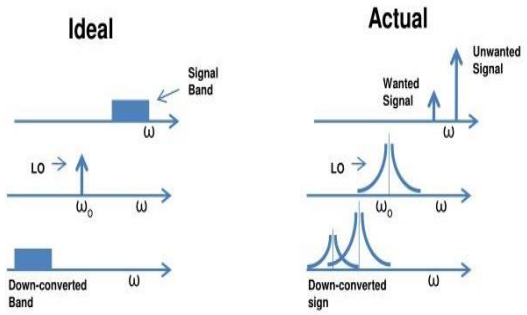


Figure II.26 : Effect of phase noise.

The remaining parameters will not be detailed hereafter; detailed studies can be found in [17], [18].

In this work, one of our objectives is to avoid generating a carrier at the receiver end. This is possible by using the duobinary passband modulation described in the previous section.

I.4.4. The demodulator

The signal at the input of the receiver is at a high frequency (60GHz precisely). Thus, it needs to be downconverted to baseband to recover the data waveforms [22]. To do so, a first approach, also known as the coherent approach, uses a mixer, i.e., the same approach used at the transmitter but to downconvert the signal. This approach requires a second LO at the same frequency (60GHz), and the same phase; implementing such LO would increase the receiver's power consumption and surface area. However, implementing a second LO can be avoided using the modulation used in this work, as introduced in section I. In addition to not requiring a second LO, the

ASK demodulators used are insensitive to small variations in the IF frequency, which reduces the constraints on the transmitter LO. Hereafter, we introduce two of the most common envelope detection techniques: the diode detector technique and the square-law demodulator.

I.4.4.1 Diode detector

The circuit shown in Figure II.27 was used in the early days of radio as a low-cost AM signal detector. It is called the diode envelope detector [23]. It provides an output proportional to the envelope of the modulated signal, as shown in Figure II.27. Its architecture is quite simple and consists of two stages. The first stage consists of a diode that enhances one half of the RF received signal. Schottky diodes with low turn-on voltage may be used for low voltage signals instead of standard junction diodes. On the positive half-cycle of the RF signal, the diode is forward biased, and thus the capacitor is charged to the peak value. On the negative half-cycle, the diode is reverse-biased, and the capacitor discharges through the load.

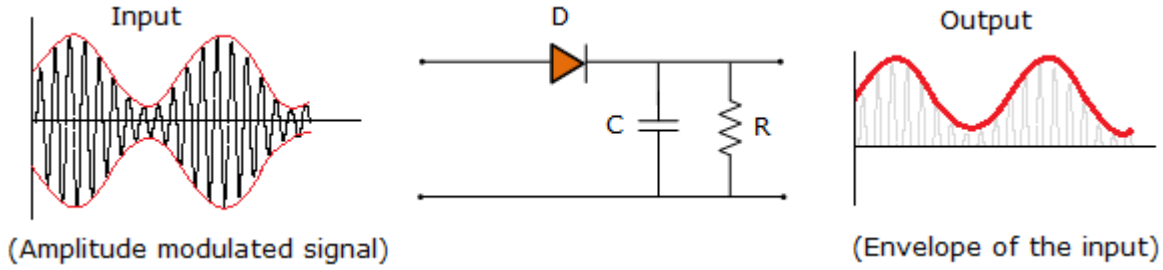


Figure II.27 : Diode envelope detection scheme.

The second stage is the low-pass filter that smooths the signal by removing the high-frequency components remaining at the diode's output. It can be implemented using a simple RC circuit. To avoid distortion, the RC values are chosen to suppress the carrier without disturbing the message bandwidth:

$$\frac{1}{f_c} \ll RC \ll \frac{1}{f_B} \tag{II-22}$$

where f_c is the carrier frequency and f_b is the message bandwidth, respectively. The simplicity and low cost of this circuit make it very interesting; however, the required turn-on voltage, even if Schottky diodes are used, makes it inconvenient for our demodulator. A more adapted circuit is proposed in the next sub-section.

I.4.4.2 Square-Law demodulator

Figure II.28 shows a block diagram of a square-law demodulator [24]. In this approach, the signal is squared before passing through a low-pass filter to remove the high-frequency components. Non-linear devices such as diodes or transistors are used for this operation, although the latter is preferred since it provides gain.

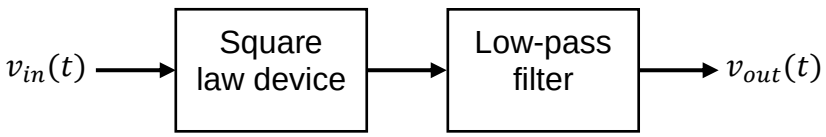


Figure II.28 : Square low demodulator.

When an input signal $v_{in}(t)$ passes through a non-linear device, it generates an output signal given by:

$$V_{out}(t) = a_1 \cdot v_{in}(t) + a_2 \cdot {v_{in}}^2(t) + a_3 \cdot {v_{in}}^3(t) + \cdots + a_n \cdot {v_{in}}^n(t) \quad \text{(II-23)}$$

where a_i are constants. If $v_{in}(t)$ is small enough, equation (II-23) can be approximated by the first two terms only:

$$V_{out}(t) = a_1 \cdot v_{in}(t) + a_2 \cdot {v_{in}}^2(t) \quad \text{(II-24)}$$

Such a device has a square-law behavior. Let us consider the amplitude modulated signal shown in Figure II.29; this modulated AM signal is given by:

$$v_{in}(t) = [A_c + m(t)]\cos(\omega_c t) \quad \text{(II-25)}$$

where A_c is the carrier signal amplitude and $m(t)$ is the modulating signal (can be digital or analog). For simplicity, we consider here $m(t)$ as a sinusoid given by $A_m \cos(\omega_m t)$.

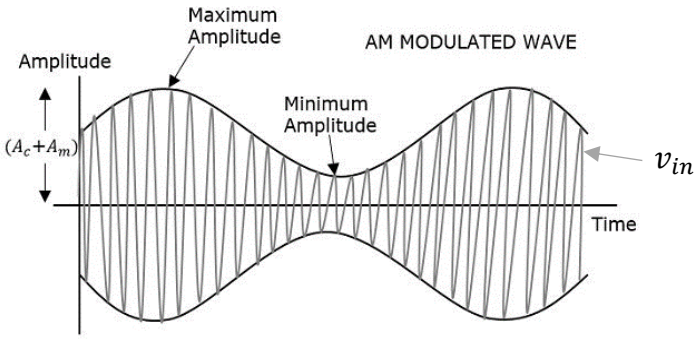


Figure II.29 : AM modulated signal.

The level of AM modulation, also known as the modulation index, is defined as the ratio of A_m, the modulating signal and A_c, the level of the unmodulated carrier. This index can also be given in a percentage called the modulation percentage by multiplying by 100. We say that an AM signal is 100% modulated if $m(t)$ has a positive peak value of $+1$ and a negative peak value of -1.

For an input signal $v_{in}(t)$, the output of the non-linear device using (II-24) can be computed as follow:

$$v_{out}(t) = a_1 \cdot [A_c + m(t)]\cos(\omega_c t) + a_2 \cdot [(A_c + m(t))\cos(\omega_c t)]^2 \quad \text{(II-26)}$$

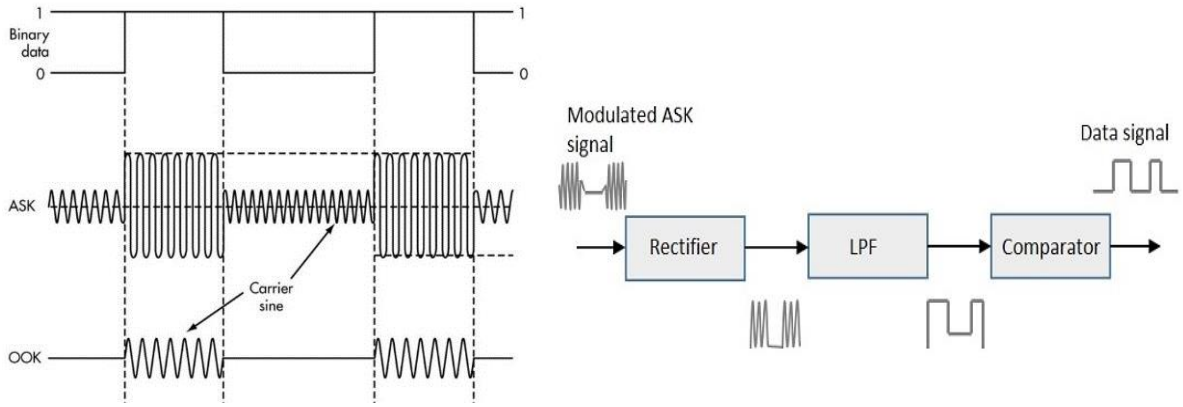
$$v_{out}(t) = a_1 \cdot A_c \cos(\omega_c t) + a_1 \cdot m(t)\cos(\omega_c t) + a_2 \cdot \left[\left(A_c^2 + 2A_c m(t) + m^2(t)\right) \cdot \left(\frac{1 + \cos(2\omega_c t)}{2}\right)\right] \quad \text{(II-27)}$$

$$v_{out}(t) = a_1 \cdot A_c \cos(\omega_c t) + a_1 \cdot m(t)\cos(\omega_c t) + a_2 \cdot \left[\frac{A_c^2}{2} + A_c m(t) + \frac{m^2(t)}{2} + \frac{A_c^2}{2}\cos(2\omega_c t) + A_c m(t)\cos(2\omega_c t) + \frac{m^2(t)}{2}\cos(2\omega_c t)\right] \quad \text{(II-28)}$$

The resulting signal described in equation (II-28) contains several frequency components at ω_m, ω_c, $2\omega_m$, $2\omega_c$, and other harmonics. We notice that $a_2 A_c m(t)$ is a

scaled version of the desired signal $m(t)$. The extra frequency components should be filtered to recover $m(t)$ correctly. A low-pass filter can eliminate all the unwanted high-frequency components by choosing a carrier frequency bigger than the modulating signal ($\omega_c \gg \omega_m$) . The DC term, however, remains.

This reasoning is identical whether the modulating signal is analog (sinusoidal) or digital (a square binary signal, for example). In the latter case, this modulation is known as Amplitude Shift Keying (ASK) modulation and is shown in Figure II.30.a. In this modulation, the transmitter transmits a significant RF signal for a binary $'1'$ and a smaller RF signal for a binary $'0'$. ASK modulation in its simplest form is known as On-Off Keying (OOK) modulation, the presence of a carrier represents a binary '1', and its absence signifies a binary $'0'$. This modulation is used for various wireless communication systems due to its simplicity and the simplicity of the transmitters.



(a) (b)

Figure II.30 : (a) ASK and OOK modulations. (b) ASK demodulator.

I.5. The channel: transmission lines

The channel connects the transmitter to the receiver; a well-designed channel should not severely degrade the signal in terms of distortion. Most of the on-chip channels are made of a conductive metal and are realized following planar processes; their shape is defined using masks, photoresists, and metal etching. We introduced RC wires in chapter I; they are the most common type of interconnections in chips nowadays. However, as we mentioned, these wires have some drawbacks for long distances, such as delay and severe distortion.

Transmission lines are a better fit for long-distance interconnections because they can carry signals at much higher speeds. Due to the necessary control of the characteristic impedance, transmission lines are much larger than simple RC wires and affect the interconnects density. Some considerations are to be respected to choose the appropriate design or transmission lines configuration, such as:

1- The transmission lines characteristic impedance, delay, and group delay must be well-defined over the band of interest.
2- The strips should be shielded from other strips to avoid crosstalk.
3- An appropriate well-defined low-ohmic ground should be defined to avoid forming and picking up signals in the substrate.
4- The transmission lines attenuation should be kept low and "constant" over the band of interest to avoid distortion.

5- The crossing of other strips over a certain distance should be acceptable and not significantly impact the transmission. For example, the strips can be implemented in the highest metal layer, their ground in a middle layer, and thus other strips can be used in the lowest metal layers.

In the next sub-sections, we discuss some of the most commonly used on-chip transmission line topologies; each has its advantages and disadvantages; a compromise may be required to choose the appropriate topology.

I.5.1. Microstrip

The microstrip line is the most common transmission line configuration; it is also the simplest from a layout point of view. In a microstrip line, the current flows in both conductors (top and bottom), the bottom conductor is the return plane.

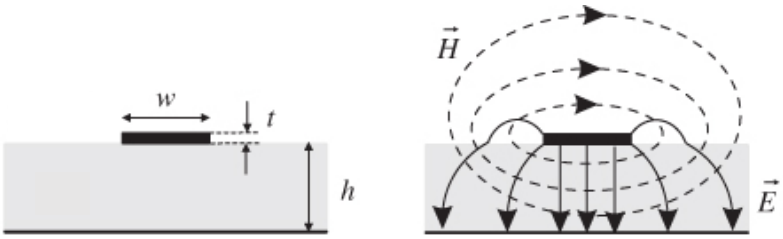


Figure II.31 : (a) Microstrip line principle. (b) Electric and magnetic fields.

Figure II.31.a gives the microstrip line principle, and Figure II.31.b shows the electric and magnetic fields. The configuration comprises two different dielectric mediums, also known as an inhomogeneous medium, the electric field being in the air and in the dielectric between the air and ground plane. It is worth mentioning that most of the electric field is constrained within the substrate, and only a small part of the energy is in the air. However, this influences the propagation mode along the microstrip; we can distinguish four modes [18]:

1. Transverse EM (TEM) waves: the electric and magnetic fields of these waves are always perpendicular to the direction of wave propagation and perpendicular between them.
2. Transverse magnetic (TM) mode: only the electric field vector is directed along the direction of propagation, the magnetic field vector is perpendicular to it.
3. Transverse electric (TE) mode: only the magnetic field vector is directed along the direction of propagation, the electric field vector is perpendicular to it.
4. Quasi-TEM mode, which is the TEM mode considering moderate losses (i.e., propagation constant much higher than the attenuation constant).

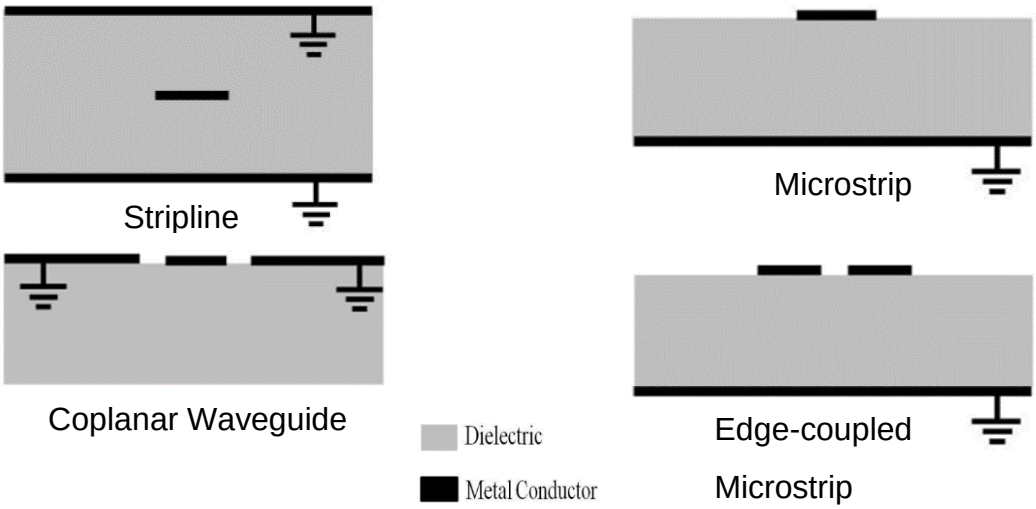


Figure II.32 : Main transmission lines topologies.

Microstrip lines have a dominant Quasi-TEM mode due to their two-medium dielectrics; as a result, its characteristics, such as characteristic impedance and phase velocity, are slightly frequency-dependent, making microstrip lines a little bit dispersive. A more detailed analysis can be found in [18]. The phase velocity of microstrip lines can be approximated by:

$$V_p = \frac{c}{\sqrt{\varepsilon_{reff}}} \tag{II-29}$$

Where ε_{reff} is the relative effective dielectric constant. We speak about relative effective dielectric constant ε_{reff} instead of relative dielectric constant ε_r when the electric field propagates in several mediums with different ε_r. In this case, it takes into account both of the air and the substrate relative dielectrics constant; it is higher than the dielectric constant of air $\varepsilon_r = 1$ and lower than the substrate dielectric constant. At higher frequencies, a larger part of the electric field is confined in the substrate, resulting in an increase of the relative effective dielectric constant approaching the relative substrate dielectric constant. To calculate ε_{reff}, the approximations given in [18] can be used. Choosing which approximation to use depends on the ratio of width to height of the microstrip line configuration.

The microstrip line characteristic impedance can be approximated by [18] :

$$Z_0 = \begin{cases} \dfrac{60}{\sqrt{\varepsilon_{eff}}} \ln\left(\dfrac{8h}{W} + \dfrac{W}{4h}\right) & W/h \leq 1 \\ \sqrt{\varepsilon_{eff}}\left[\dfrac{W}{h} + 1.393 + 0.667 \ln\left(\dfrac{W}{h} + 1.444\right)\right] & W/h \geq 1 \end{cases} \tag{II-30}$$

with W the width of the conductor strip, and h the dielectric substrate thickness. The microstrip line dielectric losses are given by the attenuation constant :

$$\alpha_d = \frac{\pi f \varepsilon_r (\varepsilon_{eff} - 1) \tan\delta}{c\sqrt{\varepsilon_{eff}}\,(\varepsilon_r - 1)} \; Np/m \tag{II-31}$$

with $\tan\delta$ the loss tangent of the dielectric. However, in CMOS technology, the majority of the microstrip line losses are metal losses due to the finite conductivity of the conductor and its small height, and they dominate the dielectric and radiation losses. High accuracy metal losses calculation usually requires EM simulation, however a good approximation is given by:

$$\alpha_c = \frac{R}{2 \cdot Z_c} \ Np/m \tag{II-32}$$

where R is the resistance per meter, and Z_c the characteristic impedance. The conductor losses are higher for narrow lines.

The relative effective dielectric constant of a microstrip line varies with frequency; this results in a non-linear variation of the propagation constant (except for the distortionless transmission line case in Chapter I). Most of the microstrip line parameters vary with the frequency, such as phase velocity or losses, resulting in a limitation of signal transmission bandwidths. Signal dispersion is one example where the variation of phase velocity over a certain bandwidth means that the frequency components propagating over this band propagate at different speeds and thus arrive at different times. Another factor is the current distribution along the conductor, as mentioned previously in Chapter I, attenuation varies significantly with frequency due to skin effect. Microstrip line EM simulation results in FDSOI 28-nm technology will be shown in the next chapter.

I.5.2. Stripline

The stripline shown in Figure II.32 is made from a central conductor submerged in a uniform dielectric; it is also limited from the top and bottom by ground planes. Contrarily to the microstrip line, its medium is homogeneous, and thus the relative dielectric constant defines the wave velocity, and the wave propagates in a TEM mode. Compared to microstrip lines, striplines are less prone to dispersion. Furthermore, striplines offer better isolation from adjacent strips, especially if vias connecting the grounds are placed along the strips. However, the apparent disadvantage is that it requires three metal layers instead of two for microstrip lines. In CMOS technology, striplines cannot be used because middle metal layers are too thin, which would result in high attenuation constant.

I.5.3. Coplanar Waveguide (CPW)

Microstrip line requires a large ground plane below the signal, and interconnects cannot be placed closeby since they impact the propagation characteristics, and crosstalk occurs. This disadvantage affects the density of the interconnections.

On the other hand, CPWs shown in Figure II.32 do not suffer from the same disadvantage since they have two ground planes closeby on the same metal layer. They provide a strong return path and isolation between adjacent strips. Furthermore, an extra ground plane can be placed below so that the electric field does not penetrate the substrate. The side grounds are connected to the ground plane below periodically, connecting them every tenth of a wavelength is a good rule of thumb [25]. The drawback of CPWs is linked to their wide surface.

I.5.4. Coplanar strips

The previous geometries are single-ended, with clearly defined signal strip and ground strips. Coplanar strips are particular in the sense that no signal or ground strips are specifically defined. They are formed of two identical adjacent parallel strips on the same metal layer. Hence it can be considered that only a differential signal can propagate, without common mode propagation, leading to excellent noise immunity. They are not or less affected by outside factors such as ground bounce or crosstalk from adjacent strips.

In our work, the edge coupled microstrips shown in Figure II.32, are used to transmit the signal differentially; the implementation and simulation results in 28nm FD-SOI technology will be shown in the next chapter.

II. Architecture simulations

II.1. System architecture simulations results

The designed system in this work includes a transmitter, a channel, and a receiver. The signal will be modulated then upconverted in the transmitter using the IF technique introduced previously. Next, it is transmitted through the channel and demodulated at the receiver side. Keeping in mind that our main objectives are high data rates (>10Gbps) and good signal integrity (BER< 10^{-12}), we study the system to identify the components' crucial parameters. Figure II.33 shows a first block representation of our system that we will develop hereinafter.

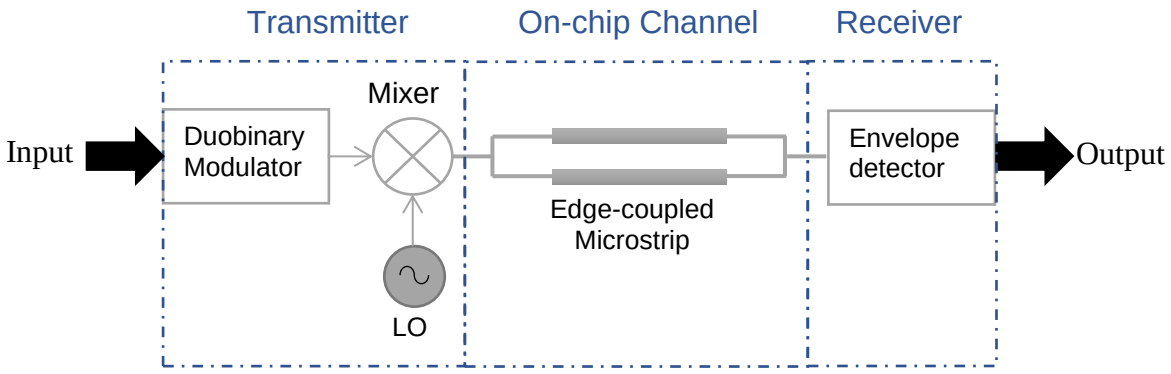


Figure II.33: Block diagram representation of the proposed transceiver.

Before implementing the system in 28-nm FD-SOI technology, we studied its specifications to check which parameters are the most crucial for proper functioning. The objective is to design a low BER robust system (lower than 10^{-12}),i.e., with high SNR (at least 18dB) and a controlled jitter. The acceptable jitter value depends on the data rate and the horizontal eye closing and will be seen in the design section. We used Advanced Design System (ADS) to model the transmitter and receiver; we sweep the components parameters to identify the bottlenecks. This modelization is useful for determining the system's gain, determining the acceptable noise level, and the appropriate receiver bandwidth. As for the channel, herein, we use 5mm microstrips for the single-ended approach and edge-coupled microstrips channel for differential

signaling; these channels were simulated using ADS FEM software; the flow will be described in the next chapter.

Although single-ended microstrips require less surface on-chip, we prefer differential signaling due to its noise immunity. The modified duobinary modulation can be used for single-ended signaling too. The system architecture is shown in Figure II.34. The first step of the duobinary modulator, i.e., the differential precoder, is not shown here and was implemented using numerical equations for simplicity. Only the second part is shown; its input called 'Precoded_binary_input" is the output of the differential precoder.

The duobinary reshaping filter was implemented using a power splitter, an adder, and a time delay (100ps in the figure for 10Gbps signal data rate). Some of the components used at this level of simulation are "ideal". The duobinary reshaping filter input and output time-domain waveforms are shown in Figure II.35 along with its frequency-domain spectrum.

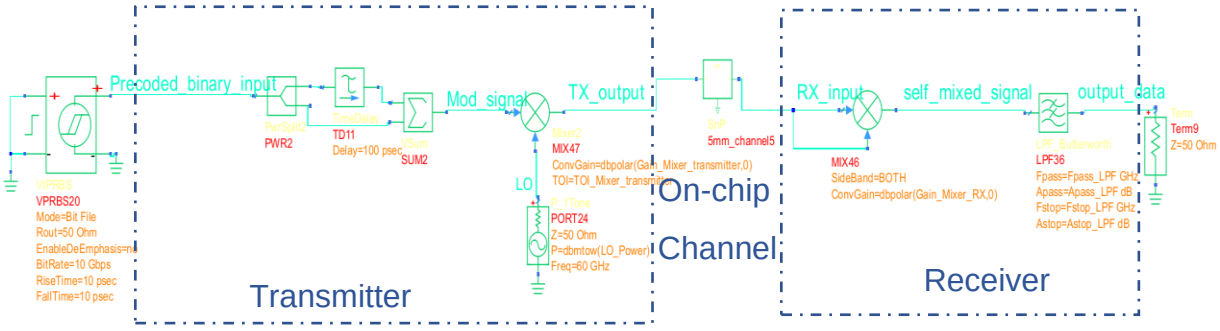


Figure II.34 : Single-ended architecture modeling for ADS simulation.

Next, the mixer is modeled, and its parameters (Gain, IIP3, leakage ....), described in the next section, can be varied. The mixer takes two inputs: the modulating signal called "Mod_signal", which consists of a three-level signal, as shown in Figure II.35, and the 60GHz carrier called "LO" in the figure. The LO power level and phase noise will be varied to check the system robustness at this simulation level.

The “LO” 60GHz carrier is mixed with the 10 Gbps signal (“Mod_signal”). This results in a signal, shown as “TX_output” in Figure II.35 and Figure II.36, with three states: 0, $Acos(2\pi f_c t)$ and $-Acos(2\pi f_c t)$.

The transmitter output “TX_output” spectrum is concentrated around the carrier frequency (60GHz), as shown in Figure II.37. We notice that the first null is only 5GHz away from the carrier frequency (60GHz) for a 10Gbps data signal. An ASK modulation with the same data rate requires larger BW. The signal is attenuated in the 5mm channel, as shown in Figure II.37; in this case, the 5mm channel has around $6dB$ loss at 60GHz and an equal Noise Figure (NF). The attenuated signal called “RX_input” is self-mixed, i.e., multiplied by itself to get a squared signal. The resulting signal is a 10Gbps signal with a larger BW of 10GHz. We notice from the time-domain signal and the frequency-domain spectrum of the “self_mixed_signal” in Figure II.37, that the resulting signal has a large amplitude at 120GHz ($2f_c$). Thus, a low-pass filter is necessary to remove the high-frequency components and recover the square BB signal shown in Figure II.37. Figure II.38 shows a “clean” eye diagram of the output signal; the signal is minimally distorted ($SNR \approx 28\,dB$). The jitter will be evaluated accurately in the next chapter.

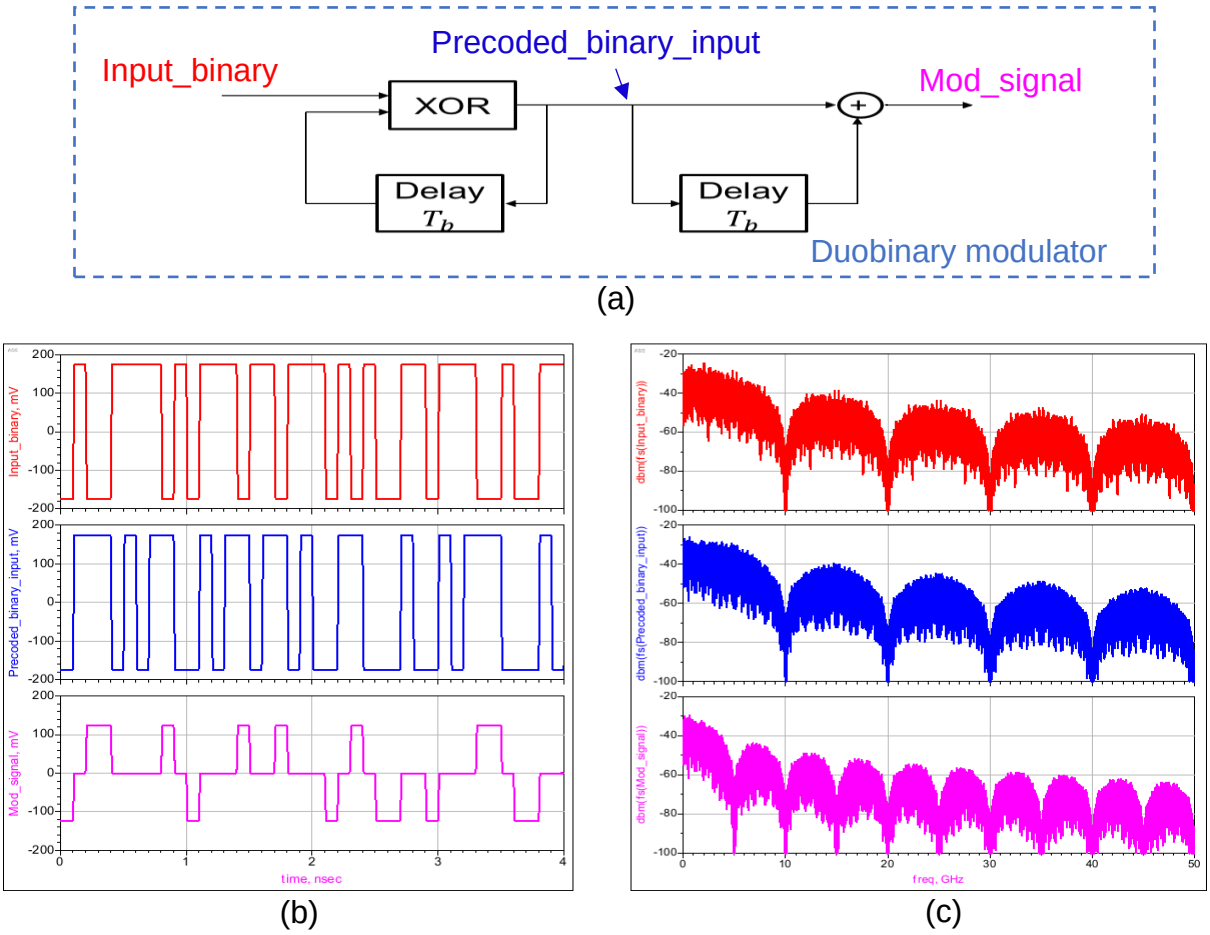


Figure II.35 : (a) Duobinary architecture. (b) Duobinary time-domain waveforms. (c) Duobinary frequency-domain spectrum.

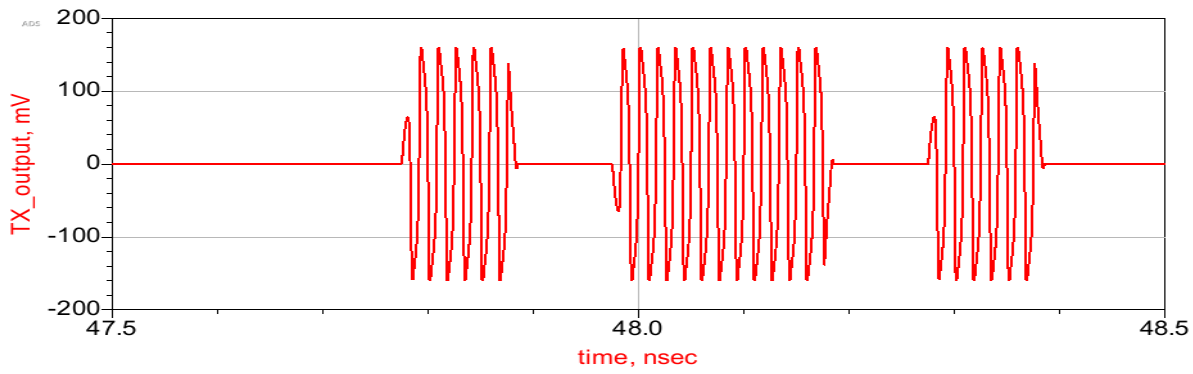


Figure II.36 : The transmitter modulated output waveform.

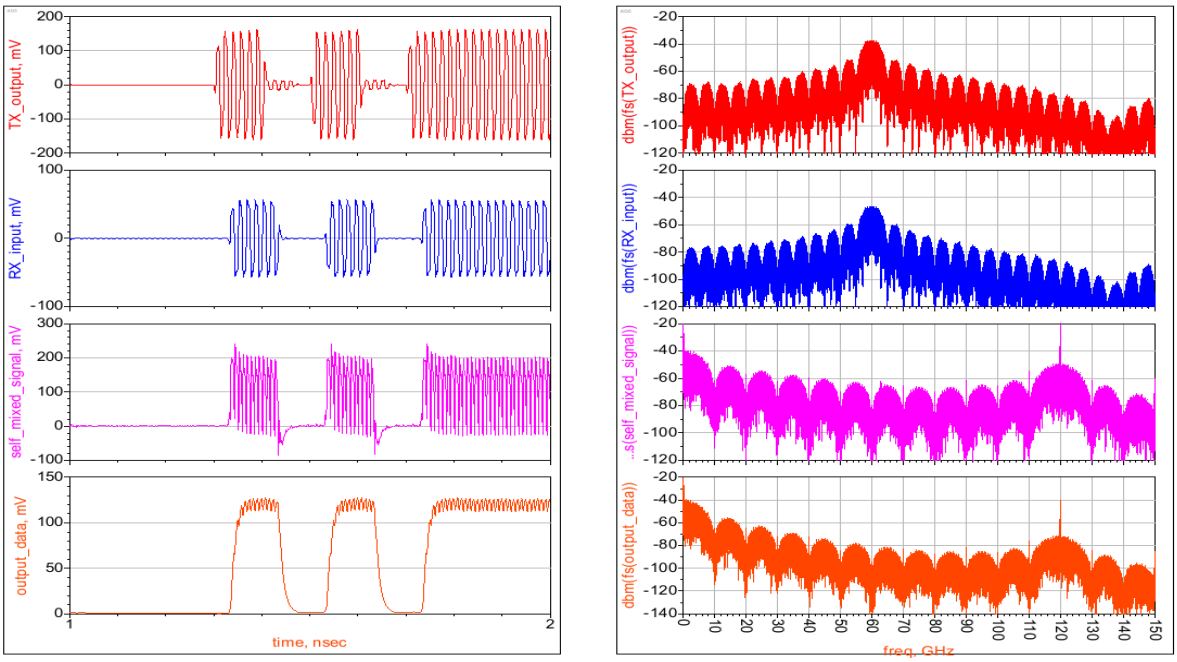


Figure II.37 : Demodulation time-domain waveforms and frequency-domain spectrum.

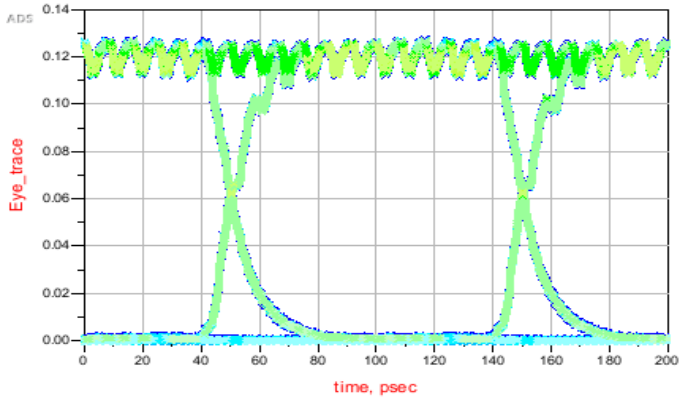


Figure II.38 : Demodulated 10Gbps signal eye diagram.

II.2. Length adaptive serial link

To make this approach attractive for large digital ICs links, we studied the link's length adaptivity. In other words, we studied the possibility of using the same transceiver for different distances without redesign. For different link lengths, several parameters vary. For a matched transmission line, the most important ones are the link's attenuation, the propagation delay, and its noise figure (equal to its attenuation, as explained below). These factors impact the SNR and, consequently, the BER of the system.

One of the main advantages of this method is the linear propagation delay. As mentioned in chapter 1, the delay is linear when transmitting in the LC region of the wire instead of the RC region, where the delay is proportional to the length squared. The delay is also much lower in the LC region.

The linearity of the delay makes it easily predictable and anticipated in the floorplan. We assume that the delay in a transmission line designed in a SiO_2 substrate

BEOL ($\varepsilon_{reff} \approx 4$) propagates at half the light's speed, resulting in a delay of around 6-7ps/mm at 60Gz. The next chapter's simulation results using 28-nm FDSOI technology confirms this assumption. Consequently, a 5mm length is crossed in 35ps while a 15mm is crossed in 105ps. Of course, the difference in delay due to the distance can be corrected with an appropriate sampling synchronization.

The second parameter is the transmission line attenuation constant that leads to insertion loss (in dB), and it can be related linearly to the distance (dB/mm). If a transmission line is designed to have an attenuation of 1dB/mm at 60GHz, extending it to a 5mm results in 5dB attenuation, and a 15mm long link results in 15dB attenuation. Because the losses are predictable, compensating them using a variable gain amplifier (VGA) is more straightforward.

The third parameter is the noise factor of the transmission line. At standard temperature ($T = 290°K$), the noise factor of a passive device is related to its noise temperature T_e, and is given by:

$$F = 1 + \frac{T_e}{T} = 1 + \frac{(A-1)\cdot T}{T} = A \tag{II-33}$$

where A is the attenuation of the transmission line. For simplification reasons, we consider a simple cascaded system of a transmitter (Gain G_1 and noise factor F_1) and a transmission line with an attenuation (A_2). The system noise factor can be calculated using the Friis formula given by:

$$F_{casc} = F_1 + \frac{(F_2 - 1)}{G_1} \tag{II-34}$$

This allows the designer to predict the system's achievable SNR and the maximum transmission line length leading to an acceptable SNR. The maximum transmission line length can also be outlined using the budget link analysis, as described in the next sub-section.

II.2.1. Budget Link analysis

To develop the budget link analysis, we suppose that the transmitter delivers an output power of $-5dBm$. Figure II.39 shows an example of two transmission lines with different attenuation values. The first is 5mm long and has an attenuation of 5dB; the second is 15mm long with a 15dB attenuation. For simplicity, we assume these attenuation values are given at 60GHz. We also consider that the receivers' gain is 3dB, and its noise figure is 6dB.

We calculate the noise power over a 10GHz bandwidth (B) using the equation:

$$N = BTk_B \approx -74\ dBm \tag{II-35}$$

with k_B the Boltzmann constant. This value represents the theoretical noise floor for an ideal receiver. In practice, the NF of the receiver impacts the receiver actual noise floor, and the output noise power is given by:

$$N_{RX} = N + NF_{RX} \approx -68\ dBm \tag{II-36}$$

In this ideal case shown in Figure II.39, we find that the margin is very large ($-68dBm + 17dBm = 51dB$) for the 15mm transmission line. In other words, changing the length of the transmission line impacts the attenuation and the noise, of course, but does not constitute a bottleneck for the SNR and the BER.

These high margins also mean that higher data rates (higher than 10 Gbps) could probably be reached. However, as we will see, this theoretical SNR can degrade very quickly when non-linearity and jitter appear. Figure II.40 shows the output signal after 5mm and 15 mm long transmission lines, respectively, for the same system parameters. The SNR degraded from 35dB for 5mm transmission to 28dB for a 15mm transmission line, respectively.

Finally, if necessary, the attenuation can still be compensated using a VGA, as shown in Figure II.41.

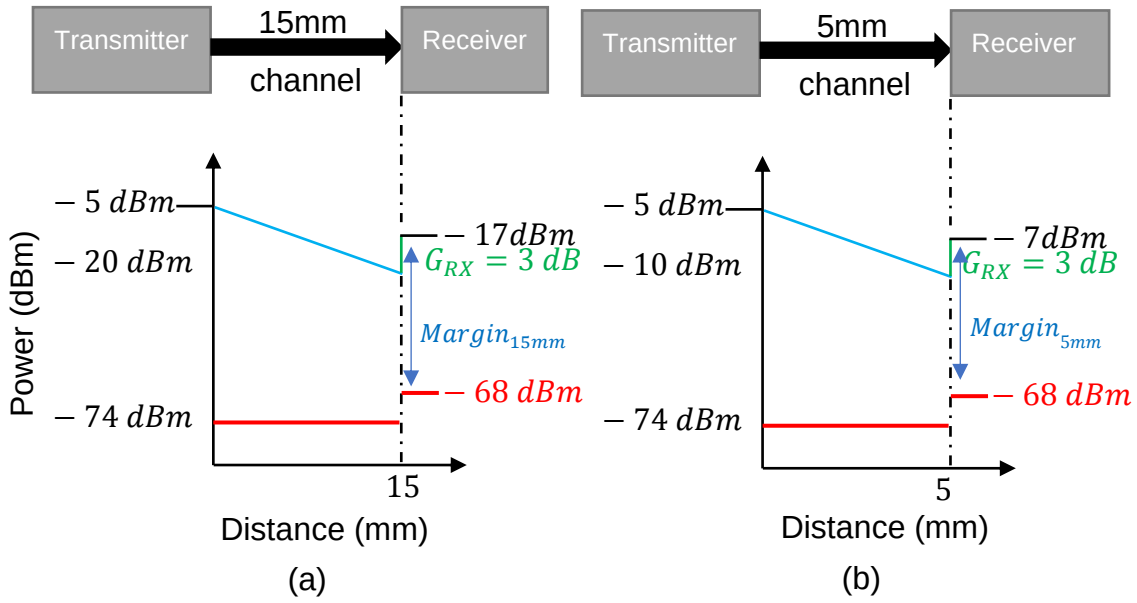


Figure II.39: Budget link analysis. (a) for a 15mm channel. (b) for a 5mm channel.

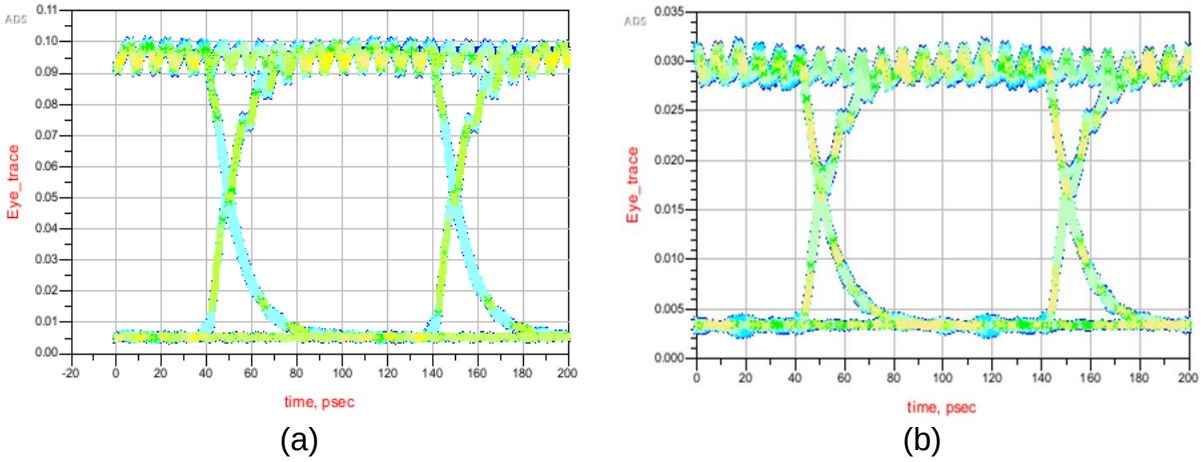


Figure II.40 : Output signal eye diagram. (a) after a 5-mm transmission line. (b) after a 15-mm transmission line.

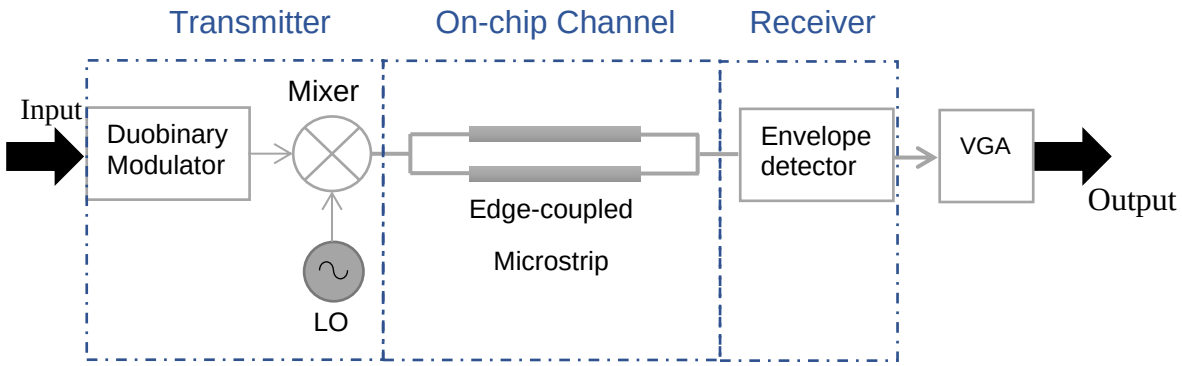


Figure II.41 : The proposed architecture for length adaptive serial link.

II.3. Impact of some system components on the signal integrity

The BER of the system can degrade very quickly for several reasons, such as non-linearity of the mixers, noise figure, impedance mismatch... The reasons are numerous. Hereafter, we evaluate the effect of some of them.

II.3.1. LO leakage impact on the SNR

The first example shown in Figure II.42 represents the effect of the LO-RF isolation in the transmitter mixer. It measures the amount of power leaking from the LO port (60GHz carrier in this case) to the RF port (10Gbps data signal). Higher isolation indicates lower port leakage. The LO-RF isolation is varied from -60dB to -15dB, and the SNR varies from 28dB down to 5dB. For appropriate functioning, we conclude that an isolation of -25dB is necessary.

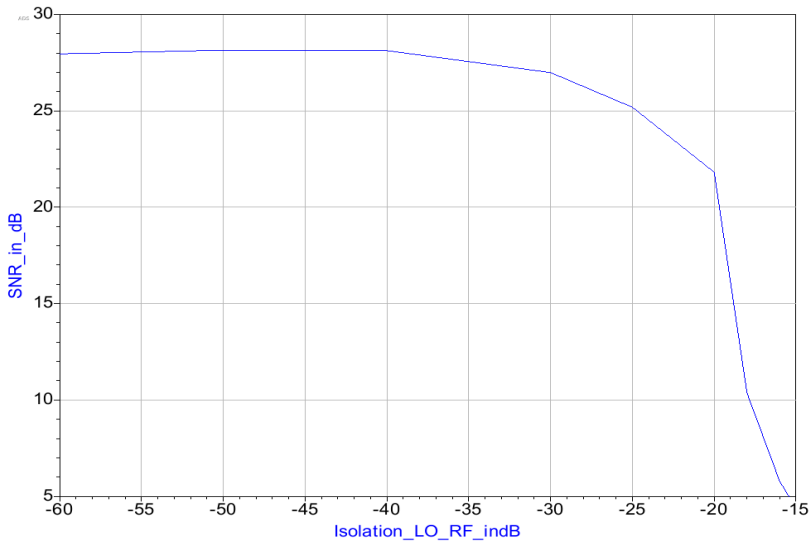


Figure II.42 : Impact of the transmitter mixer LO-RF isolation on the SNR.

II.3.2. Impact of the receiver low-pass filter on the signal integrity

Figure II.44 shows the effect of the low-pass filter cut-off frequency on the signal integrity. For this purpose, we vary the 3-dB cut-off frequency of the 1st order Butterworth low-pass filter in the systems model, and we notice that the SNR is maximal around 10GHz, as shown in Figure II.43. It is worth mentioning that for this

simulation, the jitter injected into the system is minimal. The jitter will be studied in detail in the implementation step. Figure II.44 shows the time-domain waveforms and the eye diagrams for the different cut-off frequencies. We notice that for the $f_{3dB} = 5\,GHz$ case, the signal is distorted with longer rise/fall time, contrarily to the $f_{3dB} = 20\,GHz$ where the rise/fall time is shorter. However, the eye diagram shows the significant presence of the 120GHz component shown as amplitude noise (oscillations) in level one of the eye diagrams. This explains the decrease of SNR after $f_{3dB} = 12\,GHz$, where a 1st order filter with higher cut-off frequency attenuates less the 120GHz component.

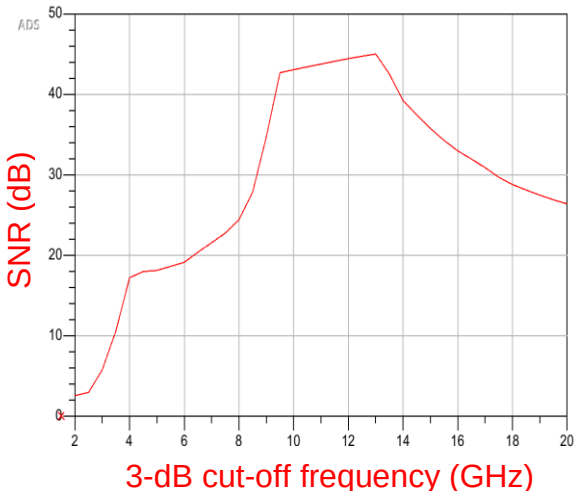


Figure II.43 : Effect of the low-pass filter 3-dB cut-off frequency on the SNR.

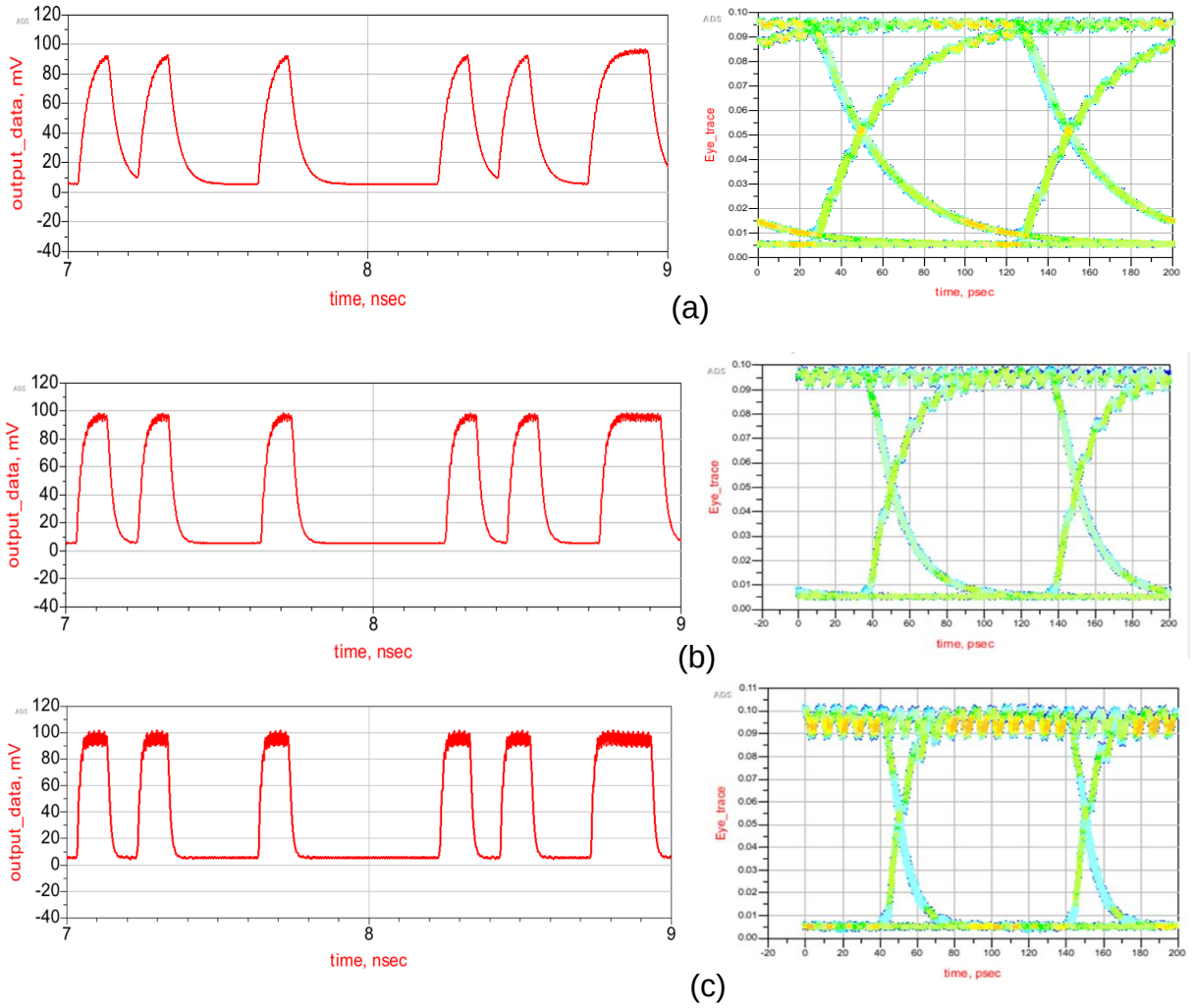


Figure II.44 : Time-domain waveform and eye diagram of the output signal for different low-pass filter 3-dB cut-off frequencies. (a) $f_{3dB} = 5\ GHz$. (b) $f_{3dB} = 10\ GHz$. (c) $f_{3dB} = 20\ GHz$.

II.3.3. Injected jitter impact on the BER

Another critical factor in BER degradation is the jitter injected into the system, either random or periodic. Herein, we inject random jitter ($RJ_{rms} = 4\ ps$) and a Periodic jitter ($PJ_{freq} = 1\ GHz$). We notice the degradation in the eye diagram of the noisy signal compared to the eye diagram without jitter. The jitter bathtub curves are shown in Figure II.45; they represent the horizontal eye closure. We notice that the eye closes "faster" when jitter is injected. The maximum acceptable value of jitter is determined during the design step in chapter III.

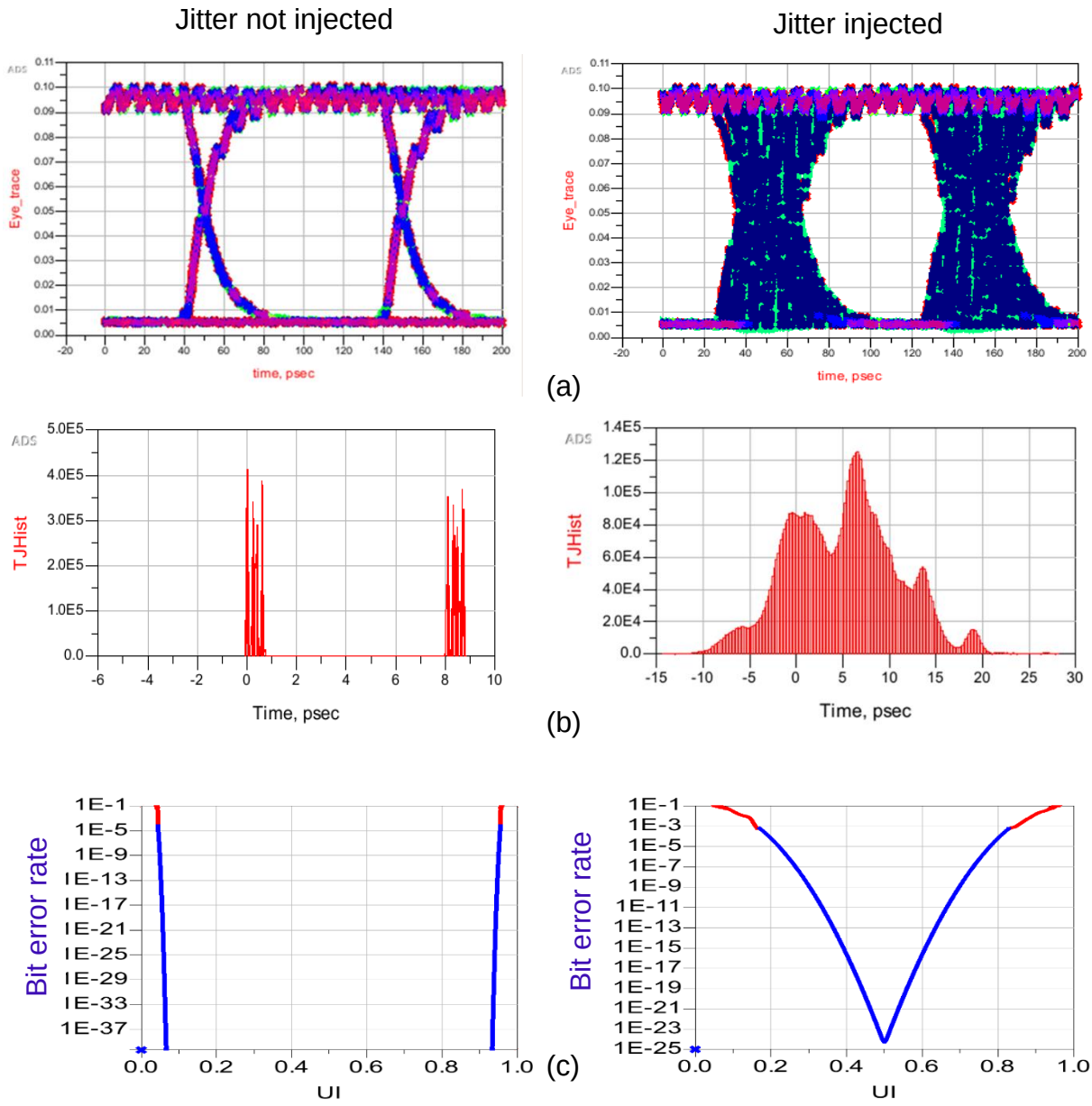


Figure II.45 : (a) eye diagram. (b) Total jitter histogram. (c) Jitter bathtub curves.

III. Conclusion

In this chapter, we introduced several serial data transmission functionalities. First, amplitude modulations are described as well as the chosen duobinary modulation. We showed that the duobinary modulation compresses the spectrum by a factor of two. Higher compression factors are achievable with higher-order (more complex) polybinary modulations. Next, we introduced the Intermediate Frequency technique, which uses a carrier at a high frequency to transmit the signal in the transmission line's LC region. The advantages of this technique are numerous. Firstly, it allows transmission in a region of the transmission line where the delay is minimal (7ps/mm) and, more importantly, linear. Furthermore, the distortion due to the channel is minimal.

Next, we introduced the proposed system architecture blocks (transmitter, channel, and receiver) and some of the used components' important parameters. In section II.2, we studied the length adaptivity of this solution. We concluded that the same transmitter and receiver could be used for a range of distances by adding a variable gain amplifier stage to the receiver when needed. This advantage is significant since it makes its implementation in digital circuits more straightforward and attractive.

Finally, the impact of some of the system's components was briefly analyzed. These "theoretical" simulations allowed us to explore the system's components and parameters to find the most critical ones.

In the next chapter, we detail the 10Gbps prototype chip design containing the transmitter, the channel, and the receiver. In this work, we implemented the transmitter and receiver in FD-SOI 28-nm technology; the following chapter shows the real implementation limitations due to the design or the technology. We will show how we implemented the different components shown in this chapter (besides the LO). We also show that the BER was lower than 10^{-12} with a high margin in SNR and Jitter in post-layout simulations.

IV. References

[1] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[2] E. Bogatin, *Signal and power integrity–simplified*. Pearson Education, 2010.

[3] B. Sklar and others, *Digital communications: fundamentals and applications*. 2001.

[4] L. W. Couch, *Digital & Analog Communication Systems*. Pearson Higher Ed, 2012.

[5] intel, “AN 835: PAM4 Signaling Fundamentals.” [Online]. Available: https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/an/an835.pdf.

[6] A. Lender, “Correlative level coding for binary-data transmission,” *IEEE Spectr.*, vol. 3, no. 2, pp. 104–115, 1966.

[7] S. Pasupathy, “Correlative coding: A bandwidth-efficient signaling scheme,” *IEEE Commun. Soc. Mag.*, vol. 15, no. 4, pp. 4–11, 1977.

[8] A. Lender, “The duobinary technique for high-speed data transmission,” *IEEE Trans. Commun. Electron.*, vol. 82, no. 2, pp. 214–218, 1963.

[9] R. Howson, “An Analysis of the Capabilities of Polybinary Data Transmission,” *IEEE Trans. Commun. Technol.*, vol. 13, no. 3, pp. 312–319, 1965.

[10] J. V. Kerrebrouck *et al.*, “100 Gb/s Serial Transmission over Copper using Duo-binary Signaling,” 2016.

[11] J. Lee, M.-S. Chen, and H.-D. Wang, “Design and Comparison of Three 20-Gb/s Backplane Transceivers for Duobinary, PAM4, and NRZ Data,” vol. 43, no. 9, p. 14, 2008.

[12] B. Min, K. Lee, and S. Palermo, “A 20 Gb/s triple-mode (PAM-2, PAM-4, and duobinary) transmitter,” *Microelectron. J.*, vol. 43, no. 10, pp. 687–696, 2012.

[13] J. H. Sinsky, A. Adamiecki, and M. Duelk, “10-Gb/s electrical backplane transmission using duobinary signaling,” in *2004 IEEE MTT-S International Microwave Symposium Digest (IEEE Cat. No. 04CH37535)*, 2004, vol. 1, pp. 109–112.

[14] J. H. Sinsky, M. Duelk, and A. Adamiecki, “High-speed electrical backplane transmission using duobinary signaling,” *IEEE Trans. Microw. Theory Tech.*, vol. 53, no. 1, pp. 152–160, 2005.

[15] F. Voineau, “Systèmes communicants haut-débit et bas coûts par guide d’ondes en plastique,” PhD Thesis, 2018.

[16] U. Alvarado, G. Bistué, and I. Adín, “Mixer Design,” in *Low Power RF Circuit Design in Standard CMOS Technology*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 129–177.

[17] R. Caverly, *CMOS RFIC design principles*. Artech House Boston, 2007.

[18] D. M. Pozar, “Microwave Engineering,” *Fourth Ed. Univ. Mass. Amherst John Wiley Sons Inc*, pp. 26–30, 2012.

[19] “Optimizing IP3 and ACPR Measurements.” http://www.ni.com/pdf/en/Optimizing_IP3_and_ACPR_Measurements_With_the_PXIe_5668R.pdf.

[20] “What’s The Difference Between The Third-Order Intercept And The 1-dB Compression Points?” https://www.electronicdesign.com/resources/whats-the-difference-between/article/21799714/whats-the-difference-between-the-thirdorder-intercept-and-the-1db-compression-points.

[21] B. Razavi, “A study of phase noise in CMOS oscillators,” *IEEE J. Solid-State Circuits*, vol. 31, no. 3, pp. 331–343, 1996.

[22] R. K. R. Yarlagadda, “Analog Modulation,” in *Analog and Digital Signals and Systems*, Boston, MA: Springer US, 2010, pp. 429–487.
[23] *Analog communication*. Technical Publications.
[24] S. A. Maas, *Nonlinear Microwave and RF Circuits*. Artech House, 2003.
[25] M. Steer, *Microwave and RF Design, Volume 2: Transmission Lines*. University of North Carolina Press, 2019.

Chapter III : Design of the chip in 28nm FD-SOI technology

In this chapter, we introduce the system's implementation results. Firstly, we briefly introduce STMicroelectronics 28nm FD-SOI technology used in this work. Secondly, we explain our choice of channel. Then, we show the designed transmitter and receiver simulation results.

Next, we show the system's results for a 10Gbps data rate; we also show that our design has a BER lower than 10^{-12} up to 14Gbps. Afterward, we present the performances for a more practical case, where digital buffers are used at the output instead of a 50Ω output impedance.

Finally, we compare the proposed system performances to the essential works from the state of the art.

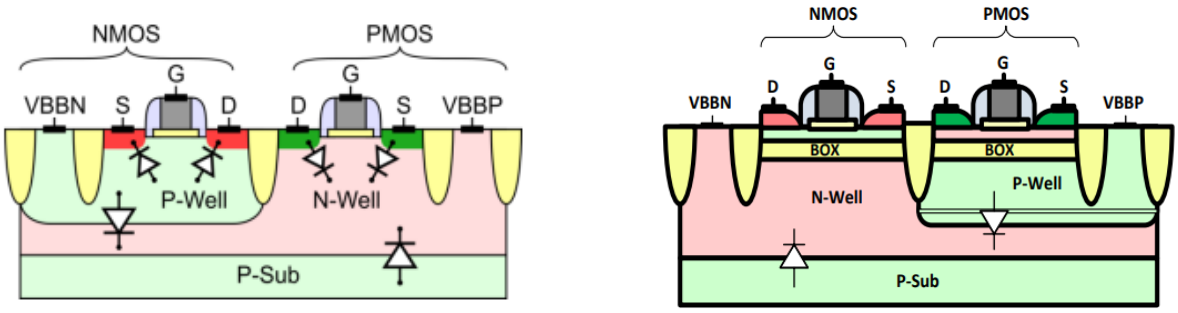
CHAPTER III : DESIGN OF THE CHIP IN 28NM FD-SOI TECHNOLOGY

I. Overview of the 28nm FD-SOI technology

I.1. The Front-end of line (FEOL)

Back in 1975, Gordon Moore, the co-founder of Intel, predicted that the number of transistors in an integrated circuit chip would double every year. His prediction was held for several decades and became the commonly known "Moore's Law". This law guided the semiconductor industry and defined its trajectory. The continuous scaledown of transistors increased the density of transistors while reducing the cost, or, as Moore said, "the cost per component is nearly inversely proportional to the number of components.", In other words, the more transistors we add, the cheaper each one gets. This cost reduction was one of the reasons that pushed the semiconductor industry to surpass the physical challenges they faced, through new fabrication techniques (new lithography techniques, for example), new transistors to better control the electrons or the use of different materials. Scaling the MOSFET transistors also allowed for faster transistors and better performances in general. However, serious physical challenges appeared for smaller transistors (90nm nodes and lower), such as Short channel effects.

Silicon-On-Insulator (SOI) technology was developed to reduce these effects and to advance to smaller nodes. Fully-depleted Silicon On Insulator (FD-SOI) transistors, that constitute the FEOL, i.e., the active devices (transistors), are shown in Figure III.1. They are built on a thin silicon film covering a buried ultra-thin oxide (BOX) layer. This BOX is absent in conventional BULK technology, as shown in Figure III.1. The BOX isolates the device's source/drain from the substrate; this eliminates the source/drain leakage current to the substrate, and results in a reduction of DC power consumption compared to the conventional bulk technology.



BULK CMOS

FD-SOI

(Flip-well flavor/LVT devices)

Figure III.1 : Conventional Bulk CMOS and FD-SOI transistors.

The fully-depleted (free of intrinsic charge carriers) very thin silicon channel improves the high K metal gate's capability to control the current flow through the channel efficiently and ultimately turn it off when the transistor is off; a better gate control results in a reduced gate leakage. Furthermore, the undoped channel eliminates the random dopants fluctuation (RDF) of the channel and reduces the V_{th} variability. The reduced variability means that the difference in power and performances between the process corners is decreased significantly.

Another FD-SOI advantage is the ability to raise or lower the threshold voltage to optimize power and performances; This is possible due to the forward and reverse body-biasing range available through the "back-gate" terminal.

It is worth mentioning that STMicroelectronics's 28nm FD-SOI technology proposes two transistors threshold levels: regular V_{th} (RVT) and low V_{th} (LVT), they both can be forward and reverse body biased (FBB and RBB) to reduce or increase the V_{th}, as shown in Figure III.2; Their V_{th} can be varied up to 250mV compared to 15mV in conventional Bulk technology.

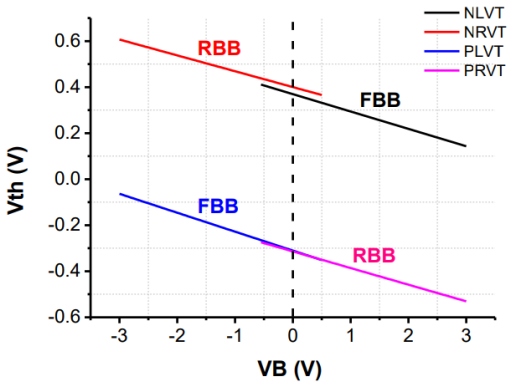


Figure III.2 : The threshold voltage variation for RVT and LVT 28nm FD-SOI technology [1].

I.2. The Back-end of line (BEOL)

The performances of an RF/mmW circuit depend on both the FEOL, and the BEOL, i.e., the transmission lines and the inductors, but also the electromagnetic properties of the substrate.

Figure III.3 shows the two different BEOL available in this technology; the first is the eight metal (8ML) layers stack that includes:

1- from M1 to M6: 6 thin copper metal layers used for digital parts routing.

2- IA and IB: 2 thick copper metal layers used for the power grid in digital circuits.

3- LB: A thicker aluminum Cap metal layer with very low resistance and parasitic capacitance.

The second BEOL available in this technology, and the one we used is the ten metal layers one (10ML), and it incorporates an extra two layers. This stack is used for large systems that require extra metal layers for routing. Furthermore, the losses of passive devices in the LB metal layer are reduced due to their greater distance from the substrate.

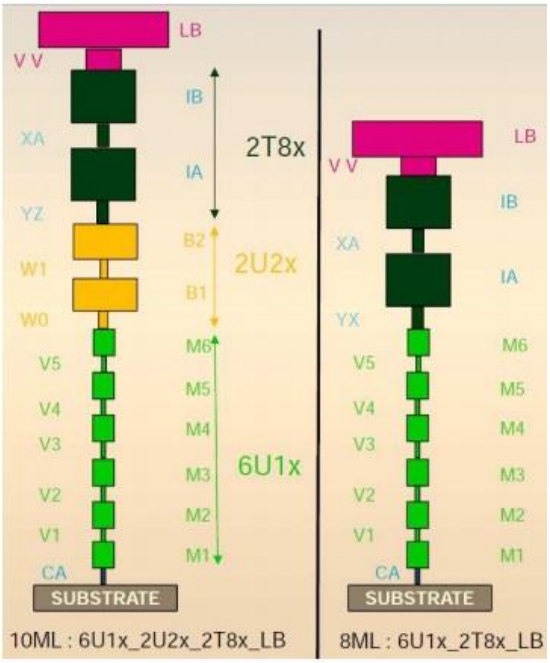


Figure III.3 : BEOL of the 28nm FD-SOI technology.

The author in [2] compared, in a school-case study, the attenuation of a 50Ω microstrip in 8ML 28nm FD-SOI technology to other STMicroelectronics technologies such as BiCMOS 55nm, CMOS 40nm, and CMOS 65nm. We notice in Figure III.4 that the 28nm technology has an attenuation lower than $0.9dB/mm$ even at 80GHz, behind the BiCMOS 55nm technology. It is worth mentioning that the BiCMOS 55nm technology is fully dedicated to RF/mmW applications with a thicker BEOL. This makes this technology suitable for both digital/mixed circuits and RF/mmW circuits.

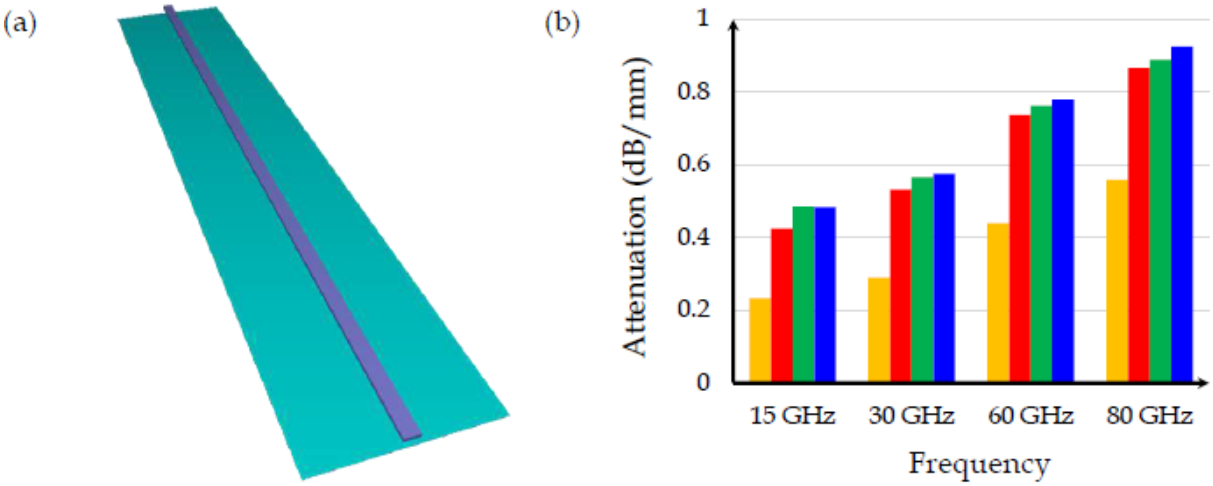


Figure III.4 : Transmission lines simulations for different technology nodes. (a) simulation model and (b) simulated attenuations. Orange, red, green, and blue refers to BiCMOS 55 nm CMOS 28 nm FDSOI, CMOS 40 nm, and CMOS 65 nm, respectively. [2]

II. The channel

II.1. Electromagnetic simulations

All of the passive devices shown in this chapter were simulated using the Keysight Momentum tool [3], where we used Momentum microwave mode due to the layout's large physical size and the high frequency required (up to 200GHz for some instances). We also used the complete foundry substrate definition; the engineers calibrated the substrate definition data in STMicroelectronics for better accuracy. In our simulations, we choose a mesh frequency equal to the maximum simulations frequency and a mesh density of at least 50 cells/wavelength.

To reduce the EM simulations time, we used a full ground plane. However, this full plane cannot be realized in practice, mainly due to the Chemical-mechanical polishing (CMP) fabrication process. The CMP flattens or smooths the surface of the metal layers; this is only possible if a specific metal density (usually between 20% and 80%) is respected; thus, a full metallic plane is not allowable (as well as the absence of metal). In practice, the ground plane is realized as mesh, as shown in Figure III.5. In some cases, to offer a low ohmic return plane, several stacked metals may be used. This mesh plane behaves similarly to the full ground plane up to a particular frequency (around 100GHz); above this frequency, the signal wavelength becomes larger than the size of the holes of the mesh.

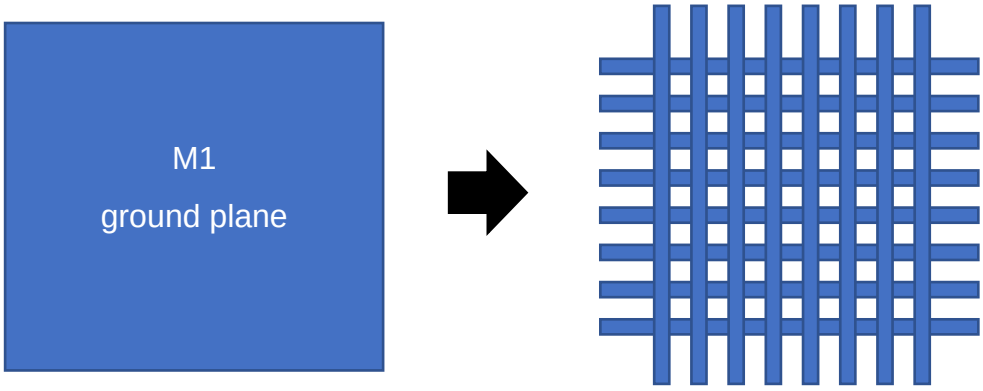


Figure III.5 : M1 full and meshed ground plane for process compliance.

The mesh is designed to respect the process design rules, e.g., the width of each strip and spacing between the metal strips. In practice, the design rules and metal densities guide the design and play an essential role in designing passive devices.

II.2. The Microstrips ground plane

Usually, for RF/mmW circuits, the signal conductor is placed in the higher layers, LB for our case due to its reduced resistivity, and the ground plane is placed in M1, as shown in Figure III.6.a, for lower losses. This is interesting for short distances; however, the transmission lines in this work are intended for several millimeters and up to twentieth millimeters. In this case, placing the ground plane in M1 will add significant constraints to the floorplan and waste silicon area, since most of the FEOL beneath the ground plane (usually larger than the signal width) cannot be used. Our first approach was to implement the ground in B1 layer (M7) of the stack; this will free the metal layers from M1 to M6 for routing. This was not enough, considering that the

layers B1 and B2 may be used for the power grid (VDD and GND), for example. Thus, it was evident that we had to use IA for the ground, as shown in Figure III.6.c, this ground can be combined with the digital chip's ground (that may be placed in the layer underneath).

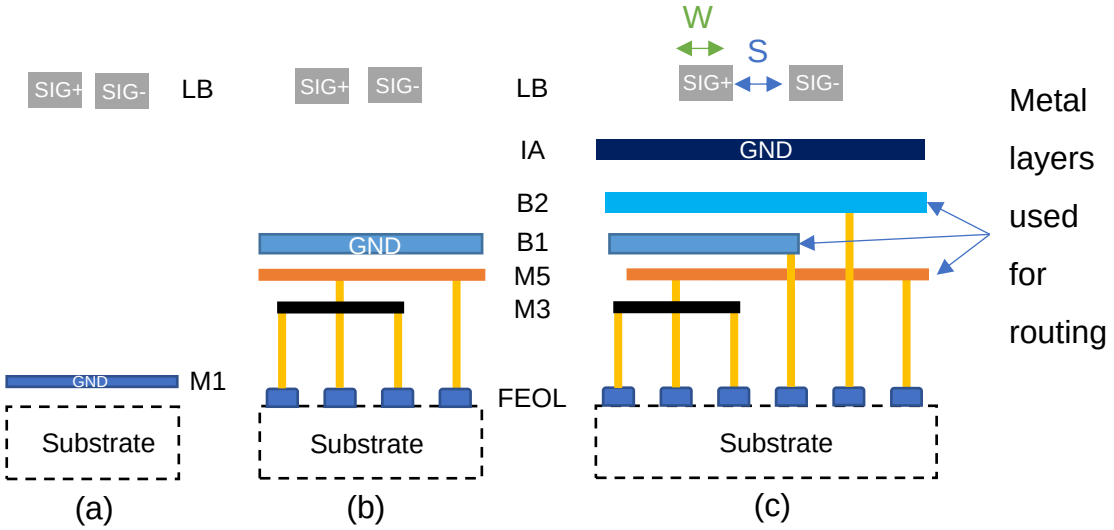


Figure III.6 : The ground plane in different metal layers. (a) the ground plane in M1. (b) the ground plane in B1. (c) the ground plane in IA.

We simulated 50Ω (at 60GHz) microstrips for the three different possibilities. We kept the signal conductor in the LB metal and varied its width. The width of the signal conductor varied from 11um when the ground is placed in M1, down to 8um for a ground placed in B1, and 4.4um when the ground is placed in IA. We compared their attenuation (dB/mm) as shown in Figure III.7; the figure also shows the disadvantage of this approach, which is the higher losses at higher frequencies, mainly metallic losses due to the significant skin effect. We notice that the losses are higher when the ground is placed in IA. The height of the BEOL is one of the most crucial microstrip parameters; it directly impacts the widths of the strips, and consequently, their resistance. In other words, for the same impedance, 50Ω in this case, finer BEOL along with tighter strips significantly increases the resistance. However, we believe these extra losses remain acceptable as it allows us to take full advantage of the stack for routing.

II.3. Edge-coupled microstrip lines

We decided to use edge-coupled microstrips for differential signaling; Firstly, because they correspond to coplanar propagation that depends less on the BEOL height for differential mode, secondly because it offers better noise immunity as introduced in chapter I.

Designing differential lines adds an extra parameter: the spacing between the lines. The spacing between the two lines can be used to couple the lines more or less, and vary their impedances (even and odd impedances). In our case, we choose to design the lines to have a $Z_{comm} = 50\Omega$ and a $Z_{diff} = 90\Omega$ at 60GHz.

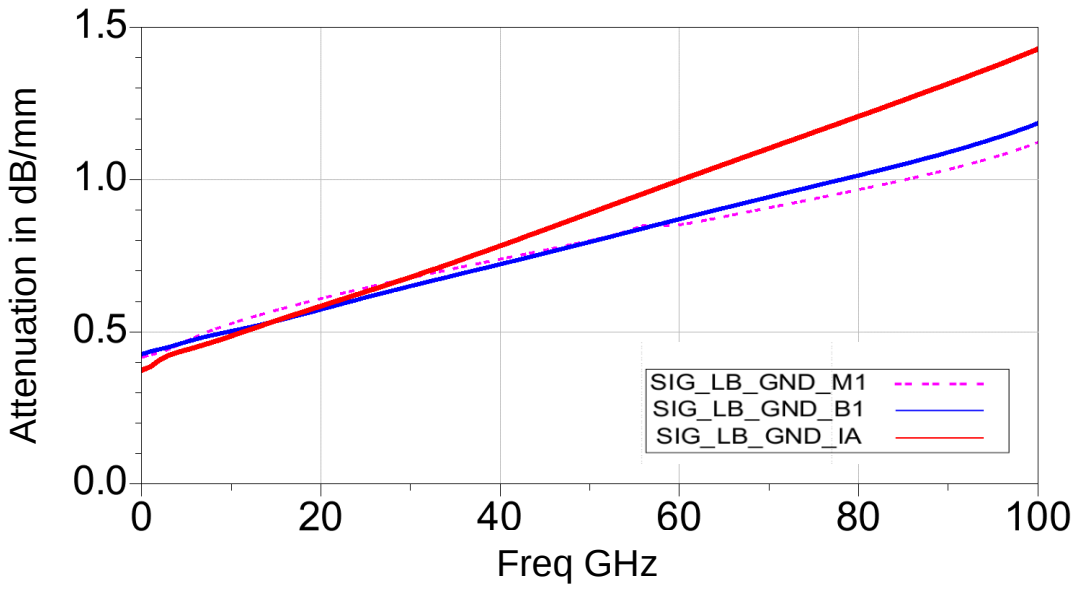


Figure III.7 : The attenuation for a 50Ω microstrip when its ground is placed in M1, B1, or IA (EM simulation results).

We kept the 4.4um wide 50Ω microstrip line with a ground plane in IA, plus we added another strip at a distance S, as shown in Figure III.6. We varied this distance for different values; Figure III.8 shows the Z_{diff} for three spacing values S equal to 4um; 7um and 9um. We notice that differential impedance increases with the spacing, the $Z_{diff} = 90\Omega$ is achievable with a 9um spacing. In Figure III.8, we show the insertion loss for differential propagation (in dB/mm) for the previous edge-coupled microstrips. The edge-coupled microstrips differential impedance was matched; we notice that the losses are lower for the 9um strips. It is worth mentioning that the coupling and hence the noise immunity is lower than the 4um spacing lines, and that for future versions of this work, the lines may be tightly coupled for better noise immunity and better density; at the expense of more losses.

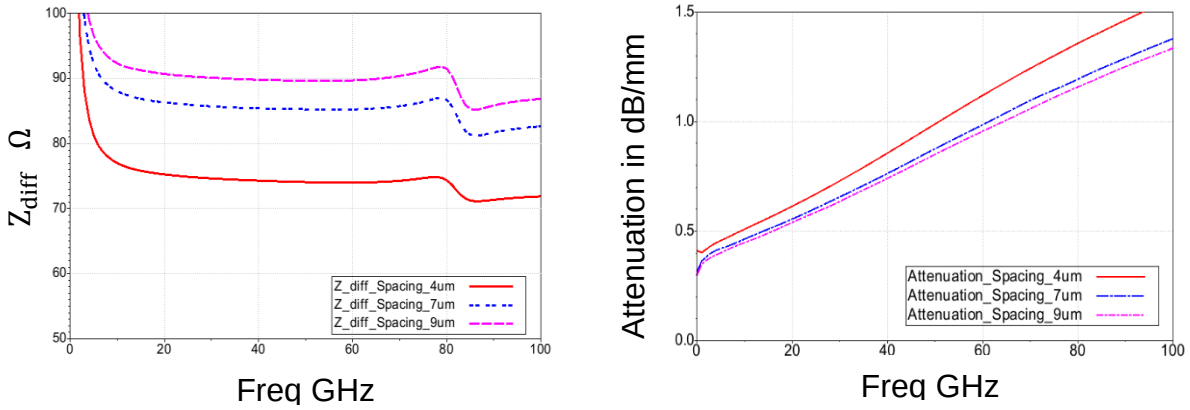


Figure III.8 : The spacing between edge-coupled microstrips impact on the differential impedance Z_{diff} and the attenuation in dB/mm (EM simulation results).

To reduce the surface used by the lines in the prototype, we decided to implement the 4.6mm long lines, as shown in Figure III.9.a; this allowed us to gain some surface. We mentioned earlier that the design rules and the density of metals

per layer complicate the ground plane's design. However, for this case, We managed, while respecting all the design rules, to draw a full ground plane under the conductor strips, and used the meshed plane elsewhere, as shown in Figure III.9.c. The insertion losses of the drawn line are shown in Figure III.9.b; we expect an attenuation of 5dB in differential propagation.

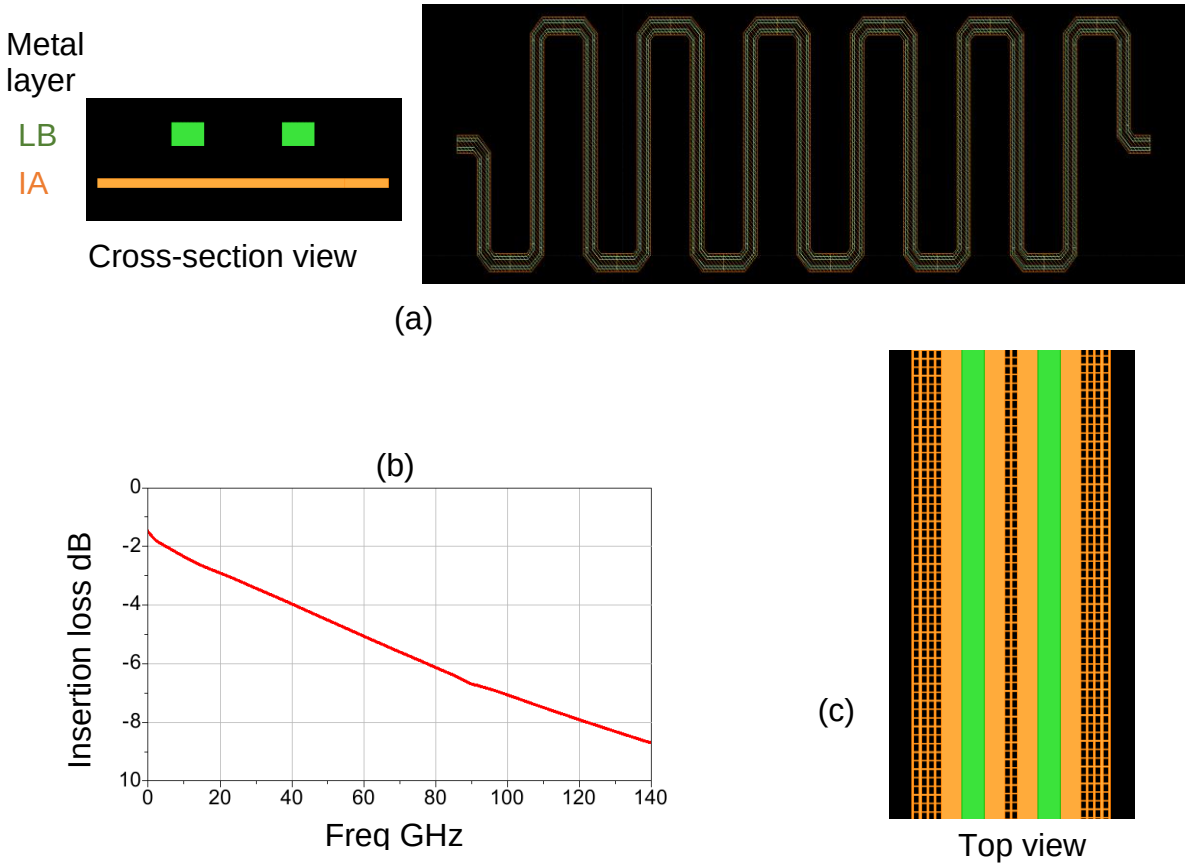


Figure III.9 : (a)The layout of the lines to reduce the used area and its cross-section layout. (b) The transmission lines insertion losses in dB (EM simulation results). (c) Layout respecting the ground plane density requirements.

III. The 10Gbps transmitter architecture

The transmitter architecture for our prototype is shown in Figure III.10. It can be decomposed into three main segments:

1. The duobinary modulator: takes a clock and data signal and delivers the desired signal with a compressed spectrum.
2. A passive mixer: takes the duobinary modulated signal (in baseband) and the local oscillator (LO) 60GHz signal to generate a modulated signal transmitted to the receiver through a channel.
3. The rest of the circuitry can be considered as an interface between the duobinary modulator's digital gates and the 50Ω probes. Consequently, we used a matching network and an LNA (+buffers) to transform the sinusoidal input clock into a squared clock. And a more straightforward matching using a resistor and buffers at the data input.

The blue squares in Figure III.10 represent the pads; we will show that the used pads were optimized to have low parasitic capacitance at 60GHz. The grey squares represent the transmission lines used to bring the signal from the pads into the gates. These passive devices were simulated and optimized using ADS Momentum. The DC pads are not shown.

For measurement purposes, the transmitter requires three external RF signals (and one DC input):

1. The clock signal is required to synchronize the duobinary modulator different digital gates.
2. The data signal to be transmitted, we choose to use an external signal instead of implementing a PRBS generator on-chip in order to test the prototype chip for different data rates.
3. The third signal is the 60GHz LO signal; we used an exterior LO generator instead of implementing a VCO for better control and flexibility in measurements.

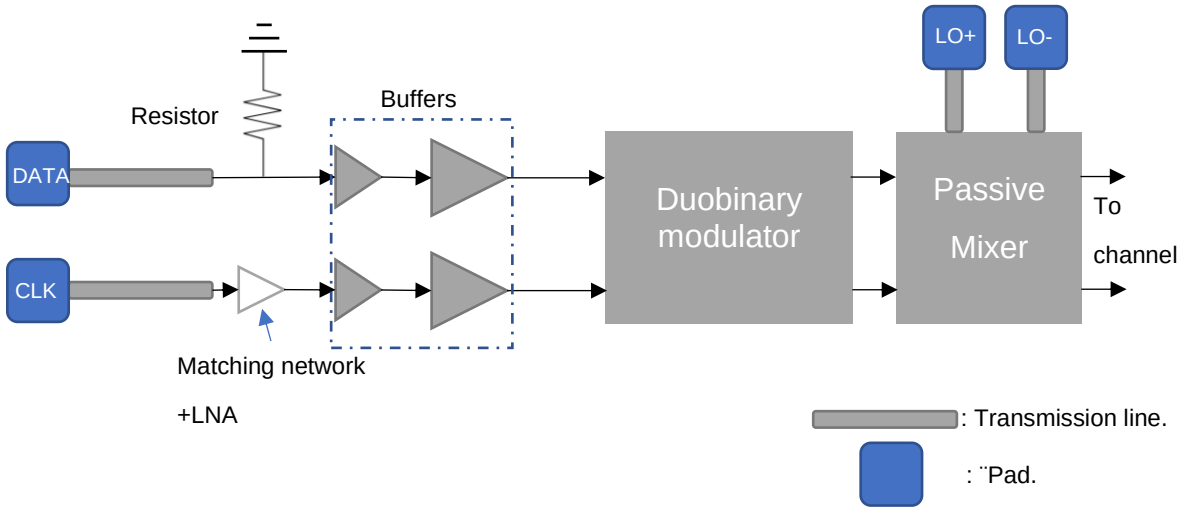


Figure III.10 : Block diagram of the transmitter implemented in the prototype chip.

III.1. Clock buffers

The external clock cycle (10GHz) is amplified at the chip's input and goes through a series of inverters to generate a square-like signal from the input sinusoidal signal. We used the circuit shown in Figure III.11 to match the probe 50Ω impedance. It is worth mentioning that the pad and the transmission line (280um) were simulated in ADS Momentum and taken into consideration for the impedance matching. A voltage supply $V_{DD} = 1V$ is used for this work. Firstly, AC coupling is used to block the DC component completely. Next, an inverter with resistive feedback is used. The resistor used in feedback allows the inverter to self-bias and increases its 3dB bandwidth by $1/R_F C_L$ with R_F the feedback resistor, C_L the load capacitance and r_o the drain output resistance, the resulting 3dB bandwidth is given by [4]:

$$\omega_{3dB} = \frac{1}{r_o C_L} + \frac{1}{R_F C_L} \qquad \text{(III-1)}$$

Combining the low-pass filter of the inverter and the AC coupling results in bandpass filter behavior, as shown in Figure III.12 for different feedback resistors. We

notice that the gain (7dB at 10GHz for $R_F > 10k\Omega$ is higher than the one for lower resistor value (1kΩ), the gain, when using $R_F = 20k\Omega$, is stable on a larger bandwidth. However, the $R_F = 10k\Omega$, is a better compromise to attenuate/not amplify the out-of-band noise and reduce the phase noise and jitter. The inverter with resistive feedback consumes 0.93mW when using a 10kΩ resistor, and is considered as a large part of the transmitter power consumption. The AC-coupled inverter with a resistive feedback overall transfer function is given in [4] as:

$$\frac{v_{out}}{v_{in}} \cong \frac{sg_m R_F C_c}{g_m + \frac{1}{r_o} + s\left(C_L + C_C + \frac{R_F C_c}{r_o}\right) + s^2 R_F C_c C_L} \quad \text{(III-2)}$$

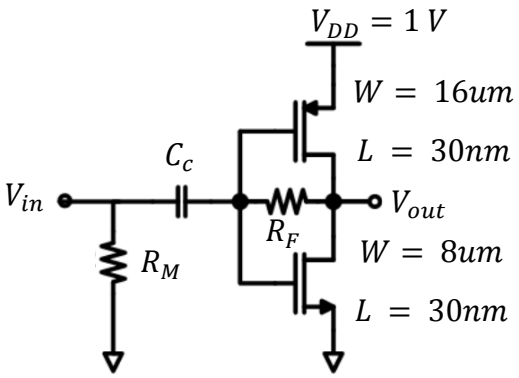


Figure III.11 : Clock buffer AC-coupled inverter with resistive feedback.

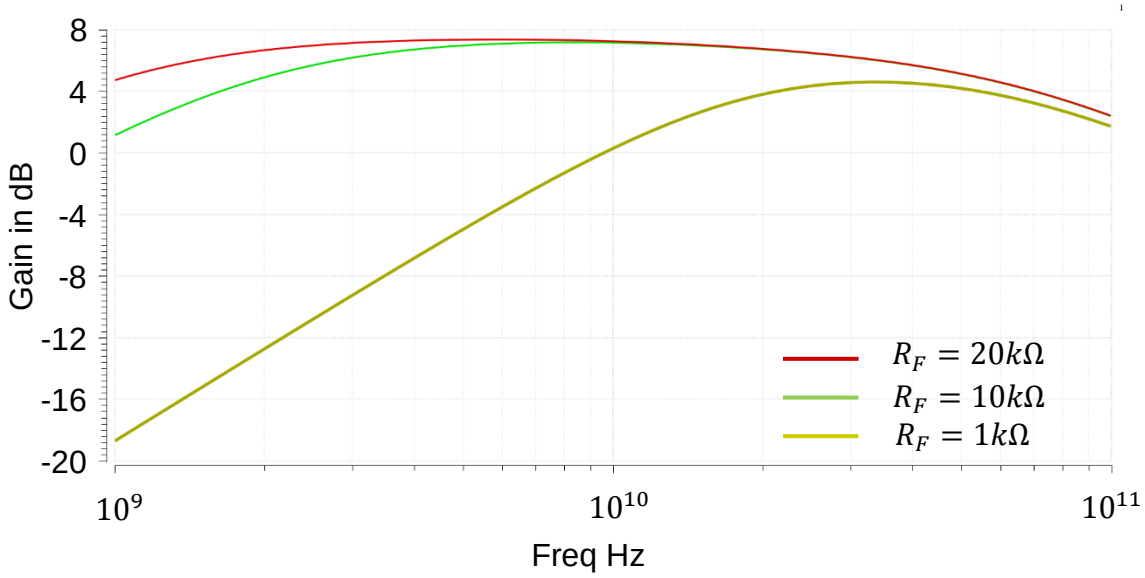


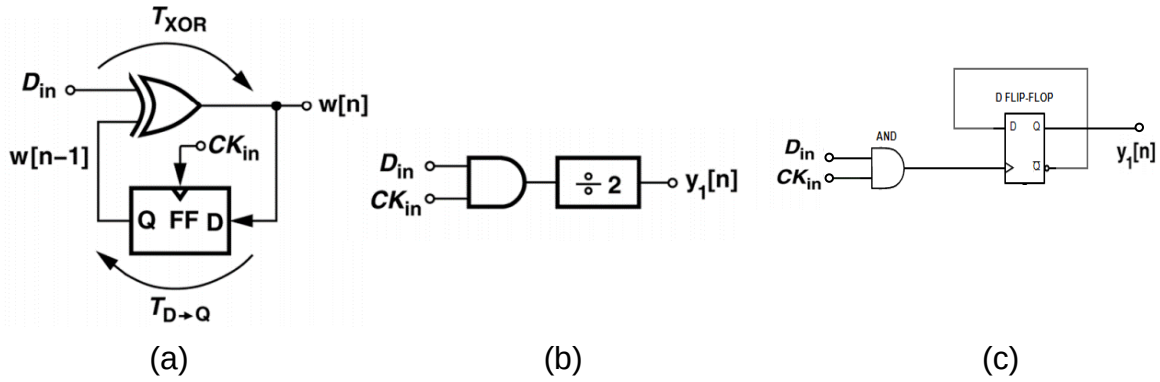
Figure III.12 : Gain of the terminated AC-coupled inverter with resistive feedback shown in Figure III.11, for different feedback $R_F$values.

Next, the clock signal goes through 4 inverters; The inverters are used to generate a square-like signal. The disadvantage of this approach is its sensitivity to PVT.

III.2. Duobinary modulator

As mentioned in chapter II, the duobinary modulator can be decomposed into two main parts. Firstly, the differential precoder that precodes the changes of bits, when the input binary data changes its state ("0" to "1", or "1" to "0"), the differential precoder produces a "1" in its output. The output is "0" when no changes occur at the input. The conventional differential precoder and the used precoder herein are shown in Figure III.13. As shown in [6], controlling the timing is difficult for high data rates due to the feedback loop. Breaking the loop is preferred. Hence, the proposed precoder consists of an AND gate and a "divide-by-2 counter"; this results in the same functionality given by [6] :

$$y_1[n] = y_1[n-1] \oplus D_{in}[n] \tag{III-3}$$



(a) (b) (c)

Figure III.13 :(a) Conventional differential precoder. (b) Differential precoder proposed in [6] (c) Used differential precoder.

The divide-by-2 counter, usually used as a clock divider, can be implemented in different ways; we used a D flip-flop with feedback, as shown in Figure III.13.c. The inverted output $\overline{Q}$ is connected to the data input terminal, the input signal is fed to the clock input, and the output is taken from Q. In this circuit, the input acts as a clock for the flip flop, and the data on the D input is clocked through to the output. The used circuit has a z-transform $H_1(z) = 1/(1 + z^{-1})$.

The second part of the duobinary modulation, i.e., the reshaping filter, requires a circuit with a z-transform of $H_2(z) = 1 + z^{-1}$. Usually, the channel is taken into consideration to recreate this response. However, for the bandpass modulation case, this is not possible, and the $1 + z^{-1}$ function must be implemented before the upconversion. Thus, we used the circuit in Figure III.14.a. The interest of choosing AND and NOR gates at the output instead of NAND and XOR, as in [7], will be explained in the mixer section.

The circuit shown in Figure III.14.a was implemented using some of the high-performance standard cells from the design kit; high-performance cells were chosen due to their enhanced response at higher frequencies (10GHz and higher), and to assess whether a simple digital implementation will be sufficient in future implementations. However, during the layout step, we modified some of them to control rising/fall time properly, and we inserted buffers, not shown in Figure III.14.a, when necessary. For accurate simulations, all of the modulator parasitics were extracted

using StarRC (RCc extraction); this allowed us to accurately see the coupling between the different nets and improve the layout dependently.

The DC wire, and the region surrounding of the modulator, were simulated in Momentum to extract the parasitic inductances' values. Decoupling capacitors of different values were implemented all along the DC input wire.

Another advantage of using digital standard cells or similar layout techniques is the compact size of the cells; the duobinary modulator shown in Figure III.14.a was implemented in a $14um \times 3.3um$ area. However, we suggest a full-custom design for higher data rates, and we believe that one of the data-rate limiting factors in our design is this part.

Each output of the duobinary modulator is then fed into a D Flip-flop to synchronize the outputs. As we noticed, some glitches might appear, especially for different process corners, at the AND and NOR gates' outputs (Out1 = BB+ and Out2 =BB-). Thus, synchronization helps remove the glitches at the expense of an extra clock cycle to the modulation. Next, the outputs of the flip-flops are fed into the mixer.

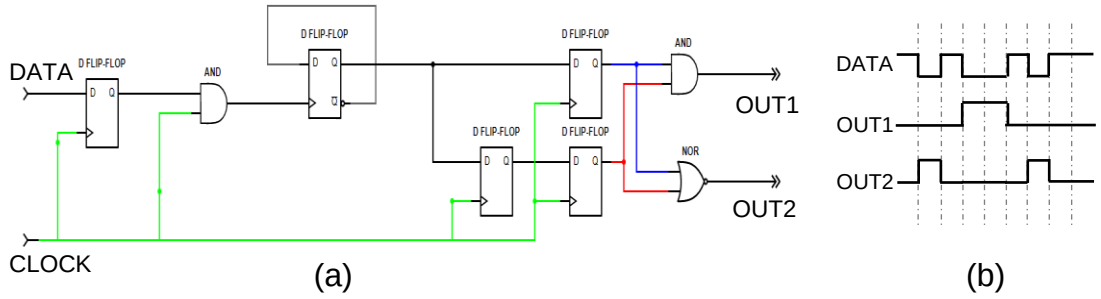
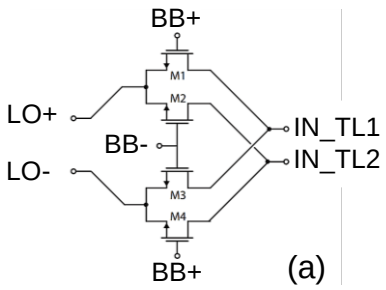


Figure III.14 : (a) Duobinary modulator. (b) input and output modulated signals.

III.3. Mixer

In this work, we used a FET quad passive mixer consisting of four switches, as shown in Figure III.15.a. It consists of two differential pairs fed with the baseband signals (BB+ and BB-) representing the duobinary modulator's output and a differential 60GHz LO (LO+ and LO-). The 60GHz is generated externally for the prototype chip. Hence a 50Ω matching was necessary. This type of mixer does not require any DC bias current and thus do not dissipate standby power, and commutes the signal in the voltage domain instead of current.

On the other hand, they cause conversion loss instead of gain for the active Gilbert cell and suffer from higher noise figure. However, in our case, we believe that these disadvantages are acceptable. Herein, we will show some of the mixer parameters; detailed studies can be found in [8].



IN_TL1 output example

BB- \ BB+	0	1
0	0	LO+
1	LO-	0 (LO- + LO+)

(b)

Figure III.15 : (a) A passive double-balanced mixer. (b) the output of a passive mixer.

This mixer is composed of four NMOS switches that switch between the ON and OFF states. The gate input modulates the channel resistance and mixes the gate and drain inputs in the voltage domain, resulting in frequency components at $(\omega_{RF} - \omega_{LO})$ and $(\omega_{RF} + \omega_{LO})$ (plus other harmonics). Figure III.15.b shows the different states each output of the mixer can take. When BB+ and BB- have different binary states, the output is either LO+ or LO-; these two states provide the phase inversion necessary for the duobinary modulation's correct functioning. When BB+ and BB- are both equal to "1" the output signals are severely attenuated due to the destructive interference, but this requires extra care to balance the signals path. If both BB+ and BB- signals are equal to "0", then the signal is not transmitted to the channel.

Since the duobinary modulation requires only three output states, we choose to use the state (BB+ = BB- = "0") and eliminate the case (BB+ = BB- = "1"). This is realized by choosing the AND and NOR gates at the output of the duobinary modulator instead of NAND and OR as in [7]. Figure III.16.b shows an example of the output for both cases.

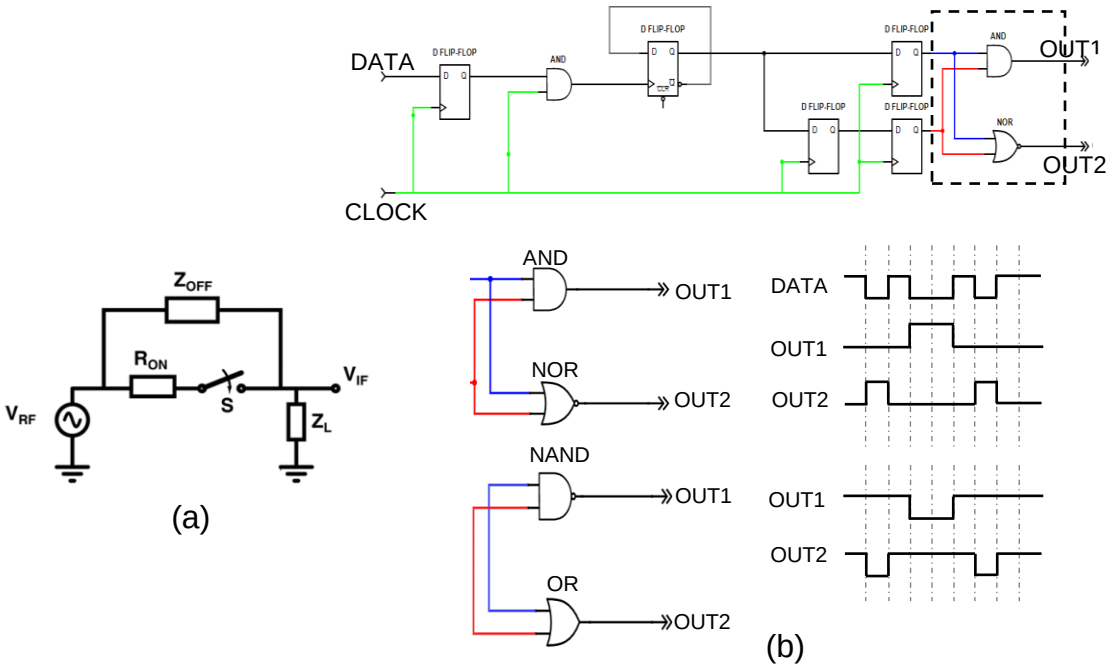


Figure III.16 : (a) One transistor passive mixer R_{ON} and Z_{OFF}. (b) Outputs of the duobinary modulator for different gates (AND and NOR) or (NAND and OR).

Each switch can be represented, as shown in Figure III.16.a. In the ON state, the signal flows through the impedance R_{ON}, it can be seen as a sum of the transistor resistances (R_D drain resistance, R_s source resistance, and r_{ds} the on-resistance, and the parasitic capacitances). Preferably, the switch is designed to have a low R_{ON} to reduce the insertion losses. In the off state, the signal sees the impedance Z_{OFF}, this impedance should be very high for better isolation between the RF and IF ports; it is mainly due to the drain and source parasitic capacitances. R_{ON} and Z_{OFF} are approximated by [8] (|| means in parallel):

$$R_{ON} = \frac{L_g}{\mu_n C_{OX} W[(V_g - g_c V_{RF}) - V_{th} - V_{ds}]} \quad \text{(III-4)}$$

$$Z_{OFF} = \frac{1}{j\omega_{RF}(C_{GS}||C_{GD})} \quad \text{(III-5)},$$

with Z_L the load impedance. We notice that R_{ON} decreases when lowering L_g (transistor channel length) or increasing the transistor's width. However, by increasing the width, the transistor's parasitic capacitances increase too and thus reduces Z_{OFF}. At high frequencies, the parasitic capacitances create low impedance paths for the RF signal and might lead to leakage. Furthermore, in the post-layout simulation, we noticed that the number of fingers used impacts (slightly) the conversion gain. The conversion gain of a double-balanced passive mixer is given in [8] as :

$$G_C = 20\log_{10}\left(\frac{2}{\pi}\left(\frac{Z_L}{Z_L + (R_{ON}||Z_{OFF})}\right) - \frac{2}{\pi}\left(\frac{Z_L}{Z_L + Z_{OFF}}\right)\right) \quad \text{(III-6)}$$

We used the smallest transistors length available to maximize f_{max} and f_t; and 25um wide transistors. Herein, we show the conversion gain of the designed passive mixer, at 60GHz, is expected around -5.3dB conversion gain, only 1.2dB higher than the conversion gain of an ideal dual-balanced passive mixer (-4dB) [8], and 1.7dB lower than a similar mixer at a similar frequency (60GHz) in 65nm CMOS technology [9]. The conversion gain varies by around 0.2dB in our band of interest (±5GHz from 60GHz), and 1.2 dB from 40GHz to 90GHz. Figure III.17 shows the conversion gain with respect to the input power in dBm. We notice that 1 dB power compression occurs at a higher input power (around 7dBm) than the input power used in our application (-4dBm).

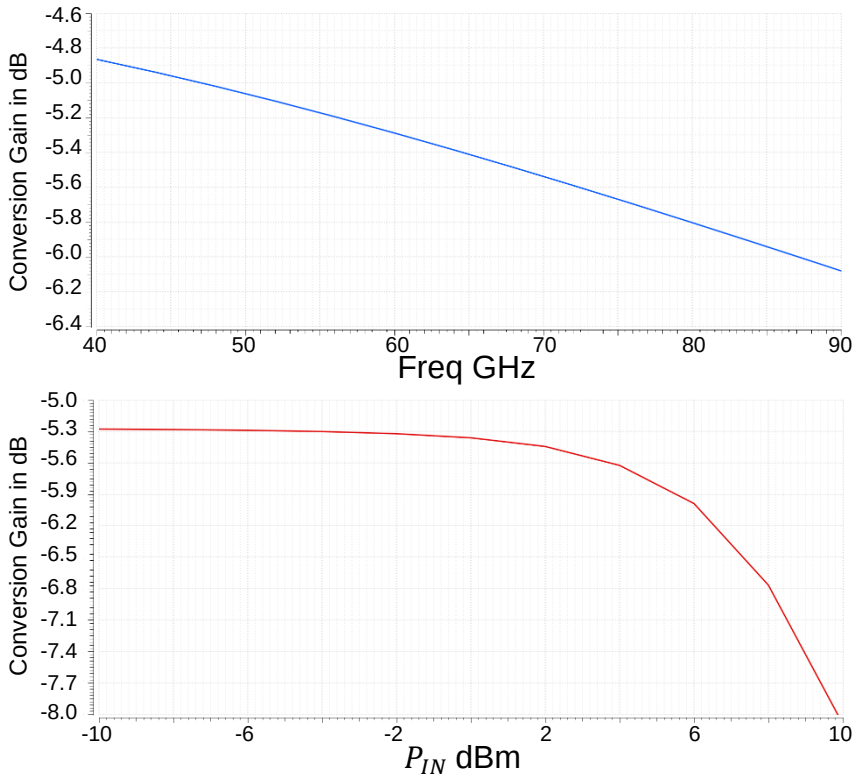


Figure III.17 : The conversion gain of the mixer with respect to frequency and input power.

The transistors' layout is shown in Figure III.18; staircase access layout was used for the transistors whenever possible to reduce the parasitic capacitors; we used both the StarRC extraction tool and Momentum EM simulations to extract the parasitics and model the interconnects appropriately.

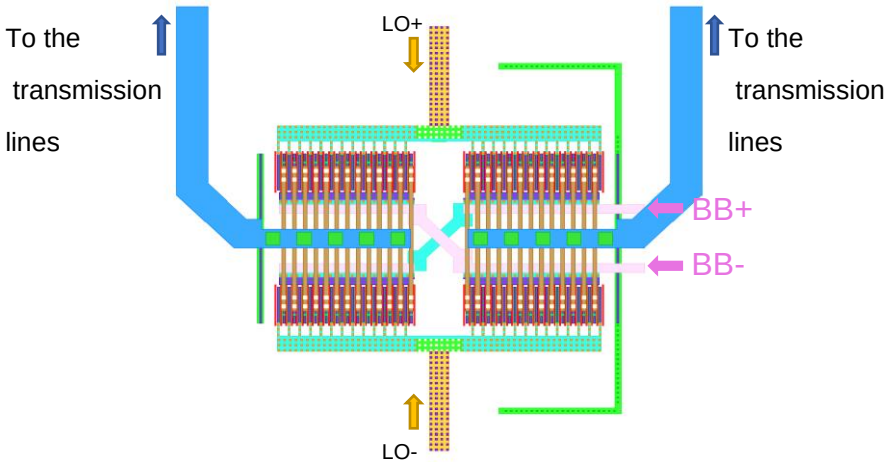


Figure III.18 : Layout of the passive mixer in 28nm FD-SOI Technology.

The three possible outputs in time-domain are shown in Figure III.19; these results are post-layout simulation results with a well-matched output in typical process corner and standard temperature 27°C.

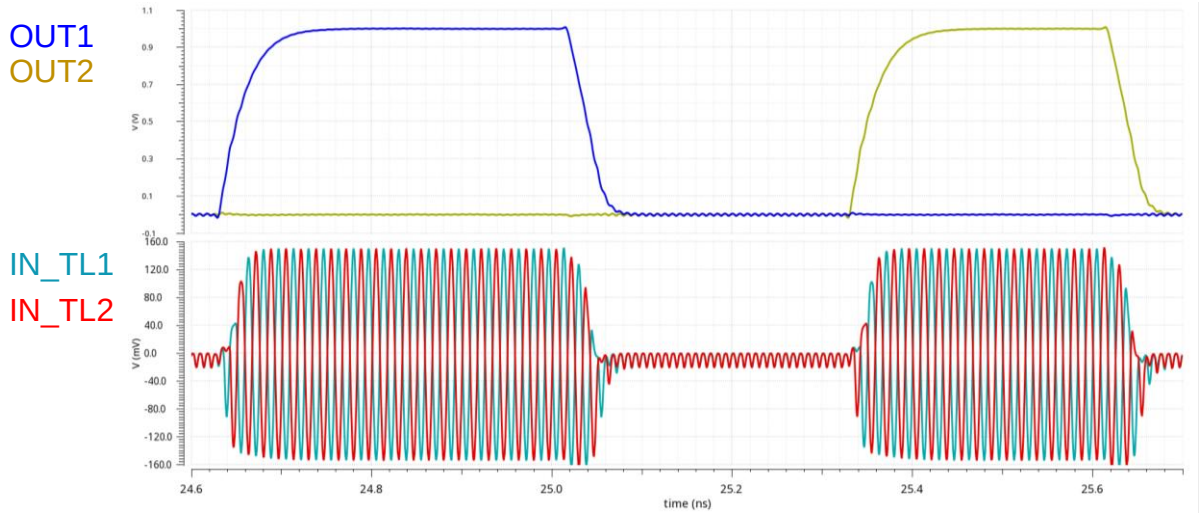


Figure III.19 : Time-domain waveforms representing: OUT1 and OUT2 the outputs of the duobinary modulator. IN_TL1 and IN_TL2 the outputs of the mixer (input of the transmission lines).

III.4. Inputs matching

The LO is generated externally in this work and is brought to the mixer through pads and transmission lines (around 350um long), as shown in Figure III.20; we found that their insertion loss is around 0.36dB at 60GHz in post-layout simulations.

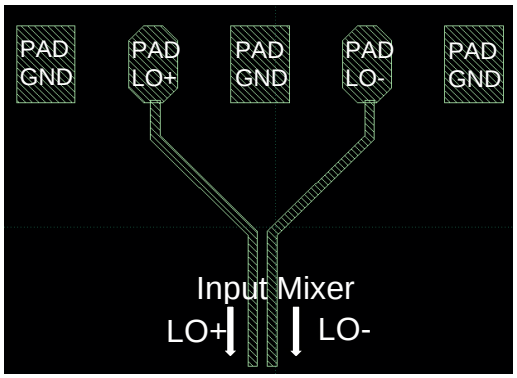


Figure III.20 : Layout of the LO input transmission lines.

The pads used in this work were optimized for a low capacitance, where the conventional shielded FC44S flip-chip RF pads shown in Figure III.21 from [2] present a parasitic capacitance of around 30fF at 60GHz. The parasitic capacitance can be reduced by proper layout. In this work, instead of using the shielded pads, we did not use a metal ground below the signal pads, revealing the substrate and resulting in a frequency-dependent parasitic capacitance. We also reduced the size of the signal pad to the minimum required. The design rules and metal density conditions were respected in the layout. Momentum simulations showed that the parasitic capacitance is around 11fF at 60GHz, as shown in Figure III.21.c.

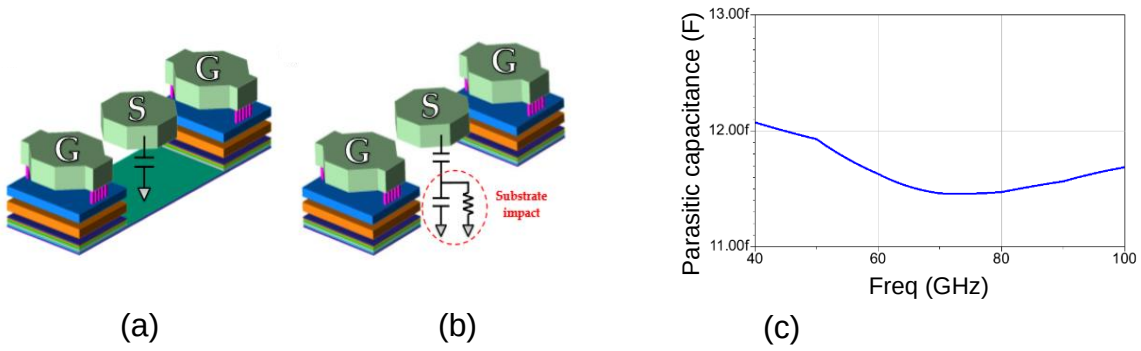


Figure III.21 : (a) Layout of a shielded RF pad. (b) The layout of a non-shielded RF pad [2]. (c) The simulated parasitic capacitance of a non-shielded RF-pad.

The LO input return loss is about -20dB at 60GHz; Figure III.22 also shows the return loss for the CLOCK and DATA inputs, which are lower than -20dB up to 65GHz.

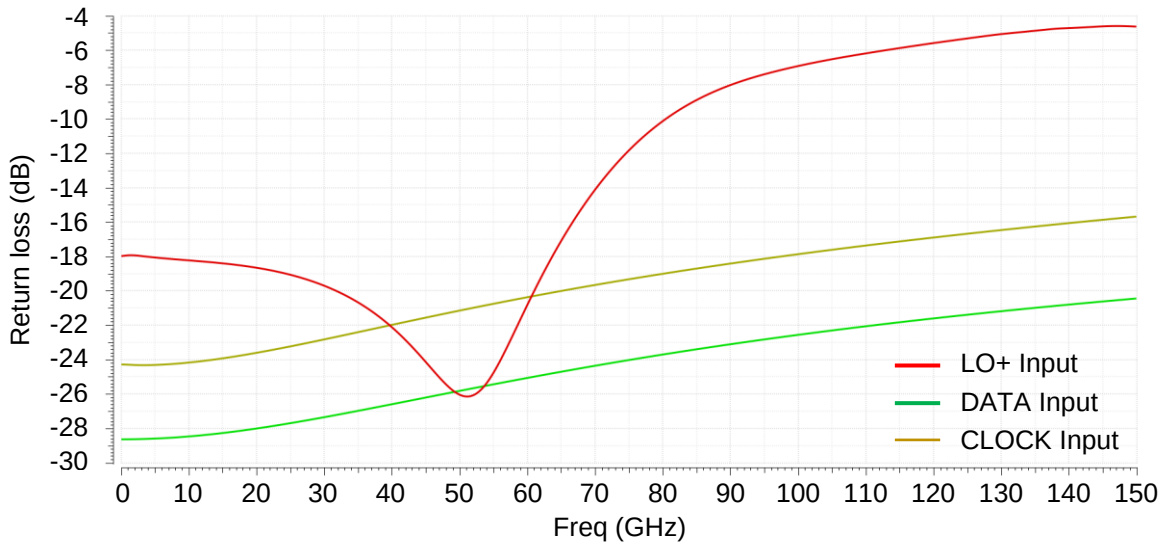


Figure III.22 : The return loss at the inputs of the transmitter for the LO+, DATA, and CLOCK inputs.

IV. The 10Gbps receiver architecture

The receiver receives a 60GHz modulated signal, and as mentioned in chapter II, we preferred to use non-coherent demodulation to demodulate the signal; hence, a simple envelope detector is sufficient to recover the data signal. Figure III.23 shows the receiver block diagram. Firstly, a high-pass filter is used to suppress any low-frequency noise (such as crosstalk or RF leakage signal); it also acts as a matching network for the square-law devices (M1 and M2). The architecture of the receiver is shown in Figure III.24. The self-mixer transistors M1 and M2 deliver a current proportional to the input power. Next, their output is filtered through the current-mirror, with resistive compensation, for higher bandwidth; this low-pass behavior attenuates the leakage signal and harmonics components generated by the mixing process. The output is matched to 50Ω for measurement purposes. However, a more practical case with buffers for a real digital application instead of a 50Ω probe was also studied during this work.

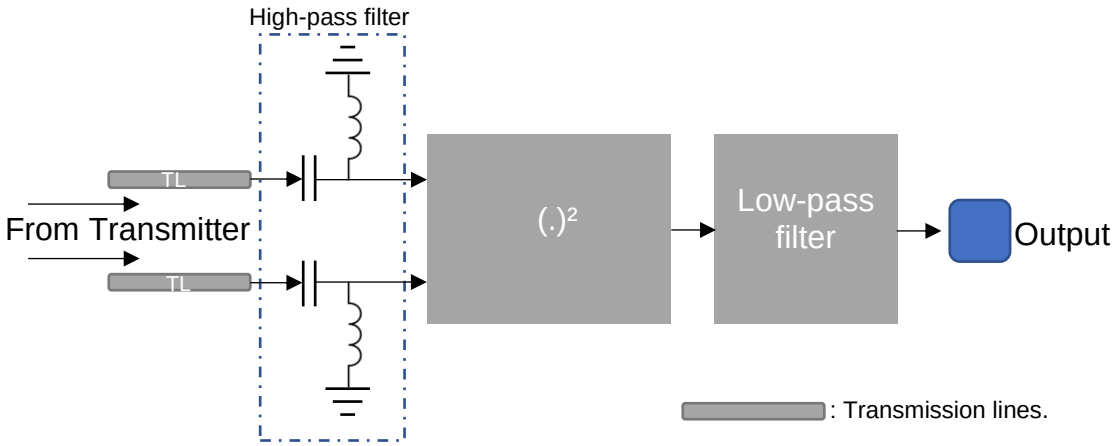


Figure III.23 : Block diagram of the receiver.

The transistors M1 and M2 input voltages can be written as :

$$V_{GS1} = V_G - V_A \cdot \cos(\omega_0 t) \quad \text{(III-7)}$$

$$V_{GS2} = V_G + V_A \cdot \cos(\omega_0 t) \quad \text{(III-8)}$$

Where V_A is the voltage swing at the source of M1 and M2. With proper bias, these transistors function in saturation to generate square-law behavior. The drain current of a MOSFET in saturation can be written as :

$$I_D = \frac{\mu_n C_{OX}}{2}\frac{W}{L}(V_{GS} - V_{TH})^2(1 + \lambda V_{DS}) \quad \text{(III-9)}$$

With λ the channel-length modulation parameter. Summing the two drain currents in node Q gives :

$$I_{DQ} = \frac{\mu_n C_{OX}}{2}\frac{W}{L}[2(V_G - V_{TH})^2 + 2(V_A \cdot \cos(\omega_0 t))^2](1 + \lambda V_{DS}) \quad \text{(III-10)}$$

This equation can be further simplified:

$$I_{DQ} = \frac{\mu_n C_{OX}}{2}\frac{W}{L}[2(V_G - V_{TH})^2 + {V_A}^2(1 + \cos(2\omega_0 t)](1 + \lambda V_{DS}) \quad \text{(III-11)}$$

Equation (III-11) shows that the resulting current is a combination of a biasing current and the squared input voltage swing. The resulting signal contains a component at $2\omega_0$; In practice, it will also include some leakage signals at 60GHz. These signals are next attenuated using the current-mirror.

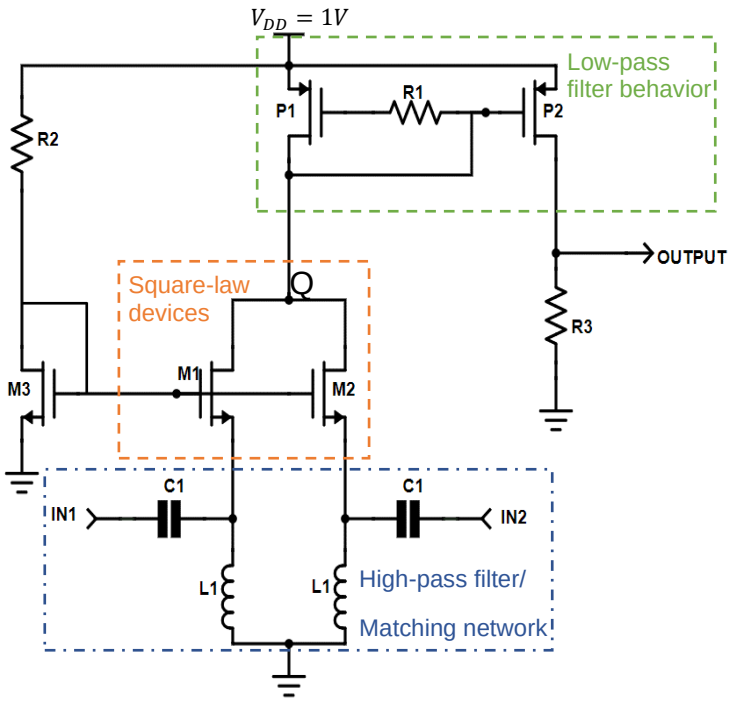


Figure III.24 : Schematic of the demodulator.

To properly size the transistors M1 and M2, several parameters can be varied (W, L, V_B, V_G). We used minimal length transistors for a better gain and faster transistors; we did not use the back bias. The transistors width should be chosen for sufficient I_D current as given by equation (III-9) and a proper input impedance matching. The common-gate used in this circuit is useful for its low input impedance and high output impedance; the input impedance Z_{in} can be approximated by:

$$Z_{in} \cong \frac{1}{g_m + g_{mb}} \qquad \text{(III-12)}$$

Where g_m is the transconductance transistor and g_{mb} the back gate transconductance. Thus, the matching LC network is tuned to match the transistors' Z_{in} and the transmission lines impedance.

IV.1. The matching network

L1 and C1 values were chosen to match the load impedance to the line's differential impedance. The chosen values are L1=150pH and C1=70fF.

We used M9 for the signal and an unshielded inductor instead of patterned ground shields or floating fingers [10], as we had better performances and a higher Self-Resonance Frequency (SRF). The challenge of such an approach is to respect the metal densities for lower metal layers M1 and M2. Hence, dummies were added in the center and around the inductor to respect the minimal required density rules. The chosen inductor, shown in Figure III.25, has a value of 150pH at 60 GHz with a quality factor of around 20, as shown in Figure III.26.b; the SRF was higher than 120GHz.

Two inductors were used instead of a differential one; special care was taken to place their ground connections as close as possible and keep the layout symmetrical, as shown in Figure III.26.a.

It is worth mentioning that the inductors use most of the surface required for this approach and are necessary for large bandwidth and high data rates. For lower data rates, they may not be required if the low-frequency noise is restrained at the transmitter side; however, further investigation is needed. The simulated return loss S_{11} at the input of the receiver is shown in Figure III.27. We expect an insertion loss lower than -10dB in the band of interest.

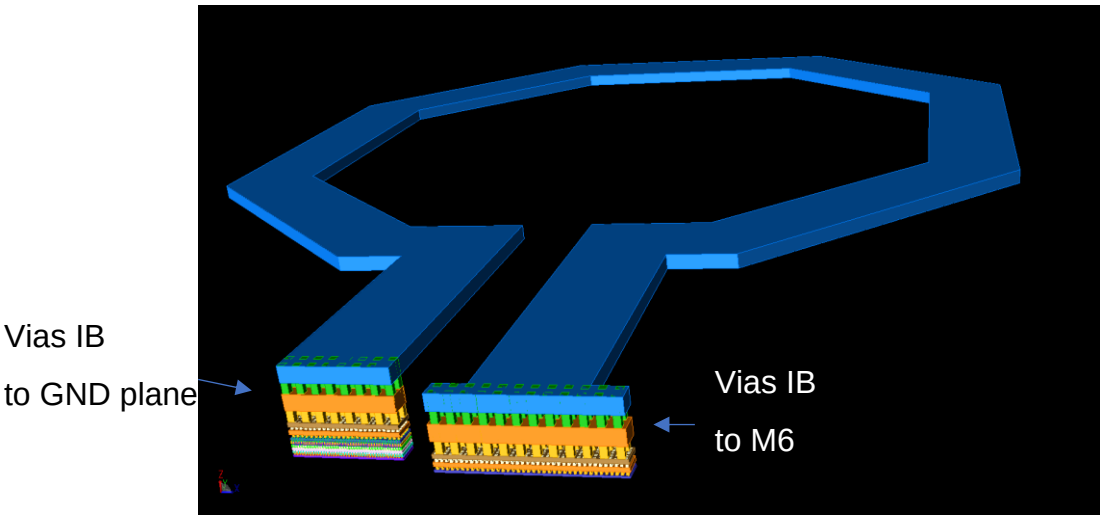
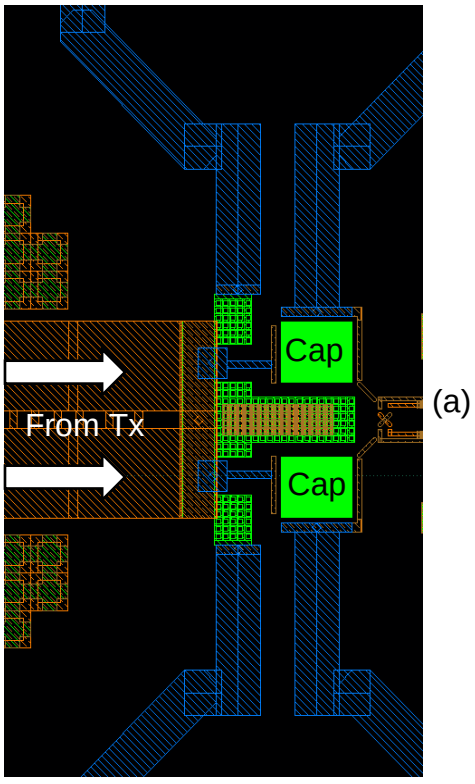
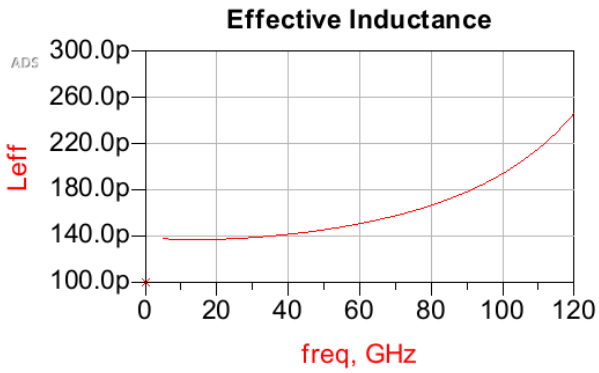


Figure III.25 : Implemented L1 inductor (Surrounding and dummy metals not shown).



(a)



(b)

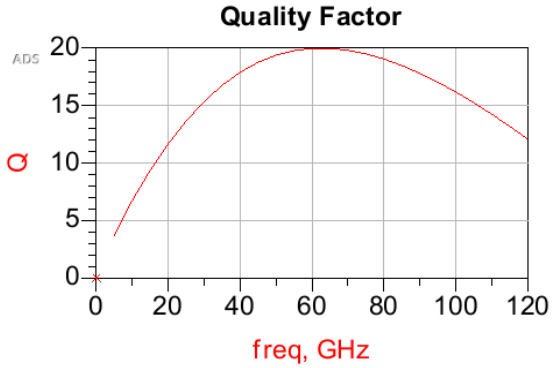


Figure III.26 : (a) Symmetrical layout for the input signals. (b) Inductance and quality factor values.

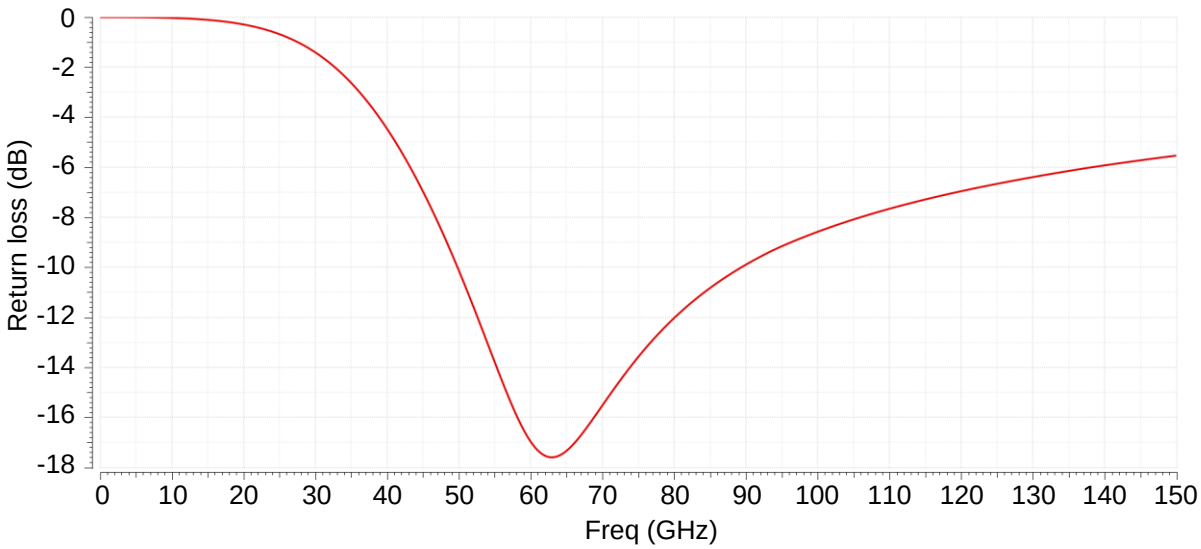


Figure III.27 : Receiver input return loss.

IV.2. The circuit's core

Two current mirrors are used in Figure III.24 circuit; the first is used to bias the transistors M1 and M2 into saturation; we chose a $V_g \approx 325mV$ for an $I_{DQ} \approx 420uA$. The second current mirror is used for current amplification.

The current at the output can be approximated by :

$$V_{out} \cong I_{DQ} \cdot B \cdot R_3 \tag{III-13}$$

Where I_{DQ} is the drain current of the transistors M1 and M2 as given by equation (III-11); and B the gain of the current mirror composed of the transistors P1 and P2; and R_3 the load resistance. This helps set the voltage gain $\propto {V_A}^2$; the load R_3 is chosen to match the output impedance (50Ω) over a broad bandwidth for measurements. The low R_3 value impacts the gain significantly.

The ratio B depends on the transistors P1 and P2 dimensions; where B is approximately given by:

$$B \cong \frac{(W/L)_{P2}}{(W/L)_{P1}} \tag{III-14}$$

Naturally, increasing the width of P2 while keeping the width of P1 constant will provide a high gain; however, by increasing its width, we increase its capacitances, resulting in a reduction of the current mirror bandwidth since, for a basic current mirror, the 3-dB bandwidth is approximated by :

$$\omega_0 = \frac{g_{mP1}}{C_{gsP1} + C_{gsP2}} \tag{III-15}$$

Thus, a compromise bandwidth/voltage gain/power consumption is required. In our case, B is equal to 8. The bandwidth of a current mirror can be improved by increasing the DC biasing current at the expense of higher power consumption. However, diverse simple techniques can improve a current mirror's bandwidth, too, such as resistive or inductive compensation techniques, as shown in Figure III.28.

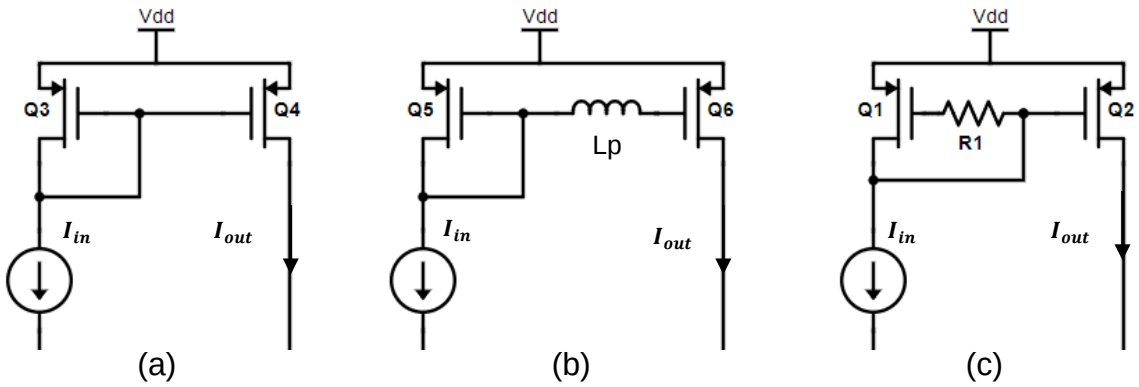


Figure III.28 : (a) Basic current mirror. (b) Current mirror with inductive compensation. (c) Current mirror with resistive compensation.

In the inductive series-peaking technique [11], an inductance is added between the current mirror transistors' gates to extend its bandwidth. This inductor is in series with C_{gs} of both MOSFETs; and the CM transfer function, in this case, is given by :

$$\frac{i_{out}(s)}{i_{in}(s)} = \frac{g_{mP2}/g_{mP1}}{\frac{s^3 Lp C_{gsP1} C_{gsP2}}{g_{mP1}} + s^2 Lp C_{gsP2} + \frac{s\left(C_{gsP1} + C_{gsP2}\right)}{g_{mP1}} + 1} \qquad \text{(III-16)}$$

If $C_{gsP2} \gg C_{gsP1}$ and $Lp = C_{gsP2}/(4g_{mP1})^2$, the bandwidth of the CM is approximately doubled compared to a basic CM as given by equation (III-17). The main inconvenience of this method is the large inductor area.

$$\omega_{0i} \approx \frac{g_{mP1}}{C_{gsP2}} \qquad \text{(III-17)}$$

A resistor is added between the MOSFETs' common gate terminals in the resistive compensation technique, as shown in Figure III.28.c. Adding this resistor can, in theory, double the CM bandwidth without impacting its DC behavior. The resistor adds a pair of pole and zero into the CM transfer function, it becomes:

$$\frac{i_{out}(s)}{i_{in}(s)} = \frac{g_{mP2}(sC_{gsP1}R + 1)}{s^2 C_{gsP1} C_{gsP2} R + s\left(C_{gsP1} + C_{gsP2}\right) + g_{mP1}} \qquad \text{(III-18)}$$

And its 3-dB bandwidth is approximated by :

$$\omega_{0r} = \sqrt{\frac{g_{mP1}}{RC_{gsP1}C_{gsP2}}} \qquad \text{(III-19)}$$

By appropriately choosing R, the dominant pole is canceled by the added zero due to the resistor's use. If $C_{gsP1} = C_{gsP2}$ and R is chosen to be equal to $1/g_{mP1}$ in equation ((III-17)), ω_{0r} is equal to g_{mP1}/C_{gsP1}. Thus, the bandwidth is doubled compared to a basic CM. The authors in [12] proposed to use a transistor whose channel resistance is equal to $1/g_{mP1}$ instead of the passive resistor. This adds extra circuitry to bias the added transistor correctly and adds another parameter to adjust the bandwidth appropriately; at the expense of extra power consumption.

Figure III.29 shows the current mirror gain in dB; we notice that the -3dB bandwidth for the basic current mirror is around 10 GHz, the bandwidth is extended by 6GHz for a current mirror with resistive compensation. However, an inadequate resistor value can cause frequency peaking, as shown for the 10kΩ resistor. Thus, the chosen 1.7kΩ is more appropriate for a flatter frequency response. Figure III.30 also shows an example of the resistor impact in time-domain, where we notice a faster rise/fall time when using the 1.7kΩ.

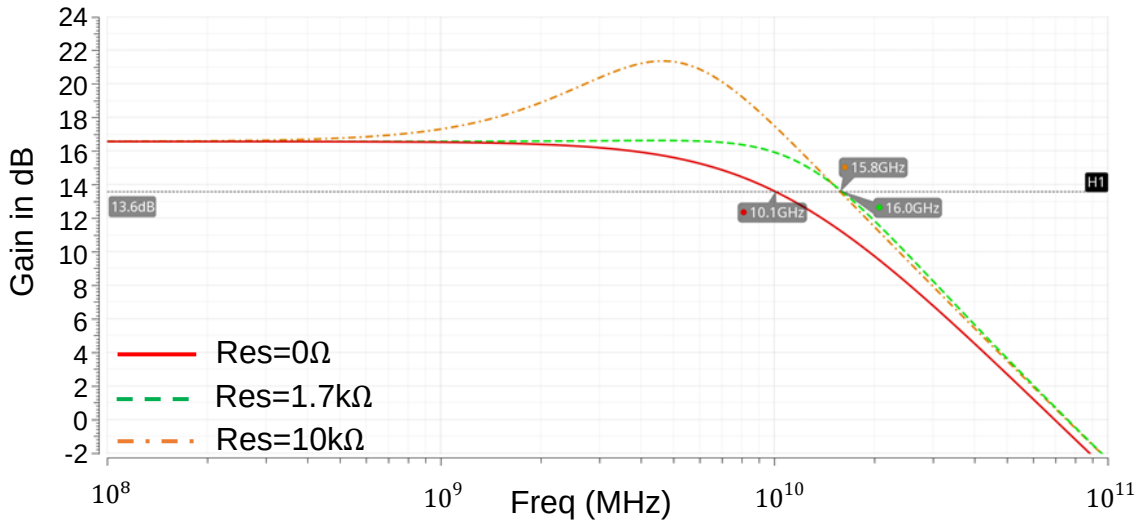


Figure III.29 : Impact of the resistor on the CM bandwidth.

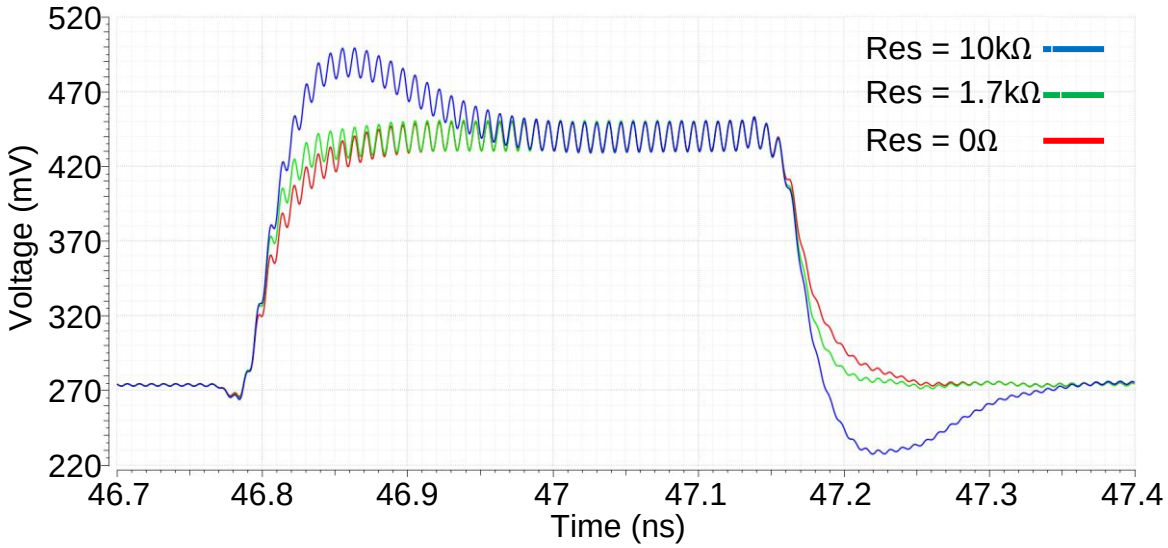


Figure III.30 : The CM resistor compensation technique impact on the output time-domain waveform.

We used a resistor R_3 (circuit shown in Figure III.24) equal to 130Ω; this resistor, along with P2 output impedance, allowed us to match the output to the 50Ω probe; Figure III.31 shows the output return loss. S_{11} is lower than -10dB and relatively stable in the band of interest. The output pads were placed at 28um distance from the circuit to measure the signal and its reference appropriately.

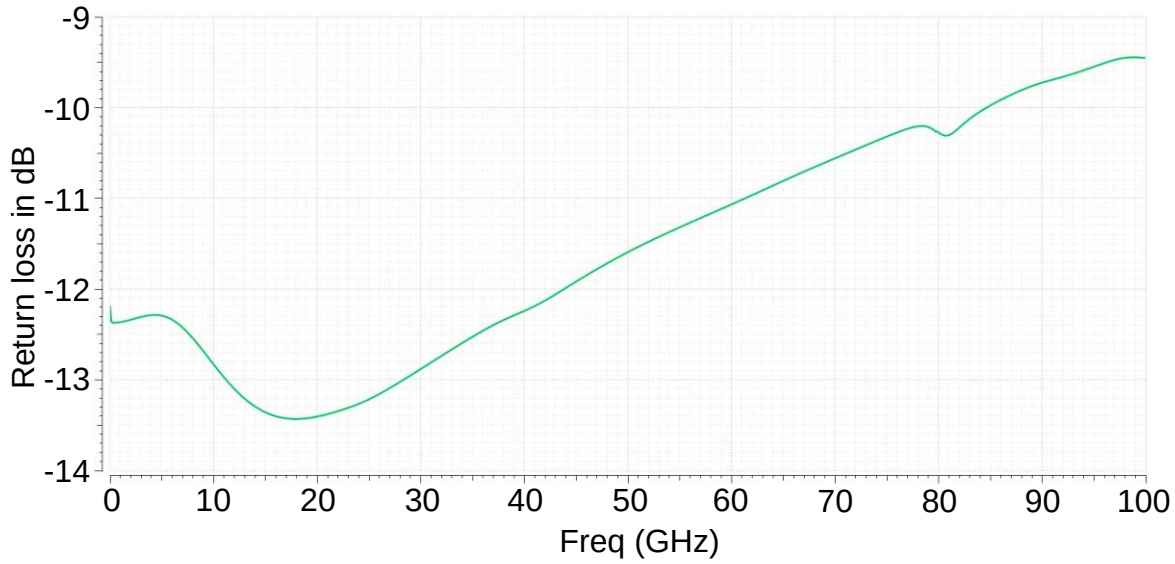


Figure III.31 : Output return loss in dB.

The receiver DC power consumption is evaluated at 8.7mW. It is worth mentioning that most of the power consumption is due to the 50Ω output impedance for measurements. In a real use case where high impedance digital gates (inverters and buffers) were used instead of a 50Ω matching, we reduced the receiver DC power consumption to around 2.5mW as described in the following section.

IV.3. 10Gbps transceiver simulation results

For a 10Gbps signal, around 500ps are required to modulate, transmit the signal, and demodulate it. This approach transmission delay depends on the data rate without considering the serialization and deserialization process. Hereafter in Figure III.32, the data clock cycle is described as the clock related to the data rate $T_{data\,rate}$. This data clock is different from the chip clock T_{CLK} (the chip clock is usually between 2-3GHz). Thus, one data clock cycle is $T_{data\,rate} = \frac{1}{f_{data\,rate}} = 100ps$.

For a distance lower than 10mm, the transmitter requires 4 data clock cycles ($4 \cdot T_{data\,rate} = 400ps$) to modulate the data using duobinary modulation correctly and then upconvert it to 60GHz. The propagation and demodulation require an extra $T_{data\,rate}$ data clock cycle. Thus, the total is $5 \cdot T_{data\,rate}$, as shown in Figure III.32, which shows the input and output data after demodulation. Accordingly, for a chip running at 2GHz, this propagation time is less than one clock cycle.

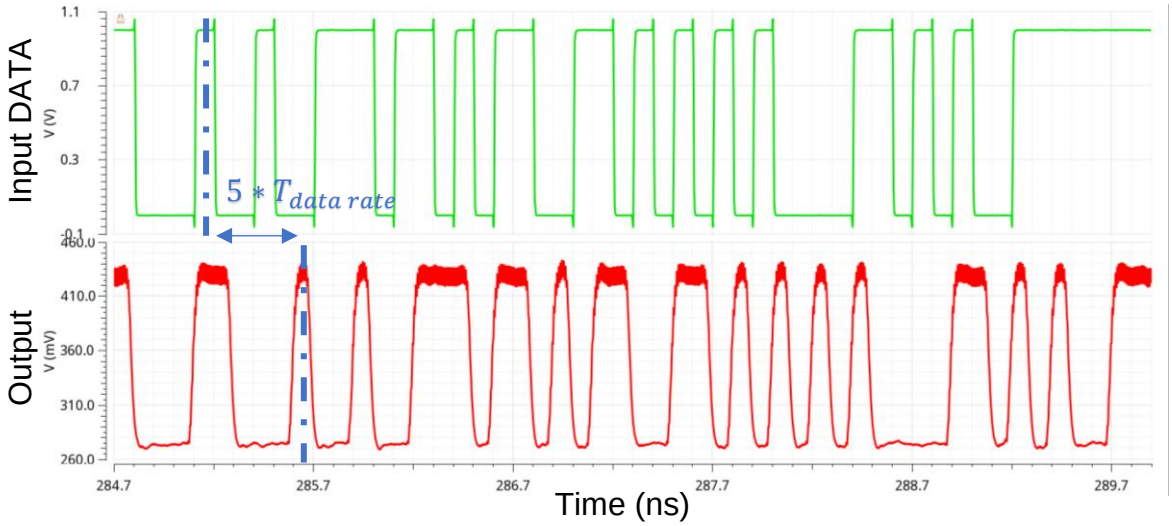


Figure III.32 : Simulated system input and output waveforms.

Figure III.33.a. shows the eye diagram of a 10Gbps data signal. The eye amplitude is around 153mV. We notice from Figure III.33 that the levels zero and one have different standard deviations; Level 1 has a more considerable variation ($\sigma_1 = 4.55mV$) due to the presence of the 120GHz frequency component not completely filtered. This variation degrades the SNR and should be filtered appropriately in the next works. However, the simulated SNR is around 28dB, 10dB higher than the targeted SNR (18dB).

The input signal has a 7ps rise/fall time, and 400fs RMS random jitter. The temperature was kept at 27°C for the simulation results shown hereafter. The eye diagram shows that the crossing eye percentage is not at 50%, but instead, it is around 37%. This is due to the output capacitance and resistance impacting this eye-crossing percentage, and it should be kept within a reasonable margin for different process corners. We observed that it stays within 30% to 70% for different process corners in this work.

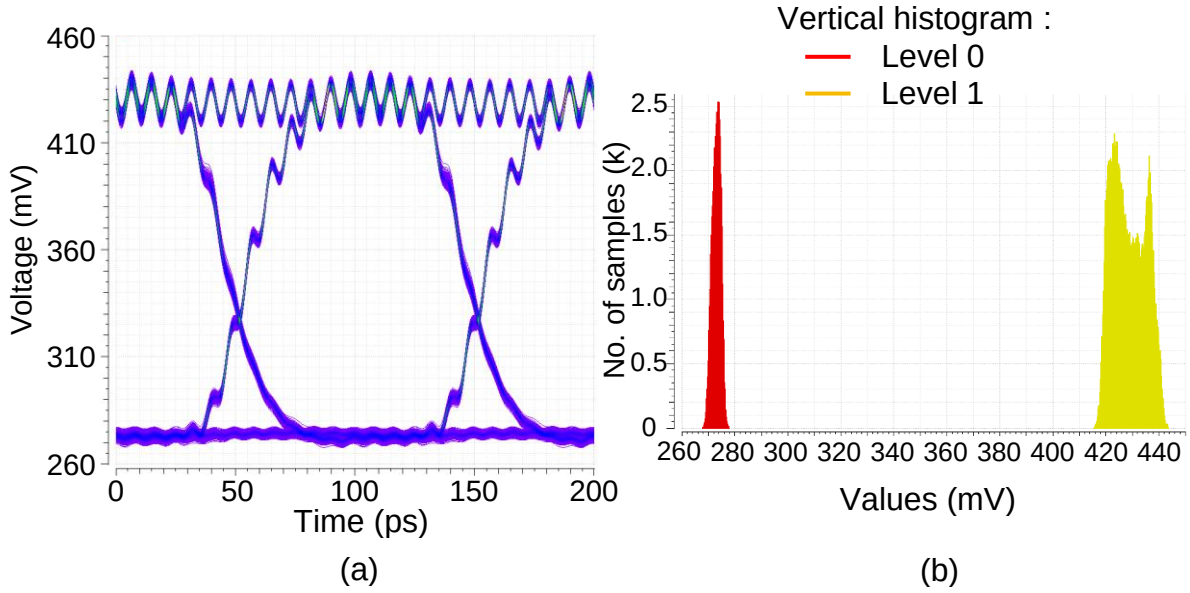


Figure III.33 : (a) 10Gbps Eye diagram. (b) Eye diagram vertical histogram.

The simulations resulted in a 1.25ps Deterministic jitter and 287fs random jitter per side (left and right) at the output. The maximum eye width is achieved at the crossing point (around 329mV) and is equal to 97.38ps, as shown in Figure III.34. The 10Gbps bathtub curve shows that the eye-opening is significantly larger than required for a BER of 10^{-12}.

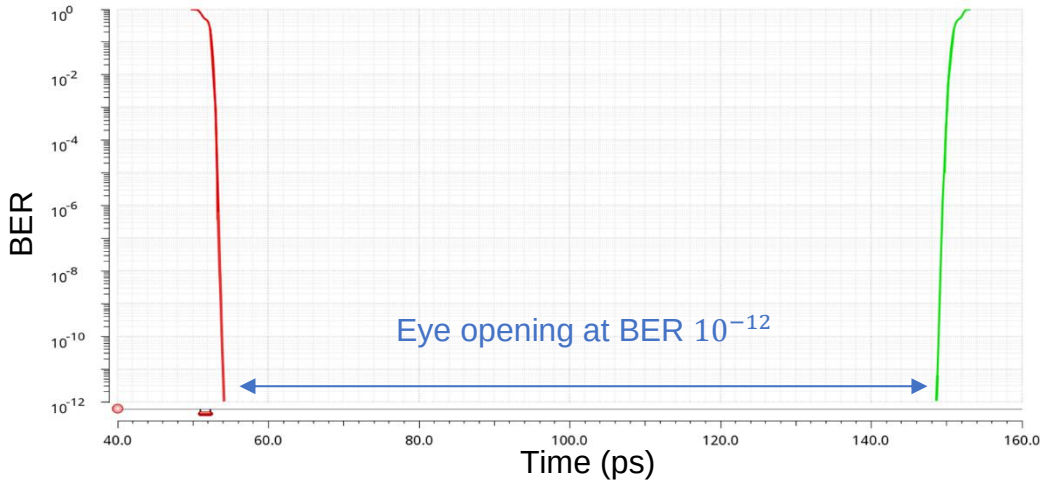


Figure III.34 : 10Gbps Bathtub curve.

The transmitter (without VCO) and receiver implementation areas are 418um² and 15643um², respectively. The receiver has a large area due to the use of inductors as shown in Figure III.35.a. Its core area, shown in Figure III.35.b, (without inductors) is around 376um².

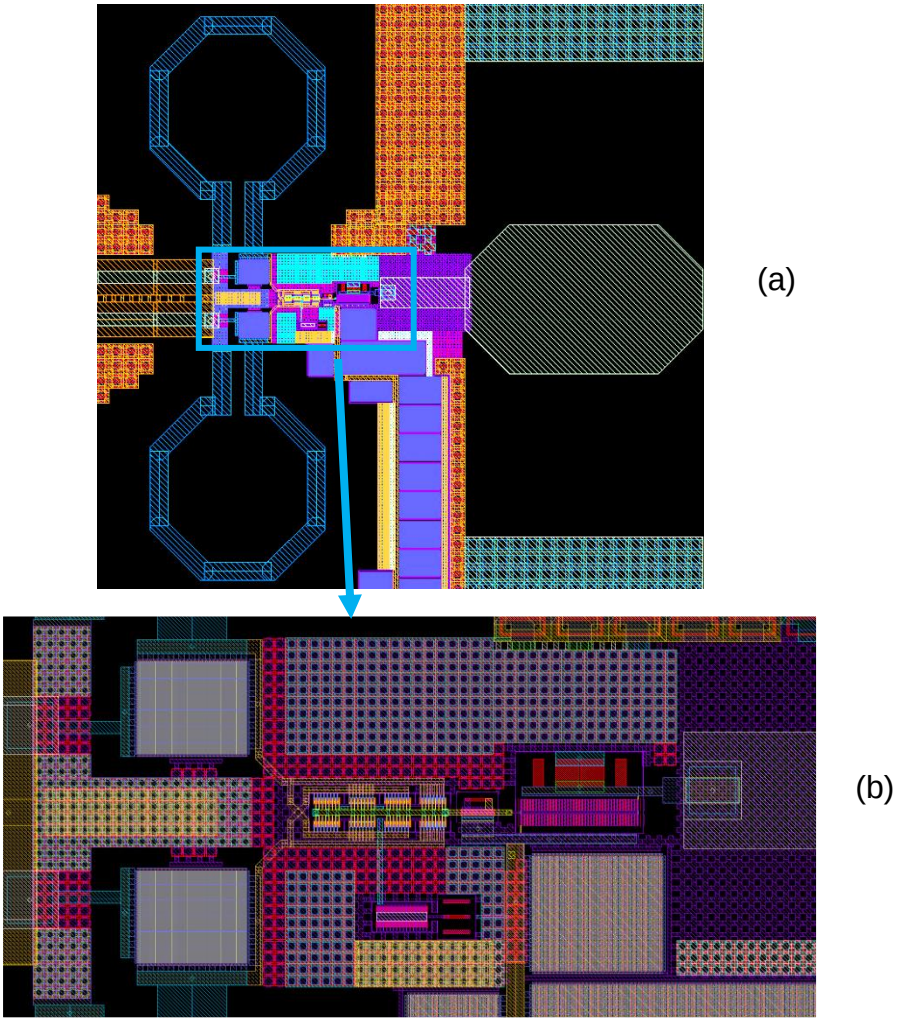


Figure III.35 : (a) Layout of the receiver (not all the layers are included). (b) The core of the receiver(not all the layers are included).

V. 14Gbps system simulation results

In this work, we designed and optimized the transceiver for a 10Gbps data rate across corners and PVT variations and a BER lower than 10^{-12}. However, in our simulations, we found that our previous design is robust up to 14Gbps; the input injected data has a rise/fall time of 8ps and 400fs random jitter. Figure III.36 shows the time-domain signal and the eye diagram of a 14Gbps data signal. The SNR is around 20dB; This is mainly due to the increase of the amplitude variation at level one, with a standard variation σ_1 of 8.2mV. We notice that the rise/fall time is degraded with a 22ps rise time, and a 11ps fall time, and the jitter presence is increased.

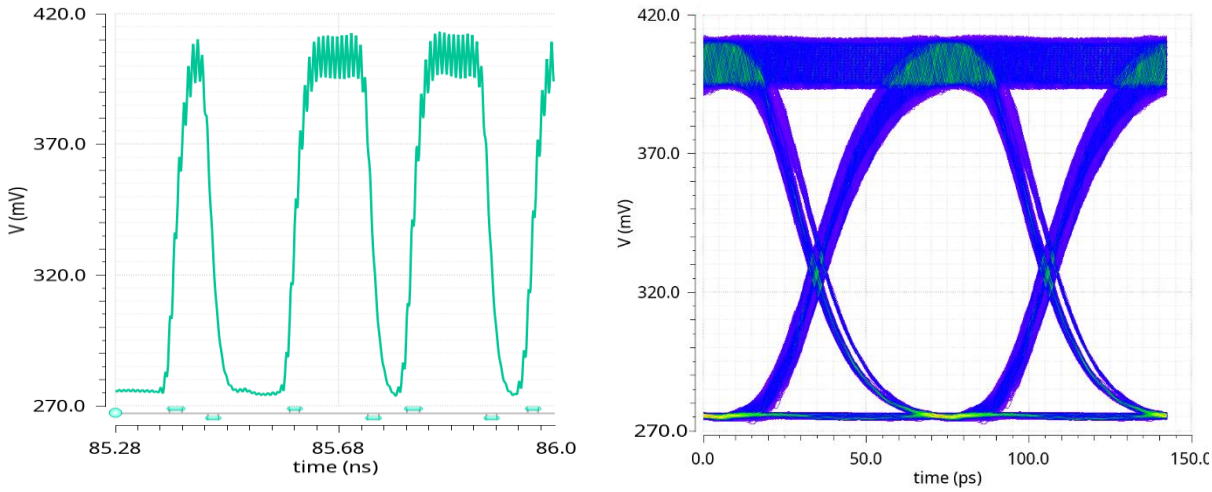


Figure III.36 : (a) Time-domain output signal. (b) Output eye diagram.

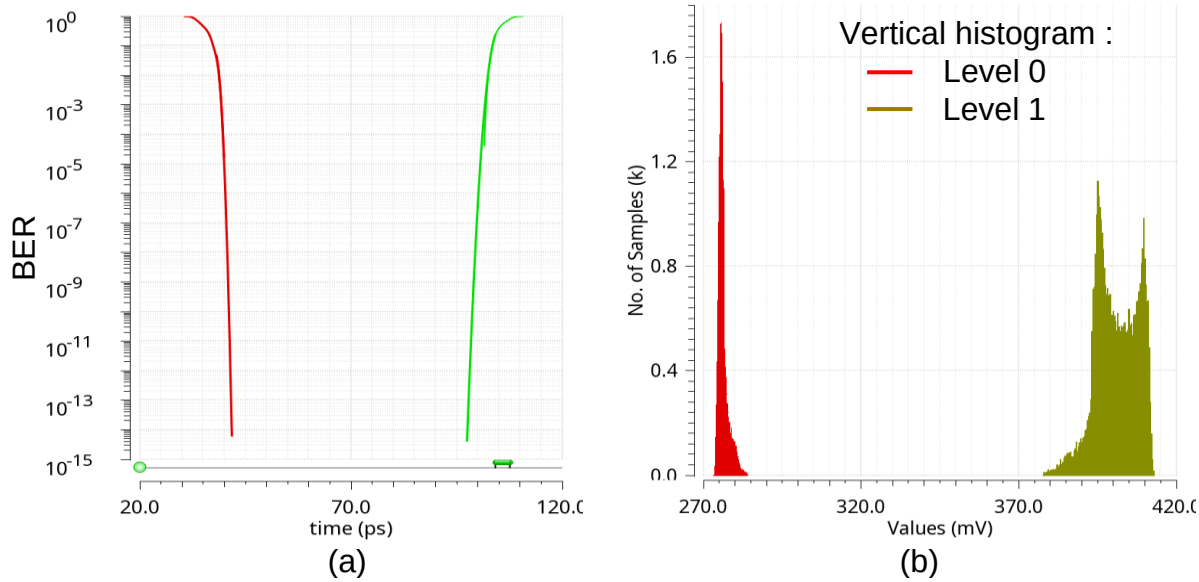


Figure III.37 : (a) 14Gbps bathtub curve. (b) Eye diagram vertical histogram.

The resulting deterministic jitter is 3.2ps, and the random jitter was evaluated at 990fs at each side of the output. The bathtub curve in Figure III.37 shows that the horizontal eye-opening at a BER of 10^{-12} is around 56.3ps. However, we believe that more work is required, especially on the duobinary modulator, to qualify this design as robust for a 14 Gbps signal through the PVT variations.

VI. Application to digital applications

As mentioned previously, the prototype chip was designed for measurement purposes, i.e., with a 50Ω output impedance. Hereafter, we discuss buffers' use at the output instead of a 50Ω output impedance, for more practical use within digital circuits, The modified demodulator is shown in Figure III.38. The transmitter was not modified. Several parameters were changed; for example, we reduced the width of the P2' transistor; we also increased the resistor R3' value to bias the first buffer appropriately.

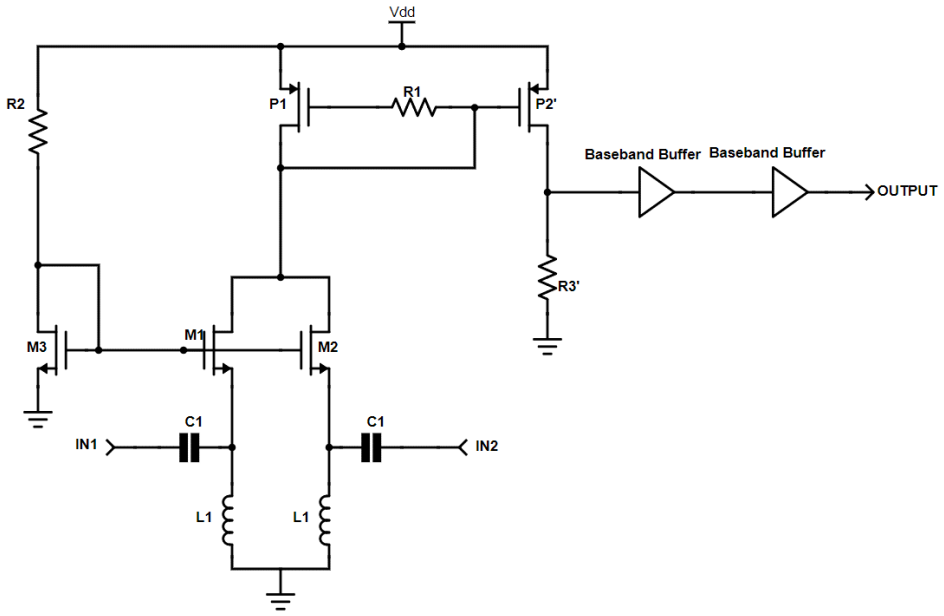


Figure III.38 : Schematic of the demodulator for digital applications.

The receiver DC power consumption was also reduced from around 8.7mW for a 50Ω output, to around 2.5mW when buffers are driven instead. The time-domain signal at the output of buffers is shown in Figure III.39.a. The eye diagram of a 14Gbps signal, shown in Figure III.39.b, indicates that the rise/fall time is around 6ps with a deterministic jitter of 4.8ps and a random jitter of around 350fs per side. The bathtub for this case is shown in Figure III.39.c, and it shows that the horizontal eye-opening is around 61ps. However, in this case, the eye-crossing percentage is sensitive to the PVT variations. Synchronization at the output is a more suitable approach.

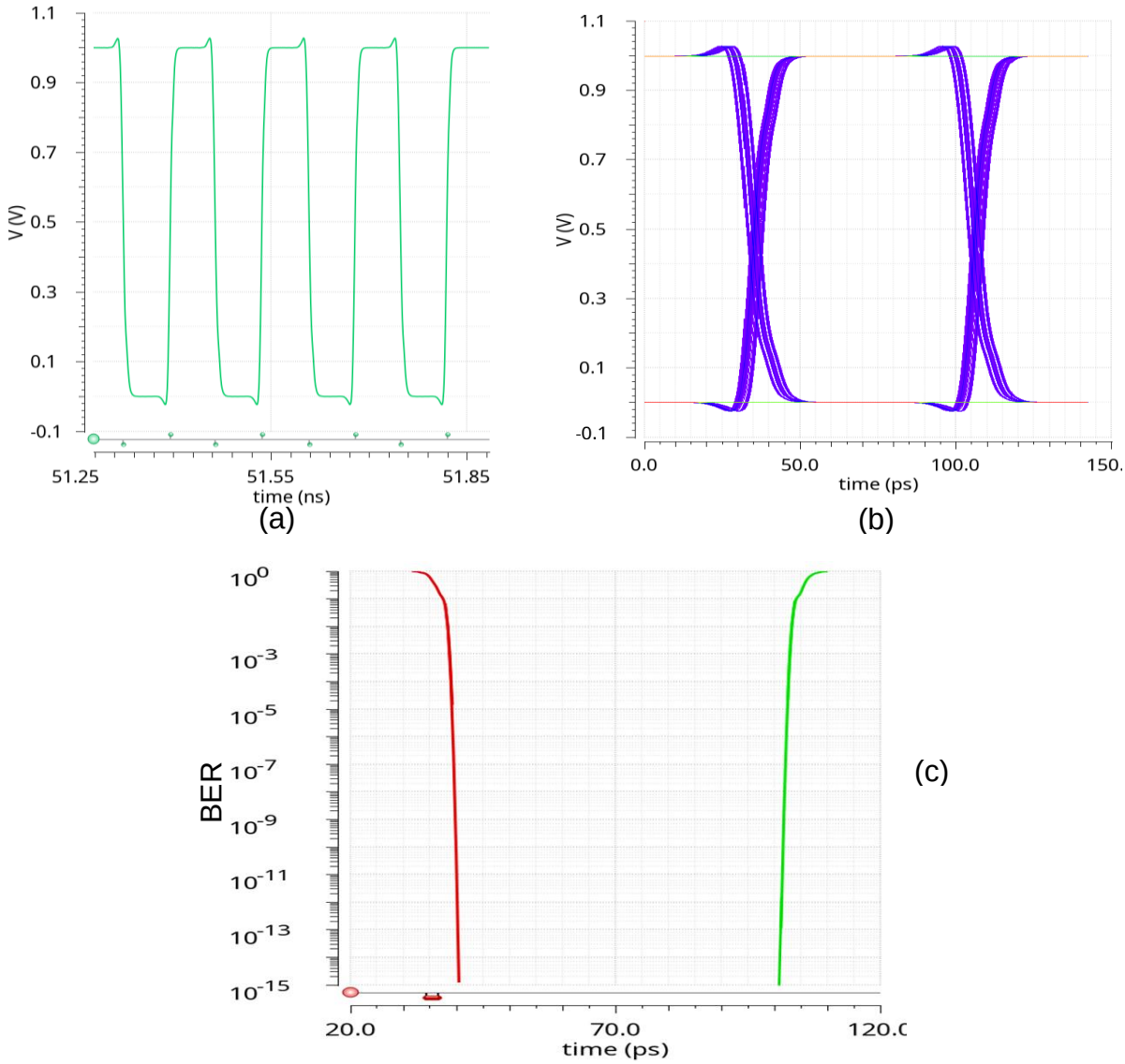


Figure III.39 : (a) Time-domain output signal. (b) Output eye diagram. (c) Bathtub curve.

VII. State of the art comparison

In this work, we managed to reach a 10Gbps data rate (and higher) with a BER lower than 10^{-12}; using RF transmission only; Contrarily to the state of the art where they combine both baseband and RF transmission to reach high data rates.

The authors in [13] used single-ended signaling to transmit a 14.4Gbps total data rate by combining a 10Gbps PAM4 signal in baseband with a 4.4Gbps ASK transmission using a 20GHz carrier. They managed to achieve a BER lower than 10^{-15}, and an energy efficiency of 1.2pJ/bit for the baseband transmission and 1.045 pJ/bit for RF transmission. This value does not take into consideration the VCO power consumption.

While in [14], they reached a 10Gbps total data rate by combining a 2Gbps baseband signal with two 4Gbps RF ASK signals at 28.8GHz and 49.5GHz. They reached a BER of 10^{-9}. They transmitted the signal over 5mm differential lines. The authors used a 90nm technology from IBM; they evaluated the energy efficiency at 0.6pJ/bit for baseband transmission and 0.45pJ/bit for RF bands; without including the

two VCOs' power consumption. It is worth mentioning that the authors did not specify the output buffer power consumption.

	[13]		[14]			[15]	This work : 50Ω output impedance	This work : Buffers at the output
Technology	65nm CMOS		90nm CMOS			65nm CMOS	28nm FD-SOI	28nm FD-SOI
Frequency Band	BB	20 GHz	BB	28.8 GHz	49.5 GHz	60 GHz	60 GHz	60 GHz
Data rate per band (Gbps)	10	4.4	2	4	4	5	10	14
Length (mm)	5		5			5.5	4.6	4.6
Modulation	PAM4	ASK	NA	ASK	ASK	ASK	Duobinary	Duobinary
Signaling type	Single-ended		Differential			Differential	Differential	Differential
Data rate (Gbps)	14.4		10			5	10	14
Energy efficiency (pJ/bit)°	1.2	*1.04	0.6	*0.45	*0.45	*1.33	*1.03	*0.25
BER	10^{-15}		10^{-9}			$<10^{-12}$	$<10^{-12}$	$<10^{-12}$
Area	NA		NA			NA	TX : 418 um² RX:0.0156 mm²	NA
NA : Not available. ***VCO Power consumption is not included.**								

Table III.1 : Comparison of our simulation results to the state of the art.

The previous works used different frequency bands to transmit a higher data rate signal; To separate the signals, they couple the signals inductively using a transformer; However, this transformer requires a large area. Using a single band to transmit the signal avoids using this transformer, as shown in [15]; In this case, the authors transmitted a 5Gbps signal using ASK modulation at 60GHz over a differential 5.5mm transmission lines. Their chip was realized using a TSMC 65nm CMOS process. They achieved a BER lower than 10^{-12} and an energy efficiency of 1.32pJ/bit (the VCO power consumption is not included).

In this work, we managed to achieve higher data rates (10Gbps and higher) by using an efficient modulation instead of ASK modulation; thus, we compressed the signal's spectrum at the input of the receiver by using the duobinary modulation. This relaxes the receiver bandwidth constraints, where the available bandwidth at the receiver's input is limited due to the components used. This bottleneck was unblocked in this case without increasing the complexity or the area of the system significantly; the extra duobinary modulator area was evaluated at only 14,4um x 3.3um. Furthermore, as shown in Table III., this work's energy efficiency was evaluated at 1.03pJ/bit for a 10Gbps signal with a 50Ω output, and 0.25pJ/bit for a 14Gbps with buffers at the output. The energy efficiency FoM given in pJ/bit is comparable to other works even though we have doubled the data rate compared to the other works. Table III. compares our work to the previous more related works.

VIII. Conclusion

In this chapter, we briefly introduced the 28nm FD-SOI technology used in this work. We then explained our channel choice and its ground plane positioning, where we choose to place the ground plane at a higher metal layer to reduce the routing constraints below the transmission lines.

Afterward, we presented the proposed transmitter and receiver and their performances using the 28nm FD-SOI technology. We also presented the whole system results at room temperature using a 10Gbps signal; the simulation results shown included the post-layout parasitics, and jitter was injected at the input. The system proved robust for PVT variations at 10Gbps, but we believe it still requires more work at 14Gbps.

We also presented the performances for more practical applications, i.e., using digital buffers at the output instead of a 50Ω output impedance for the measurements. Finally, we compared this work to previous works; we showed that we managed to double the data rate compared to state of the art for RF transmission by using the duobinary modulation while keeping the energy efficiency, complexity, and area usage acceptable.

IX. References

[1] A. Cathelin, “Fully Depleted Silicon on Insulator Devices CMOS: The 28-nm Node Is the Perfect Technology for Analog, RF, mmW, and Mixed-Signal System-on-Chip Integration,” *IEEE Solid-State Circuits Mag.*, vol. 9, no. 4, pp. 18–26, 2017.

[2] F. Voineau, “Systèmes communicants haut-débit et bas coûts par guide d’ondes en plastique,” PhD Thesis, 2018.

[3] “Momentum 3D Planar EM Simulator.” https://www.keysight.com/en/pc-1887116/momentum-3d-planar-em-simulator?nid=-33748.0.00&cc=FR&lc=fre.

[4] W. Bae, “CMOS Inverter as Analog Circuit: An Overview,” *J. Low Power Electron. Appl.*, vol. 9, p. 26, 2019, doi: 10.3390/jlpea9030026.

[5] W. Bae, H. Ju, K. Park, S. Cho, and D. Jeong, “A 7.6 mW, 414 fs RMS-Jitter 10 GHz Phase-Locked Loop for a 40 Gb/s Serial Link Transmitter Based on a Two-Stage Ring Oscillator in 65 nm CMOS,” *IEEE J. Solid-State Circuits*, vol. 51, no. 10, pp. 2357–2367, 2016.

[6] B. Min, K. Lee, and S. Palermo, “A 20Gb/s triple-mode (PAM-2, PAM-4, and duobinary) transmitter,” *Microelectron. J.*, vol. 43, no. 10, pp. 687–696, 2012.

[7] Y.-L. Chen, C. Kao, P.-J. Peng, and J. Lee, “A 94GHz Duobinary Keying Wireless Transceiver in 65nm CMOS,” *Symp. VLSI Circuits Dig. Tech. Pap.*, p. 2, 2014.

[8] K. Komoni, S. Sonkusale, and G. Dawe, “Fundamental performance limits and scaling of a CMOS passive double-balanced mixer,” in *2008 joint 6th international IEEE northeast workshop on circuits and systems and TAISA conference*, 2008, pp. 297–300.

[9] M. Ercoli, M. Kraemer, D. Dragomirescu, and R. Plana, “A passive mixer for 60 GHz applications in CMOS 65nm technology,” in *German Microwave Conference Digest of Papers*, 2010, pp. 20–23.

[10] B. Leite, “Design and modeling of mm-wave integrated transformers in CMOS and BiCMOS technologies,” PhD Thesis, 2011.

[11] B. Sun and F. Yuan, “A New Inductor Series-Peaking Technique for Bandwidth Enhancement of CMOS Current-Mode Circuits,” *Analog Integr. Circuits Signal Process.*, vol. 37, no. 3, pp. 259–264, Dec. 2003.

[12] T. Voo and C. Toumazou, “High-speed current mirror resistive compensation technique,” *Electron. Lett.*, vol. 31, no. 4, pp. 248–250, 1995.

[13] M. Jalalifar and G. Byun, “A 14.4Gb/s/pin 230fJ/b/pin/mm multi-level RF-interconnect for global network-on-chip communication,” in *2016 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2016, pp. 97–100.

[14] S.-W. Tam, E. Socher, A. Wong, and M.-C. F. Chang, “A simultaneous tri-band on-chip RF-interconnect for future network-on-chip,” in *2009 symposium on VLSI circuits*, 2009, pp. 90–91.

[15] H. Wu *et al.*, “A 60GHz on-chip RF-interconnect with λ/4 coupler for 5Gbps bi-directional communication and multi-drop arbitration,” in *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*, 2012, pp. 1–4.

Conclusion and Perspectives

Conclusion

In this work, we worked on on-chip interconnects, and more specifically, long-range interconnects for Network on chips (NoCs).

In Chapter I, we started by introducing the standard as well as the emerging on-chip interconnects, such as the wireless solution (WiNoC), the optical solution (ONoC), and the radiofrequency interconnect (RF-I). We discussed the different emerging on-chip interconnects and presented some of their advantages and challenges.

Next, we studied the standard wires and their limitations for long-range links, namely the high latency. To mitigate this issue, we proposed to use the radiofrequency solution (RF-I); this approach transmits the signal in the LC region (at high frequencies) of the transmission lines, where the latency is minimal due to limited dispersion. Moreover, The RF-Interconnect state of the art is presented; it offers a promising opportunity for high data rates serial links. We noticed that all works presented are based on the same type of modulation: ASK modulation.

Thus, in chapter II, we introduced additional (simple) modulations and summarized the advantages' duobinary modulation. The main one is the reduction of the used spectrum by a factor of two compared to classical NRZ modulation; In other words, we double the data rate for the same bandwidth. More importantly, this spectrum compression translates to lower constraints on the RF components (linearity in particular). Furthermore, the demodulation process is as simple as simple ASK modulation.

We proposed the system architecture and showed that in theory, this approach could be easily length adaptive, i.e., the same transmitter and receiver can be used for different interconnect lengths without redesign; this makes this approach very attractive for digital integration.

Chapter III is dedicated to the design of the different components: the transmitter, the channel, and receiver using the 28 nm FD-SOI technology from STMicroelectronics. Firstly, we presented our choice of Edge-coupled microstrip lines and their ground placement for reduced floorplanning constraints; the channel simulation results are shown next. Afterward, we showed the 10 Gbps transmitter and receiver simulation results; We showed that although the designed system was optimized for a 10 Gbps signal, it is robust up to 14 Gbps with a BER lower than 10^{-12}. It is worth mentioning that this prototype was designed for measurement purposes with a 50Ω output impedance. This added certain constraints that are absent for more practical cases, e.g., when high impedance digital gates/buffers are placed at the output instead. The latter case was simulated in the third chapter, and we evaluated the energy efficiency for a 14 Gbps signal at 0.25 pJ/bit with a BER lower than 10^{-12}. This performance overpasses the state of the art, that corresponds to a maximal 5 Gbps data rate.

In the state of the art works, we found works that combined both Baseband and RF transmission to reach higher data rates. However, this approach does not take full advantage of the RF bandwidth; and does not lead to higher data rate.

Furthermore, the combined baseband and RF interconnects have two different latencies (baseband transmission is much slower than the RF one), which adds certain complexity. The simulated energy efficiency of 1.03pJ/bit for a 50Ω output is comparable to similar works even though we doubled the data rate, and is lower than all similar works when digital gates are placed at the output. It is worth mentioning that the output impedance and whether the output buffers power consumption were included in the state of the art works.

A patent on this innovative high data rate structure of electrical serial links was filled. The prototype chip is shown in Figure A, and its measurement results will be published in the near future.

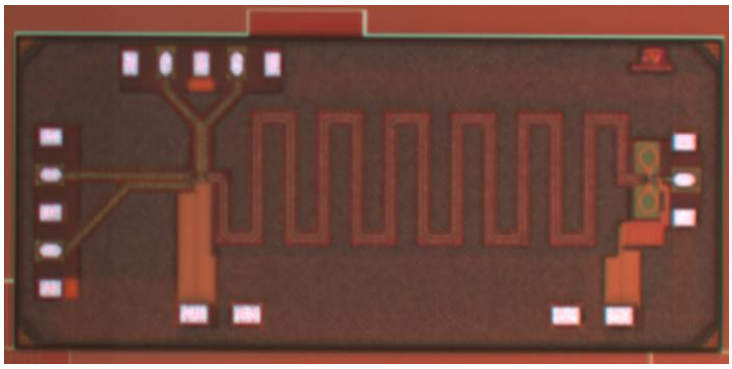


Figure A : Prototype chip in 28nm FD-SOI technology.

Perspectives

We presented our work on an on-chip interconnections solution for long-range serial links. We showed the possibility of transmitting the signal with high reliability over several millimeters in the same chip to connect, for example, its edges with minimal latency.

We believe that this work can be used as a groundwork for several improvements. For example, we believe that the data rate can be improved significantly by modifying the design, i.e., using faster gates for the duobinary modulator and improving the receiver output bandwidth; we expect this will allow this RF-approach to surpass the 20Gbps cap for on-chip interconnects.

We also believe that the same approach/design can be used for a slightly more different application, where we proposed in [1] to use the same approach to transmit a 10Gbps signal over a 7-mm interposer channel for extra dies serial links.

Nowadays, 3D/2.5D IC implementation is a genuine prospect and is becoming essential to achieving better performance. 2.5D/3D IC integration promises to go beyond Moore's Law. From a manufacturing point of view [2], incorporating different dies rather than designing one large chip ensures a better yield and lower cost by testing each die independently before assembly. It also guarantees a shorter time to market; since some chips can be reused without redesign. Designers also gain

flexibility in developing new architectures and prototyping; they can focus on their strength and complete their chips with "off the shelf" dies (CPUs, memories, RF transceivers...) from specialized companies. These dies may come from different manufacturing technologies (CMOS, BiCMOS, SOI …) and nodes.

Nowadays, researches developed several 2.5D/3D techniques using different materials. These techniques usually rely on interposers.

The interposer shown in Figure B is a substrate shared by different dies. The dies are placed next to each other and are connected through the interposer; also known as side-by-side mounting. Thus, the interposer's main role is to conduct the signal from one chip to another chip. It can also convey the signal to a wider pitch µbump via Through silicon vias (TSVs), as shown in Figure B.

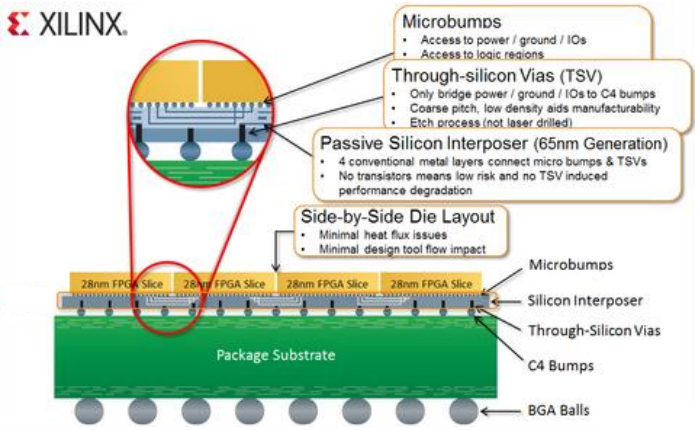


Figure B : The Xilinx 2.5D FPGA Product.

Interposers are very common for packaging applications, and we can distinguish three types: organic, glass, and silicon.

The organic interposer is considered the cheapest of the three options due to its low-cost and lower complexity (wet etching) fabrication process compared to silicon or glass processes. Unfortunately, organic interposers suffer from lower inputs/outputs density due to a larger pitch, limiting their prospective applications.

Glass interposers are the second contender. They offer higher interconnects density, act as a good isolator, and have an excellent loss for high-frequency RF applications and provide optical links implementation capability. However, they suffer from several issues during processing, such as surface defects and localized cracking. The thermal conductivity is poorer than silicon interposer, which leads to substrate heating. As a matter of fact, poor thermal dissipation is one of the interposer approach bottlenecks, where substrate heating leads to significant performance degradation.

Among the three options, silicon interposers are the most expensive. They are used in several high-performance chips such as server chips, for instance. They offer high interconnections density due to their fine pitch since they are fabricated using the typical CMOS (BiCMOS...) nodes. Hence, they offer a pitch with a sub-micrometer resolution. However, alignment mismatch is still a severe issue in the fabrication process. In a silicon interposer, the metal layers are facing upward, while the chiplets metal layers are facing downward; and the two sides are connected using a µbump. However, for silicon interposers, they are usually called µbump due to the pitch being in the um range; in this work, the pitch was fixed by the foundry technology and is equal

to 20um. A more detailed comparison of interposer technologies can be found in [3]. However, only silicon interposers will be considered in this work.

One clear advantage of putting the dies closer to each other [4], either in 2.5D or 3D integration, is reducing the communication distances. This results in a significant reduction of power required to drive the die's inputs/outputs [5]. Furthermore, it reduces transmission latency through wires. It also allows for higher data rates by decreasing the wire's bandwidth limiting factors, e.g., total resistance and capacitance.

In this work, we propose to reduce the latency even more, especially for long-range links (several millimeters). As explained in chapter II, RF interconnects provide a delay linear with the distance instead of a quadratic one for standard baseband wires. We believe the proposed approach is more attractive to connect, with very low latency (around 7 ps/mm), chiplets that are several millimeters away from each other.

This approach can be applied to both passive and active silicon interposers; passive interposers only include metal layers (the back end of line) used for interconnections and passive devices. They also may be used for passive equalizers [6]. Thus, they have a higher yield and cost much less than their active counterparts. A case study is presented in [7]. Active interposers include transistors and require the use of the full fabrication processes. The transistors can be used, for instance, to handle the power resources of the dies, along with the routers required for an extra layer of interconnects. It is worth mentioning that an interposer's area utilization is much lower than that of a chiplet.

For signal propagation, the main difference between passive and active interposers is the repeaters' implementation. As a reminder, a wire exhibits a quadratic delay with distance, resulting in high latency and slower interconnects. To mitigate these issues, repeaters are inserted periodically to transform a long link with a quadratic delay function into small linear delay subsections, as explained in Chapter I.

Repeaters insertion is not possible in passive interposers, thus, the repeaters are implemented in the chiplets. Alternatively, only the edges of the chiplets are connected, as shown in Figure C.2 and Figure C.3. This approach does not take advantage of the available metals in the interposer.

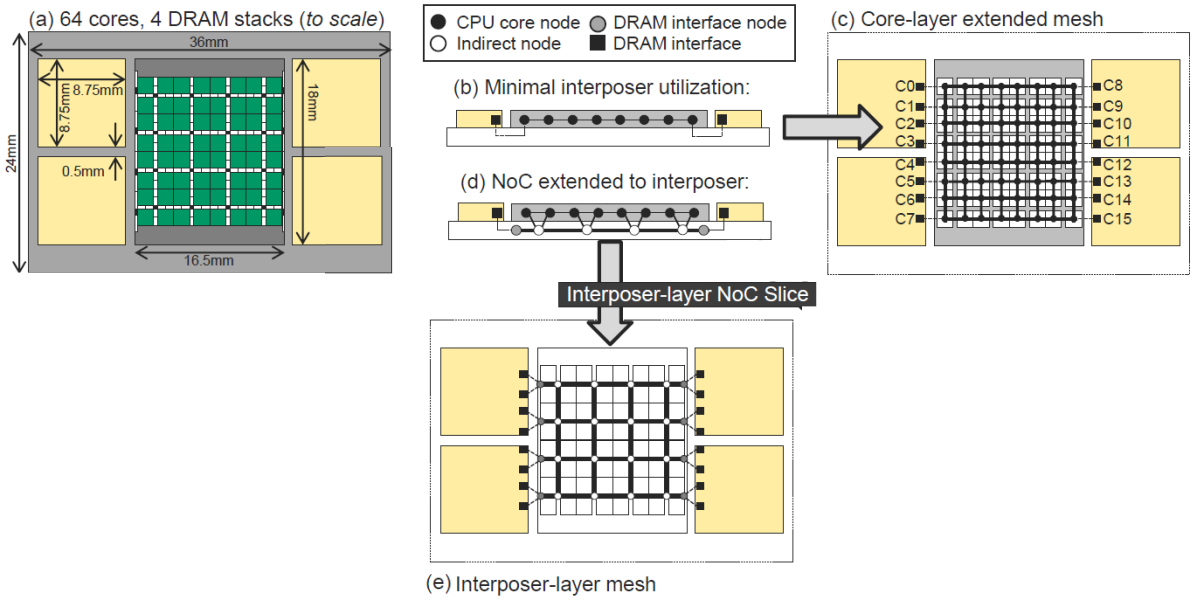


Figure C : (1) Top view of the 2.5D multi-core system with a 64-core CPU chip in the center with four DRAM stackes placed on either side of the multi-core die. (2) Side view of a simple interconnect implementation minimizing usage of the interposer. (3) The multi-core NoC slice uses a mesh topology. (4) Side view of a NoC logically partitioned across both the multi-core die and an interposer. (5) Interposer-layer NoC mesh. [4]

However, when using an active interposer, the repeaters (and routers) can be moved into the interposer, reducing the total link delay and adding an extra layer of interconnections to the NoC and excluding the router's area from the total chiplet area, as shown in Figure C.4 and Figure C.5.

The proposed approach can be implemented in both passive and active interposers, as shown in Figure D, where two chiplets 7-mm apart are connected. In the first case, we connect the two chiplets using the BEOL of the passive interposer, thus implementing the transmitter and receiver in the chiplets.

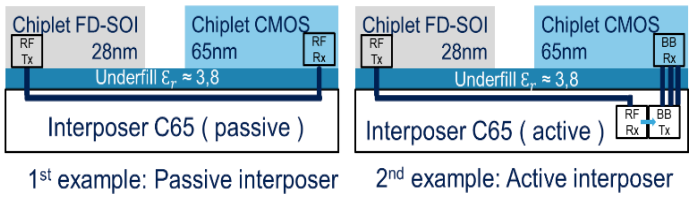


Figure D : Examples of possible system applications.

In the second case, we implement the transmitter (or receiver) in a chiplet. In contrast, the receiver (or transmitter) is implemented in the active interposer and then transmits the data in Baseband (BB) after demodulation into the chiplet. The second approach's advantage is the lower surface needed on the chiplet since the RF receiver will be implemented in the interposer. Furthermore, the TX (or RX) can be implemented in the interposer process, which is usually cheaper than the chiplet process. In the presented work, only the first case is demonstrated, the second being under investigation.

An interposer channel is very similar to the previously studied channel in chapters II and III, except for the µbump used to connect the two sides. Thus, we propose to use the same approach demonstrated for on-chip links to silicon interposers. We propose using duobinary modulation modulate a 10Gbps data, transmit it at higher frequencies in the channel's LC region, and significantly reduce the latency. Another advantage of this approach is its lower sensitivity to the channel distortion due mainly to the variation of group delay at low frequencies, Figure E represents an example of a wires group delay; it shows the prompt variation up to 3-4 GHz, equalization is usually used to compensate for these effects. We notice that the group delay is more stable at higher frequencies.

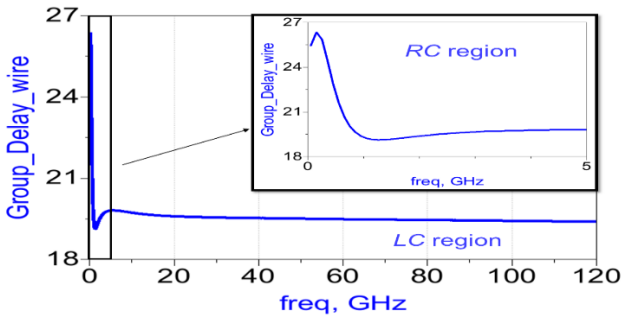


Figure E : Group delay of a wire.

A downside of using high frequency such as 60 GHz is associated to metallic losses due to the skin effect, which is significant. To reduce the strips' resistance, a width of several µm is required; and hence limiting the density. We used differential signaling for the same reasons explained in Chapter I and II.

We first studied the different components of the channel separately, then all together. For this study, we used EMPro 3D FEM tool from Keysight.

We used 10ML 28-nm FD-SOI technology for the chiplets, and BiCMOS 130 nm back-end-of-line for the interposer channel, both from STMicroelectronics. We used the full BEOL stack definition from the foundry, i.e., no approximation in dielectrics or metal conductivity was used, along with the appropriate underfill properties.

We decided to keep the chiplets interconnects as short as possible to have a minimal effect on the chiplet floorplan and surface consumption. Thus, we used 60–70 µm long transmission lines, before passing through the µbump and then through the interposer. The signal strip was implemented in M10, its ground in M9 instead of M1, to reduce its effect on the overall routing of the chiplet. With the distance of the signal strip to ground fixed, the remaining parameters for characteristic impedance tuning are the strips' width and the spacing between them.

We evaluated the insertion loss for differential propagation at 0.07dB at 60 GHz, as described in the simulation results shown in Figure F.

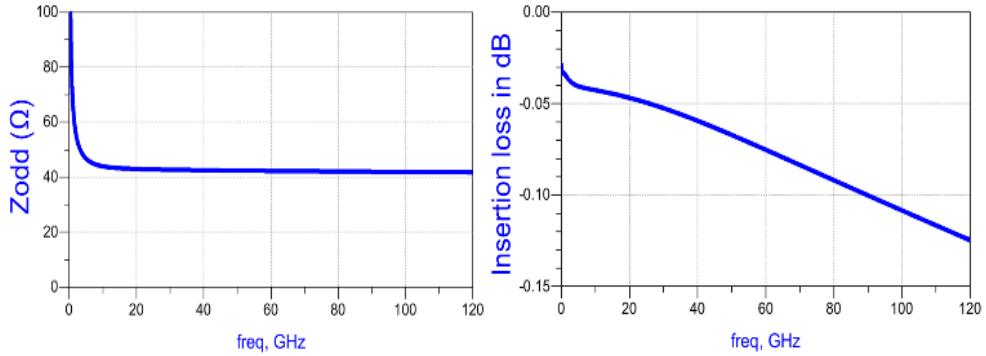


Figure F : Differential characteristic impedance Z_{odd} and insertion loss of the 28-nm FD-SOI chiplet transmission line.

Minimum strips width was used to respect the design rules and certain layout constraints. The odd-mode characteristic impedance (that must be considered for differential mode) is shown in Figure F to be 42.3 Ω at 60 GHz. Higher characteristic impedance could not be achieved with our configuration. Due to the small length of these transmission lines, the overall return loss's impact is negligible.

µbumps are used to connect the chiplet to the interposer as shown in Figure G.b. We simulated the µbump shown in Figure G.a following the available process at STMicroelectronics, which is defined as a stack of different materials with different heights such as copper, Nickel and SnAg, with SnAg having the lowest conductivity. The µbump height is 12 µm, and its diameter is 10 µm. The pitch is 20 µm.

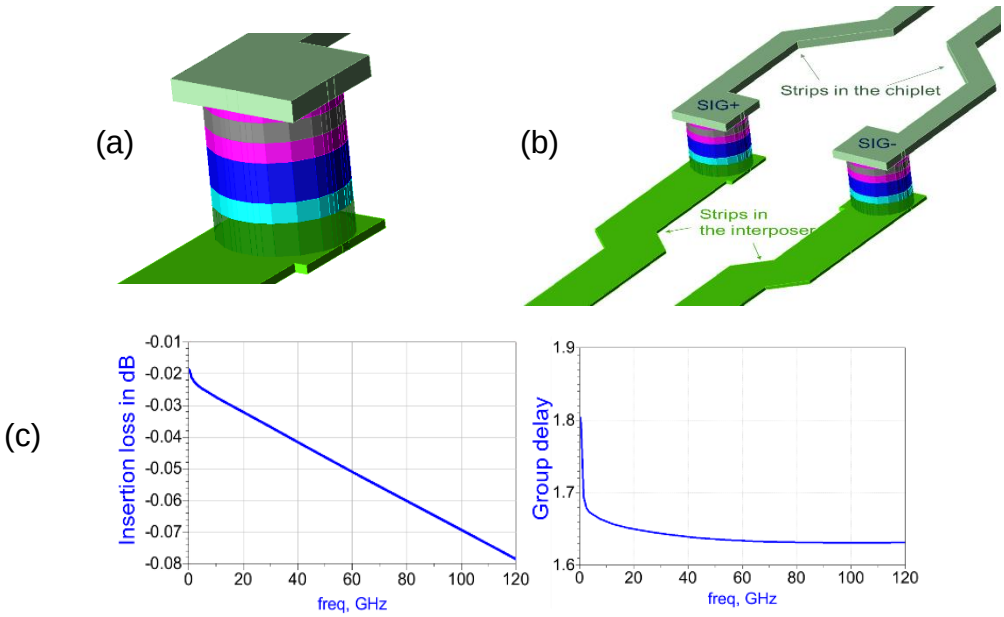


Figure G : 3D view of: (a) the µbump. (b) the µbump and the signal strips (the ground is not shown). (c) Differential mode insertion loss and group delay of the µbump.

Figure G.c represents the differential mode insertion loss and group delay of the µbump. The group delay slightly varies at low frequencies and is almost flat at higher frequencies above 20 GHz. Therefore, the µbump will not lead to dispersion around the operating frequency of 60 GHz. The differential mode insertion loss of the µbump is 0.05 dB at 60 GHz, which is negligible compared to the overall link insertion loss, as will be shown later.

As for the 7-mm interposer channel, Edge-coupled differential microstrip lines were used in a six metal layers 130-nm BiCMOS technology.

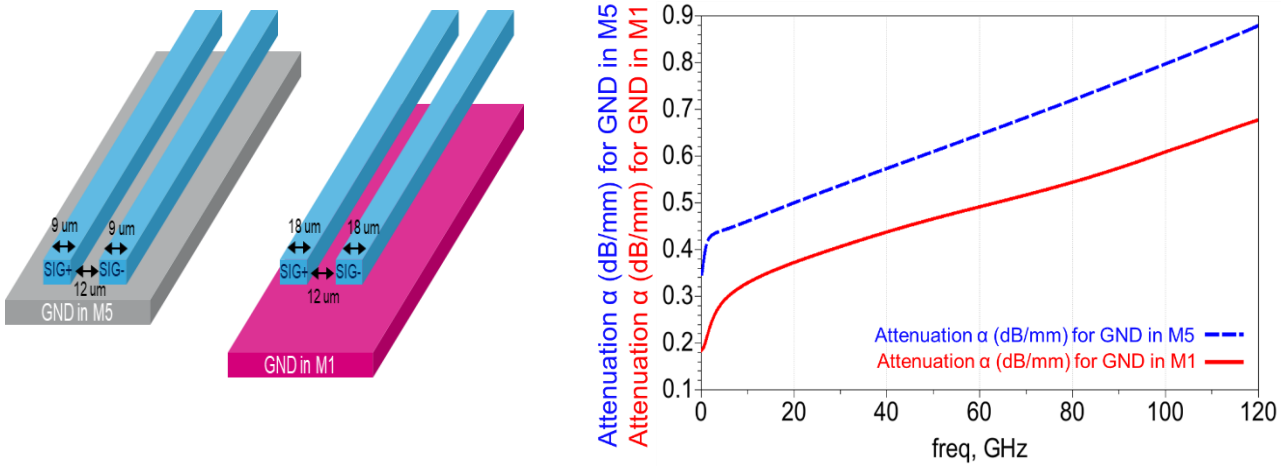


Figure H : Transmission lines configuration for ground in M1 or M5 and their attenuation constant.

We decided to put the ground in a higher metal instead of M1 bottom metal layer to diminish floorplan constraints. This choice has little effect on insertion loss due to the fact that differential mode is used. Figure H shows two edge-coupled microstrip lines configurations, where the ground is in M1 or in M5 metal layer, respectively. They both have a spacing of 12 um and a differential impedance $Z_{odd} = 45\ \Omega$.

We notice that, at 60 GHz, the attenuation constant is around 0.64 dB/mm when the ground is placed in M5, as compared to 0.49 dB/mm when the ground is in M1. This is simply due to wider strips when the ground is put in M1. Nevertheless, the difference of insertion loss is only 0.15 dB/mm, i.e., around only 1 dB for a 7-mm long interconnection, which corresponds to a negligible impact on the signal transmission but reduces the metal routing constraints significantly.

Figure I shows the differential mode insertion and return loss of the total 7-mm link from chiplet to chiplet (through the interposer). The total predicted insertion loss at 60 GHz is 4.7 dB.

We simulated the overall system, i.e., the previously transceiver (duobinary modulator, transmitter, and receiver) designed in chapter III, and the 3D EM simulated 7-mm interposer channel. The time-domain input and output for a 10Gbps signal are shown in Figure J.

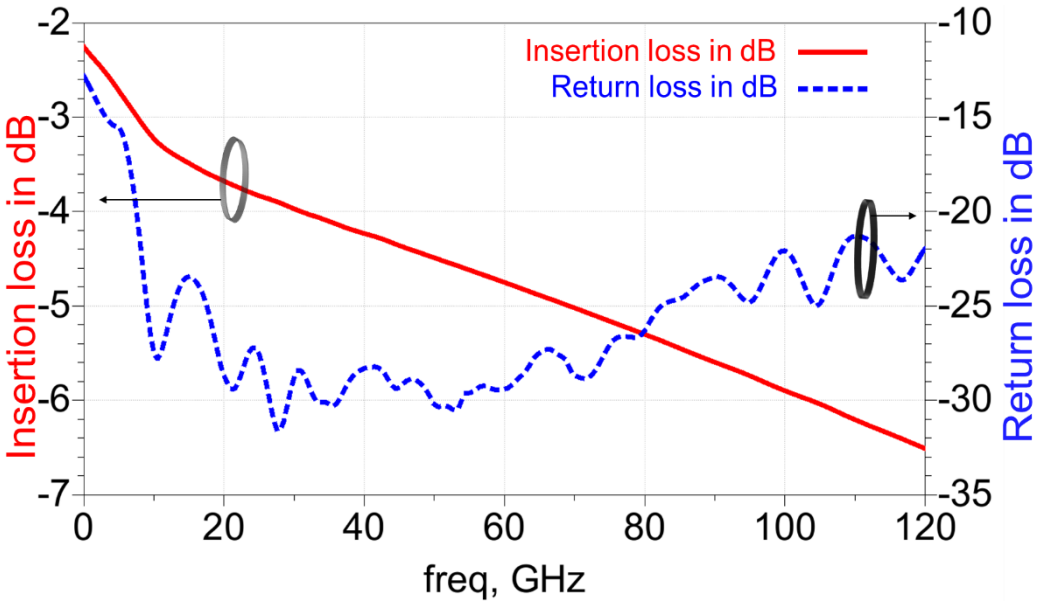


Figure I : Differential mode insertion and return loss of the 7-mm long channel.

The propagation delay follows the same reasoning explained in chapter III, where the latency decreases for higher data rates (the latency is $\propto 1/T_{data\,rate}$). We believe this solution offers a very low-latency (7-8ps/mm) solution compared to the baseband transmission that requires repeaters or complex equalization schemes. Furthermore, the duobinary modulation makes it more efficient, i.e., compressing the spectrum allows for more transmitted data. Different carriers may be used efficiently too. Moreover, this solution offers high signal integrity where a BER lower than 10^{-12} was achieved; this is mainly due to the very low distortion caused by channel, thus not requiring equalization or ECC.

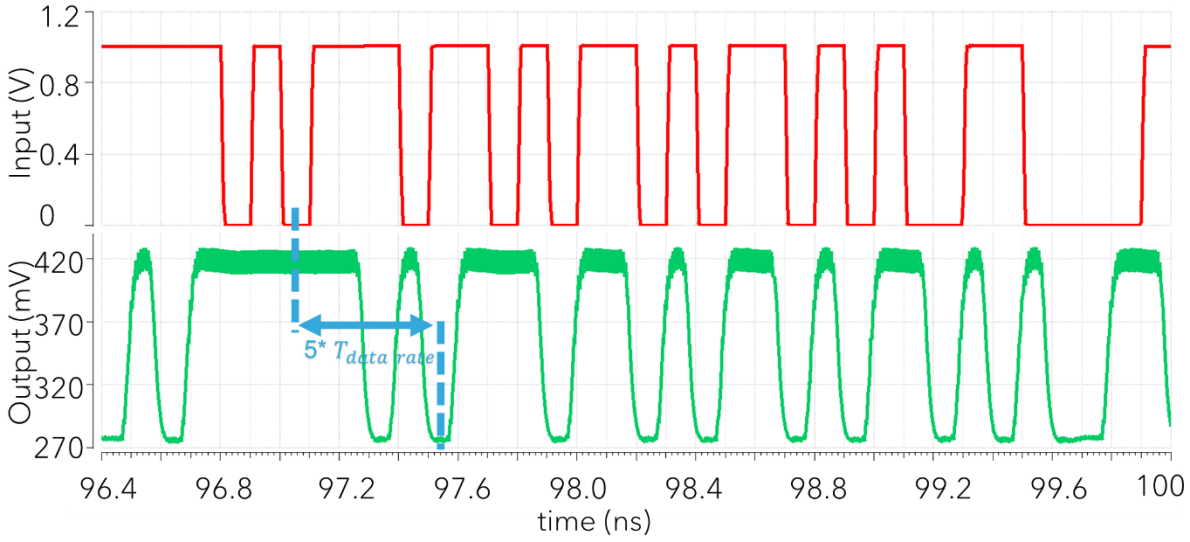


Figure J : Simulated system input and output waveforms.

We believe that this approach's main advantages, compared to the standard baseband transmission, are the lower latency and higher available bandwidth. For example, the authors in [8] achieved a 30Gbps transmission over 10mm silicon interposer differential lines with a BER lower than 10^{-13}. They used baseband

transmission, and we believe that the 10mm transmission lines were crossed in at least 650ps; this value is an estimate deduced from their time-domain figures at the input and output of the T-lines. While a transmission using the RF approach will result in a 70-80ps propagation latency for the same distance. To achieve the high data rate and signal integrity, they used a FFE equalizer, requiring significant effort for high data rates, as introduced in chapter I. While the RF approach relies on the transmission lines' physical behavior in higher frequencies to counter the distortion. This approach promises higher data rates if additional carriers are used or more complex modulations are used, for example, QAM modulation. Baseband interposer links are still an active research area; the same can be said about optical interposer links. Where the authors in [9] used optical waveguides to transmit a 12.5Gbps signal, they managed to integrate the required components (modulator and photodetector) on a single silicon substrate; Optical interconnects offer higher bandwidths than the other candidates; and latency comparable to the RF-approach. However, the processes required for silicon photonics are still complex and have a low yield; furthermore the surface used is relatively higher than an RF-interconnect. Each technique of the previous techniques has its advantages and limits; a fair comparison at these different maturity levels is impossible.

To conclude, on-chip (or interposer) interconnects are becoming a severe bottleneck for on-chip communications; In this work, we did our best to push the RF-interconnect limits in what we consider the right path by developing a simple way to double the data rate. We expect (and hope) that this work will be surpassed soon, and we wish good luck to the next innovators.

References

[1] M. Tmimi, S. D'Amico, J. Duchamp, P. Ferrari, and P. Galy, "7-mm low-latency inter-chiplet serial link with silicon interposer," in *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2020, pp. 683–686.

[2] B. Black, "Die stacking is happening," 2013.

[3] A. Usman *et al.*, "Interposer Technologies for High Performance Applications," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. PP, 2017, doi: 10.1109/TCPMT.2017.2674686.

[4] N. E. Jerger, A. Kannan, Z. Li, and G. H. Loh, "Noc architectures for silicon interposer systems: Why pay for more wires when you can get them (from your interposer) for free?," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*, 2014, pp. 458–470.

[5] M. A. Karim, P. D. Franzon, and A. Kumar, "Power comparison of 2D, 3D and 2.5D interconnect solutions and power optimization of interposer interconnects," p. 7.

[6] W.-S. Zhao, K. Fu, and G. Wang, "A Passive Equalizer Design for On-Interposer Differential Interconnects in 2.5 D/3D ICs," in *2019 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT)*, 2019, pp. 1–3.

[7] D. Stow, I. Akgun, and Y. Xie, "Investigation of Cost-Optimal Network-on-Chip for Passive and Active Interposer Systems," in *2019 ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, 2019, pp. 1–8.

[8] M. A. Karim and P. D. Franzon, "A 0.65 mW/Gbps 30 Gbps capacitive coupled 10 mm serial link in 2.5D silicon interposer," p. 4.

[9] Y. Urino *et al.*, "Demonstration of 12.5-Gbps optical interconnects integrated with lasers, optical splitters, optical modulators and photodetectors on a single silicon substrate," *Opt. Express*, vol. 20, no. 26, pp. B256–B263, 2012.

Publications

Tmimi, M., D'Amico, S., Duchamp, J.M., Ferrari, P. and Galy, P., 2019, June. 10Gbps Length adaptive on-chip RF serial link for Network on Chips and Multiprocessor chips applications. In *2019 International Conference on IC Design and Technology (ICICDT)* (pp. 1-4). IEEE.

M. Tmimi, S. D'Amico, J. Duchamp, P. Ferrari, and P. Galy, "7-mm low-latency inter-chiplet serial link with silicon interposer," in *2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2020.

Patent

M.Tmimi, P. Galy "Data transmimssion, In particular on a serial link having a great length." US patent filled in April 2020.