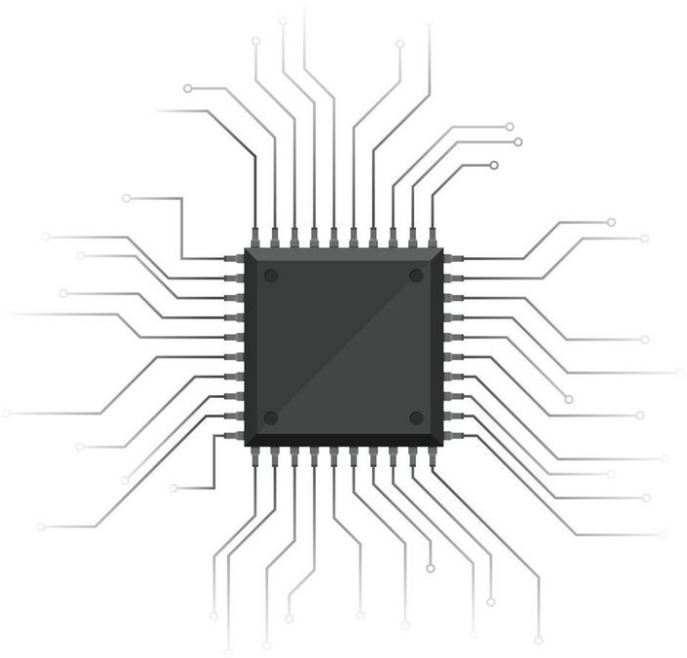
Achieving Optimal Power Performance with Advanced FinFET Technology for NAND Memory Applications up to 8Gb/s I/O Bandwidth

Master Thesis

by

Arvis Jomerts

St. nr. 4840917



University Supervisors:	Prof. Dr. ir. S. Hamdiuoi, Prof. Dr. ir. G. Gaydadjiev
Imec Supervisor:	Dr. A. Spessot
Imec Advisors:	Dr. I. G. Guerra, Dr. N. Pantano
Faculty:	Faculty of Electrical Engineering, Delft

Image:	vecteezy.com/Timplaru Emil
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

This page was intentionally left blank

Acknowledgements

After 9 months of sheer dedication to this project, I present this letter of appreciation with great pleasure.

First of all, I would like to express sincere gratitude to my Imec Project Manager Alessio Spessot for giving me a chance to obtain valuable experience in the industry environment, as well as continuous support and guidance throughout the project. I am extremely grateful to my daily Imec Supervisors Nicolas Pantano and Ivick Guerra Gomez for providing me with a continuous feedback loop, technical and practical advice, which helped me to elevate the quality of this thesis to a higher level. Their availability for Q&A sessions is highly appreciated, as without them, several doubts would be hard to straighten out.

I am also thankful to Delft University of Technology and its staff members for the quality of education they provide on a daily basis. Special thanks to Said Hamdioui and Georgi Gaydadjiev for accepting me under their supervision for this master thesis, and their inputs on the project during progress update meetings.

Lastly, I would like to mention Elma Lūcija Ulmane, who provided me with useful recommendations on thesis layout and figure visual representation. I would be remiss in not mentioning Kavitha Soundrapandiyan, with whom we had constructive discussion sessions on technical aspects of the project and on practical day-to-day matters.

Arvis Jomerts
Delft, October 2023

Abstract

3D NAND memory devices are intrinsically very cost sensitive, implying that their size, and hence logic area must be limited in order to acquire a chip which is able to conquer the competitive market price. Market forecasts of upcoming NAND products predict Input/Output (I/O) speed increase well beyond 2 Gb/s that is the current industry standard. I/O bandwidth strongly correlates with the device technology used for logic and the current state-of-the-art planar devices are predicted to reach their maximum capabilities in the near future. Use of FinFET transistor is expected to substantially enhance I/O area-performance of 3D NAND memory logic, alleviating area restriction severity and providing a foundation for future periphery generations to come.

In this thesis, a 3D NAND compatible I/O able to achieve 8 Gb/s throughput has been developed using simulation of thermally stable Imec in-house developed 14 nm FinFET technology equivalent. To validate throughput quality, industry defined eye diagram standards are used. To determine area savings provided by utilizing FinFET devices, FinFET active transmitter area is benchmarked against 45 nm planar device setup achieving the same 8 Gb/s data rate performance. To ensure an unbiased comparison, two signaling topologies are used - single ended signaling (SES) and differential signaling (DS). To extend analysis, sensitivity of the design against various parameters such as data rate, voltage and temperature is explored.

It is concluded, that active area of FinFET driver is several times lower than that of similar planar transmitter (same power and throughput) for both SES and DS. Additionally, suitable use cases of DS and SES have been evaluated depending on environmental conditions investigated during sensitivity analysis. All in all, this research provides a baseline for planar-to-FinFET scaling in I/O system and guidelines in choosing signaling topology appropriately, depending on system constraints.

Contents

List of Figures

Nomenclature

Abbreviations

Abbreviation	Definition
AC	Alternating Current
BER	Bit Error Rate
BGA	Ball Grid Array
BJT	Bipolar Junction Transistor
BW	Bandwidth
CD	Common Drain
CEM	Compatibility Eye Mask
CG	Common Gate
CM	Current Mode
CML	Current Mode Logic
CMs	Current Mirrors
CMRR	Common Mode Rejection Ratio
CMOS	Complementary Metal-oxide Semiconductor
CoB	Chip-on-board
CS	Common Source
CTT	Centre Tapped Termination
CuA	CMOS under Array
DC	Direct Current
DCS	Differential Current Steering
DDR	Double Data Rate
DP	Differential Pair
DR	Data Rate
DS	Differential Signaling
ECL	Emitter-coupled Logic
EM	Elect-magnetic
EMI	Electro-magnetic interference
EMS	Embedded Microstrip
EOT	Equivalent Oxide Thickness
ESD	Electrostatic Discharge
FC	Flip Chip
FEM	Finite Element Method
FET	Field-effect Transistor
FF	Fast-fast
GBWP	Gain Bandwidth Product
GF	Gate First
GND	Ground
HDD	Hard Disk Drive
HK/MG	High κ Metal Gate
HSTL	High Speed Transceiver Logic
IC	Integrated Circuit
ISI	Inter-symbol Interference
I/O	Input/Output
LTT	Lower Tapped Termination
LVDS	Low-voltage Differential Signaling
MS	Microstrip
NGS	Near Ground Signaling
ODT	On-die Termination
ONFI	Open NAND Flash Interface
OTA	Operational Transconductance Amplifier
PACK	Package
PCB	Printed Circuit Board

Abbreviation	Definition
PDS	Pseudo-differential Signaling
PDN	Pull-down Network
PODL	Pseudo-open Drain Logic
PSRR	Power Supply Rejection Ratio
PU	Processing Unit
PUN	Pull-up Network
PVT	Process, Voltage, Temperature
QFN	Quad Flat No-lead
QFP	Quad Flat Package
RC	Resistance-capacitance
RMG	Replacement Metal Gate
RX	Receiver
SA	Sensitivity Analysis
SFD	Saturated FET Driver
SES	Single Ended Signaling
SL	Stripline
SLVS	Scalable low-voltage Signaling
SNR	Signal-to-noise Ratio
SOI	Silicon on Insulator
SS	Slow-slow
SSD	Solid State Drive
TERM	Termination
TL	Transmission Line
TSV	Through-silicon Via
TT	Typical-typical
TX	Transmitter
UI	Unit Interval
VM	Voltage Mode
ZTC	Zero Temperature Coefficient

Symbols

Symbol	Definition	Unit
A	Area	[m^2]
C	Capacitance	[F]
C_{ds}	Drain-to-source Capacitance of Transistor	[F]
C_{gd}	Gate-to-drain Capacitance of Transistor	[F]
C_{gs}	Gate-to-source Capacitance of Transistor	[F]
C_{in}	Input Capacitance	[F]
C_L	Load Capacitance	[F]
C_{mut}	Mutual Capacitance	[F]
C_{out}	Output Capacitance	[F]
C_{ox}	Transistor Gate Oxide Capacitance	[F m^{-1}]
C_{vb}	Via Body Capacitance	[F]
C_{vp}	Via Pad Capacitance	[F]
c	Speed of Light in Vacuum	[ms^{-1}]
d	Separation Between Two Adjacent Wires/Lines	[m]
f	Signal Frequency	[Hz]
G	Conductance	[S]
g_m	Transistor Transconductance	[S]
h	Dielectric Height of PCB	[m]
h_{via}	Via Height	[m]
I	Current	[A]

Symbol	Definition	Unit
I_D	Drain Current	[A]
I_{OL}	Current Flow Through Pull-down Transistor	[V]
I_{sink}	Current Sunk by SFD	[A]
J_0	Bessel's Function of First Kind of Order Zero	[-]
k_a	Correction Factor for Return Path Resistance	[-]
k_p	Correction Factor for Proximity Effect	[-]
k_r	Correction Factor for Copper Roughness	[-]
L	Transistor Length	[m]
L	Inductance	[H]
L_{mut}	Mutual Inductance	[H]
L_{slef}	Self Inductance of a Component	[H]
l_bw	Bondwire Length	[m]
l_{prop}	Distance Signal has Propagated during Rise/Fall Time	[s]
l_{TL}	Transmission Line Distributed Instance Length	[m]
m	Number of Capacitive Increments	[-]
m	Number of Gaps Between Signal and Guard Lines	[-]
N	Number of Inverter Chain Stages	[-]
N_0	Neumann's Function of Order Zero or Bessel's Function of Second Kind of First Order.	
n	Number of Signal and Guard Lines	[-]
$P_{dynamic}$	Dynamic Power Consumption	[W]
P_{static}	Static Power Dissipation	[W]
P_{tot}	Total Power Consumption	[W]
p	Perimeter of Cross-section	[m]
Q	Charge	[C]
R	Resistance	[Ω]
R_{AC}	Alternating Current Resistance	[Ω]
R_{avg}	Average Surface Roughness	[μm]
R_{bot}	Bottom Termination Resistance	[Ω]
R_d	Drain Termination Resistance	[Ω]
R_{DC}	Direct Current Resistance	[Ω]
R_{PD}	Pull-down Resistance	[Ω]
R_{top}	Top Termination Resistance	[Ω]
r_{ap}	Via Anti-pad Radius	[m]
R_p	Via Pad Radius	[m]
r_{bw}	Bondwire Radius	[m]
r_o	Transistor Output Resistance	[Ω]
r_{via}	Via Radius	[m]
T_{bit}	Bit Width	[s]
t	Time	[s]
t	Trace/Copper Layer Thickness	[m]
t_r	Rise Time	[s]
V_{cm}	Common Mode Voltage	[V]
V_{dd}	Power Supply Voltage	[V]
V_{ds}	Drain-to-source Voltage of Transistor	[V]
V_g	Transistor's Gate Voltage	[V]
V_{gs}	Gate-to-source Voltage of Transistor	[V]
V_{in}	Input Voltage	[V]
V_{OH}	Threshold Voltage of High State	[V]
V_{OL}	Threshold Voltage of Low State	[V]
V_{out}	Output Voltage	[V]
V_s	Source Voltage	[V]
V_{th}	Transistor's Threshold Voltage	[V]

Symbol	Definition	Unit
V_{Z_0}	Voltage drop across termination resistance	[V]
V_+	Voltage of RX Positive Terminal	[V]
v	Propagation Velocity of Signal	[ms^{-1}]
W	Transistor Width	[m]
W	Trace Width	[m]
Z_0	TL Characteristic Impedance	[Ω]
α	Scaling Factor Between Two Consecutive Inverter Chain Stages	[-]
Δ	Variation/Deviation	[-]
δ	Skin Depth	[m]
tan δ	Loss Tangent of Dielectric Material	[-]
ϵ_{eff}	Effective Dielectric Constant	[-]
ϵ_r	Relative Dielectric Constant	[-]
η	Input-to-Output Capacitance ratio	[-]
κ	dielectric constant	[-]
λ	Channel Length Modulation Parameter	[V^{-1}]
μ	Mobility of Majority Charge Carriers	[m^{2}V^{-1}s$^-1$]
μ	Material Permeability	[Hm^{-1}]
μ_0	Permeability of Free Space	[Hm^{-1}]
μ_r	Relative Permeability	[Hm^{-1}]
ρ	Resistivity	[Ωm]
τ	RC Constant	[s]
χ_{01}	First Root of Via Pad Determination Chain	[-]
ω	Angular Frequency	[rad s^{-1}]

1. Introduction

This chapter introduces the subject covered in the thesis. The motivation to investigate the topic is discussed first and can be found in Section 1.1. Brief overview of the state-of-the-art NAND I/O performance can be found at the end of Section 1.1. The main thesis objective together with contributions can be found in Section 1.2.

1.1. Motivation

Unrelenting thirst for higher capacity data storage with faster transfer capabilities has been ceaselessly driven by continuous generation of data in large quantities (Big Data [1]) that has to be processed and retained in the system [2]. With processing speeds developing rapidly and memory bandwidth (BW) lagging behind, data transfer bottleneck has grown immensely in conventional processor-centric architectures, causing designers to use multi-level cache memory to hide transmission latency [3]. Several solutions exist which can resolve the lack of data rate (DR) compatibility between memory and processor - one of the most straightforward fixes is increasing input/output (I/O) speed performance [4].

A promising candidate for Big Data applications is solid state drive (SSD) based on 3D NAND flash memory, as it provides high memory density ought to further increase [5] and relatively high I/O data rates in comparison to hard disk drives (HDD) [6]. With SSD niche being portable devices as smartphones and laptops, further improvement in NAND I/O speed opens possibilities for high performance, light-weight and small-size device manufacturing [7].

Conventional 3D NAND I/O logic is commonly located under memory die stacks (defined as complementary metal-oxide semiconductor (CMOS) under array (CuA)) making periphery area strongly size restricted [8]. Thereafter, using a lower technology node able to provide higher drive strength at iso-area conditions and meeting thermal stability requirement imposed by 3D NAND manufacturing technique, is a prominent solution to I/O BW limitation [4]. Area-demanding transistor devices have already been shown to be the culprit in I/O speed limitation [9].

Current state-of-the-art 3D NAND I/O achieves DR as high as 2.4 Gb/s for CuA configuration [10], while recently, staggering 3.2 Gb/s I/O transfer has been presented for memory-logic wafer-to-wafer bonded structure [1] [2]. However, as I/O speed is expected to roughly double every 3 consequent years [11], investigation of next generation I/O for 3D NAND is crucial. With current trends leaning towards CuA system integration [12], one can guess that closest descendants of latest commercial products will employ alike topology. Thereby, it is particularly interesting to investigate CuA compatible high-performance, thermally stable devices with mature manufacturing flow.

1.2. Contributions

With all the aforementioned in mind, this thesis aims to develop and analyse 8 Gb/s 3D NAND compatible I/O employing thermally stable Imec in-house developed Fin field-effect transistor (FinFET) 14 nm technology equivalent. Such configuration achieves $\approx$3 times higher BW, implying that it corresponds to 2nd generation of current I/O standards. The main contributions of this thesis are:

- **A journal submission on optimal performance NAND I/O platform for future generations:** This thesis provides a design methodology and analysis for FinFET based NAND I/O able to reach 8 Gb/s. Industry defined signal quality requirements [13] were used to validate system throughput, ensuring that common signalling protocols are not violated. With this, a systematic approach to develop 8 Gb/s interconnect is developed, which can be repeated for any FET technology. The paper has been submitted to IEEE Journal of Solid-State Circuits (JSSC), see Appendix A.
- **Benchmarking of FinFET based I/O against planar alternatives:** In this thesis it was concluded that transmitter active area footprint can be reduced multiple times if switch from planar 45 nm tech. to 14 nm FinFET node is performed. Iso-performance conditions and shared power supply voltage were applied. Determination of exact area savings attained allows to evaluate whether shift towards using lower, more costly, technology pays-off.
- **Exploration of high-speed I/O system sensitivities against environmental conditions and system parameters:** In this thesis I/O analysis is broadened by exploring I/O design sensitivity to

[1] URL `https://www.anandtech.com/show/18799/kioxia-and-western-digital-debut-218layer-3d-nand-1tb-tlc-with-32-gts-io-speed` [cited on 5th of May 2023]

[2] URL `https://www.tomshardware.com/news/kioxia-and-western-digital-unveil-worlds-fastest-3d-nand?utm_medium=social&utm_source=twitter.com&utm_campaign=socialflow` [cited on 5th of May 2023]

various parameters, such as DR, voltage, temperature and so forth. Sensitivity analysis provided indication of which system parameters require higher amounts of focus, to ensure proper operations in case environmental condition or system parameter variations are expected. Sensitivity is evaluated in terms of how much area has to be increased/relaxed to compensate performance degradation/enhancement to keep iso-performance conditions.

- **Use of advanced FinFET technology in strictly area limited and temperature sensitive system:** In this thesis CuA structured 3D NAND I/O was explored, inferring a low periphery size and thermal stability requirements. It is worht noting that equivalent technology was used. Equivalent technology implies that characteristics of NAND thermal annealing compatible logic devices would match those of current 14 nm FinFET Dynamic random access memory (DRAM) annealing process tolerant transistors, when manufacturing flow for required NAND devices would be established. At the moment, development of NAND thermally stable logic devices is under way, thereafter, best provided alternative is used in this project assuming that almost exactly the same characteristics will be obtained for the product yet to come.

1.3. Thesis Outline

This thesis is structured in the following manner. Chapter 2 discusses state-of-the-art I/O topologies, covers sensitivity analysis scope, and defines theory related to transmission line, termination and packaging. In Chapter 3, design limitations, constraints and requirements are defined, choice of design topology made and target case simulations provided and discussed. Sensitivity analysis of various parameters is performed in Chapter 4, which is accompanied with discussion on observed outcomes. The thesis is concluded in Chapter 5, which is supplemented with recommendations for future research.

2. 3D NAND I/O Interface and Topology Overview

This chapter discusses theory related with I/O system and provides both qualitative and quantitative analysis of I/O underlying components. In Section 2.1 various 3D NAND system specifics are covered. Section 2.2 provides an overview of commonly used I/O topologies, including both benefits and drawbacks. Options for receiver amplifier are given in Section 2.3. Theoretical analysis of the transmission line can be found in Section 2.4, while high-speed applicable packaging characteristics are given in Section 2.5. Further, discussion on termination is provided in Section 2.6. Lastly, the setup of sensitivity analysis is provided in Section 2.7.

2.1. 3D NAND I/O General Considerations

Conventional memory to processing unit (PU) interconnect architecture consists of three main components - memory, PU (including memory controller) and external link in combination with its supporting data transfer enabling logic. The key elements of the logic are the driver and receiver located at I/O pins, which characterise *Data* signal strength and quality upon transmission and reception. Other supporting signals as *Data Strobe*, *Address Line Enable*, *Chip Enable*, etc. are required to ensure that data is properly sampled, written/read in/from the right place in memory, and all of that is done at the right time [14]. Even though all the aforementioned signals have different electrical requirements [13], the same interconnection topology can be used for their transmission from one integrated circuit (IC) to the other.

Note, NAND I/O topology has a generic design, meaning that the design could be shared with other purpose I/O interconnects. Nevertheless, the key difference for state-of-the art NAND devices is the compatibility with technology node used. Commonly 3D NAND memory periphery (e.g. I/O logic) is produced on the same wafer as memory array, with the memory being located on top - the structure is more frequently referred to as CuA [8]. The former imposes several limitations on the logic circuitry - first, area available for periphery is restricted by chip size and required memory density [15]. Thereafter, further scaling of NAND chips to smaller sizes to satisfy industry's growing demand for miniaturization requires rapid scaling of periphery logic.

Second, CuA requires manufacturing of logic circuitry first - it has to withstand not only temperature annealing of itself, but also the memory. The former implies that the logic devices have to be thermally stable, since memory and logic subjection to temperature is different, with memory temperature being higher. Thereafter any conventional device technology cannot be directly transferred to stacked system applications. [16]

Temperature stability also allows to fix system BW and settling response to a constant value more accurately, as resistance-capacitance (RC) contributions of on-chip components are more process, voltage, temperature (PVT) resilient [15]. The bandwidth is further enhanced by stacked memory-logic placement, as trace length reduction is enabled relatively to 2D NAND memory, lowering overall interconnect trace resistivity [8].

Current CuA state-of-the-art NAND I/O employs planar technology, reaching 2 Gb/s I/O speed [17], however, as transmission speed is expected to follow Moore's law, it is inevitable that other more performing solutions will be required. It is crucial to investigate such means timely to determine how much device performance is going to be increased when advancing to a smaller technology node. Thereafter, the novelty of this thesis lies in investigating CuA NAND I/O attained area when FinFET devices are used and benchmarking this result with planar device size for iso-performance conditions. Two wafer bonded chips using through-silicon vias (TSV) is left as future research as it provides relaxed temperature and area requirements.

When considering CuA with regards to FinFET devices, thermally stable, and as mature as possible, technology manufacturing flow has to be selected. Bear in mind, FinFET devices have switched from using SiO_2 as gate insulating material and doped poly-silicon gate conductor to thin high dielectric (κ) metal gate structure (HK/MG), which tends to be more susceptible to degradation upon thermal anneal [18]. However, without HK/MG, further down-scaling of transistors and their oxide thickness would be impossible due to immense SiO_2 device leakage current [19].

There are two HK/MG compatible manufacturing techniques: gate-first (GF) and gate last or replacement metal gate (RMG). As the name implies, in GF the HK/MG is implemented before full device

is formed, while in RMG first a dummy gate is made, which is removed after thermal annealing and later replaced with actual HK/MG. From the above, one can derive that GF structure is significantly more temperature resistant, as it has to survive its own annealing while RMG bypasses temperature processing associated challenges almost entirely. The former makes GF a more compatible technology with CuA type NAND memory, however, GF gate dielectric material choice is severely limited as device has to tolerate high temperature exposure. [20]

With all the above in mind, both RMG and GF can be viable options for 3D NAND I/O if manufacturing and device tolerance modifications are achieved. For the given moment, as technology development cannot be precisely predicted, design investigation is performed using current thermally stable Imec 14 nm FinFET technology equivalent with both RMG and GF stack-ups. In this chapter various state-of-the-art I/O interconnects suitable for 3D NAND memory I/O interface are covered, while discussion on devices is continued in Section 3.6.

2.2. I/O Interconnect Topologies

Signal propagation in the system can happen over two different configurations - point-to-point or multi-point interface. In point-to-point interconnect, the signal is sent from one IC directly to the other, while in multipoint configuration, signal created by one IC can be simultaneously used by a multitude of other components. To simplify the analysis, only point-to-point interface is analysed further in this chapter, as multipoint topology can be generated using point-to-point elements as the most basic building blocks. Performance degradation due to reflections from idle component interface would be observed in multi-point configuration - investigation of this effect is left as recommendation for future research. Moreover, only a unidirectional interface is considered, assuming that bidirectional signalling mode would not cause significant system performance loss with well designed clocking network [21].

Conventional point-to-point signalling interface is composed of 4 main components: transmitter (TX), receiver (RX), termination (TERM) and transmission line (TL). Nevertheless, to make a clear distinction between TX and RX on-die and external components, packaging is introduced as an additional element. Note, the packaging of RX is assumed to be a mirrored version of TX package as now the signal is coming in the IC rather than going out of it. With this, an example of the overall point-to-point system for differential and single ended signalling look as depicted in Figure 2.1 (TERM is not depicted as it differs per choice of signalling topology, e.g series/parallel)

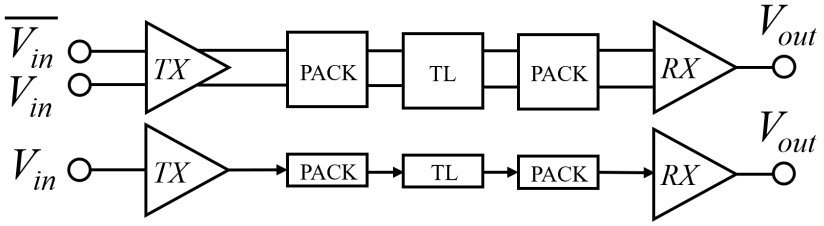


Figure 2.1: An example of an I/O interconnect for differential (top) and single ended (bottom) signalling

Electrical data transfer between two I/O pins can be split into two main classes - voltage mode (VM) and current mode (CM) signalling. The former scheme propagates information in the form of nodal voltages, while the latter performs the information processing via branch currents. The system's termination serves as the main indication of the class - VM signalling uses series or no termination, with signal propagation relying only on the charging and discharging of systems intrinsic capacitances. CM signalling on the other hand mostly uses parallel termination, with a possibility of combined series-parallel termination, thus, directly utilizing the voltage rails to source/sink current. [22]

The structural differences of the signalling classes lead to a variation in circuit's attainable speed performance, with the current driven circuits being superior. The CM signalling outperforms voltage mode with its increased noise and power supply variation rejection, higher data transfer bandwidth and improved impedance matching. Nevertheless, the drawbacks of using a CM signalling is relatively high static power dissipation and increased area, if weak saturation current sources are used. [14]

Note, exploration of VM circuits is omitted in this thesis as the required data rate of 8 Gb/s for a single I/O pin renders VM signalling practically infeasible [22]. Thereafter, for complete theoretical coverage

only a mention of basic voltage mode circuit operation mechanism is deemed sufficient.

CM circuits can be further divided in single ended (SES) and differential signalling (DS) modes. Each of the data transfer modes can be realized with a variety of topologies providing different advantages. First, the SES topologies are discussed in Subsection 2.2.1, which is followed by Subsection 2.2.2 where DS topologies are covered. Observe, this chapter discusses sizing of the devices only qualitatively. Also, as the objective of the thesis stands as designing an I/O able to achieve continuous 8 Gb/s transmission speed, only the driver of the TX is going to be analysed. It is assumed that prior circuitry performs all the required actions to ready the signal for transmission. Thus, potential pre-circuitry components are neglected in this thesis, implying that full system integration is required to verify the findings obtained in this work. Note, only input drive strength limiting elements are included to ensure that TX is not under-designed.

2.2.1. Single Ended Signalling Topologies

Single ended signalling is the current standard for communication between NAND memory and a PU due to its compact size and integration simplicity [13] [10]. Nonetheless, it has to be kept in mind, that state-of-the-art bonded NAND I/O interface can achieve up to 3.2 Gb/s transfer per line [1] [2] utilizing planar technology, while CuA reaches only 2 Gb/s margin [17]. Hence design choices have to be critically reevaluated when investigating DR of up to 8 Gb/s due to more than doubling/quadrupling in frequency.

For instance, the crosstalk and reflection caused signal degradation becomes more severe with increase in signalling speed [23] and intrinsic device gain and thus also stage gain reduces with higher switching frequency. Reduced gain in combination with lowered power supply voltage of smaller device technology node (14 nm FinFET vs. Planar 45 nm) leads to more susceptibility to both amplitude and time noise, causing higher likelihood of signal errors to arise [24]. Nevertheless, it is assumed that SES topologies can reach the target speed, therefore they are considered as potential candidates for the 8 Gb/s interconnect.

Center-Tapped Termination

Center-tapped termination (CTT) is one of the most common transmission standards used in small chip applications, where area is crucial and can come with the sacrifice of power dissipation [22]. A schematic of CTT interconnect can be seen in Figure 2.2. The TX is composed of a push-pull circuitry, the TERM is in Thevenin configuration (parallel to both rails) and the RX is differential-to-single ended with the negative terminal being applied to the reference voltage. Since RX can be common for various topologies, it is not going to be covered in the individual subsections - a discussion on several receiver options can be found in Section 2.3. Note, the value X at negative RX terminal is the ratio between reference voltage (here, the common mode voltage (V_{cm})) and power supply voltage (V_{dd}).

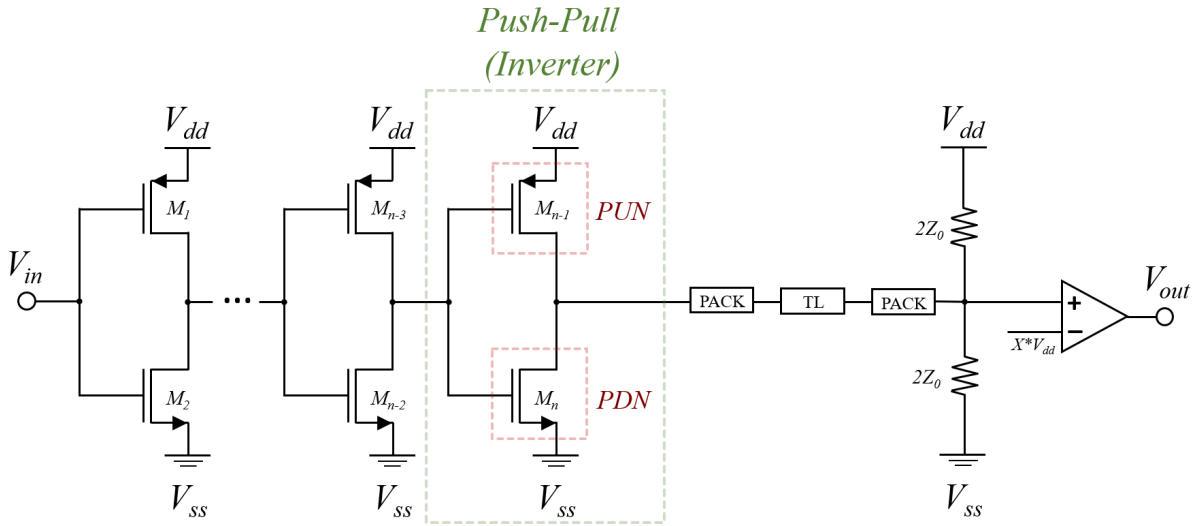


Figure 2.2: Schematic of CTT topology

Commonly push-pull circuitry consists of a pull-up network (PUN) which is responsible for sourcing current from the power supply and pull-down network (PDN) which sinks the current into the ground [3]. For CTT, PUN is composed only by a single p-FET (PFET) and PDN by a single n-FET (NFET). Such configuration allows to interchangeably charge or discharge next stage depending on the input received, making the circuit fully transparent.

The push-pull circuitry of the driver can be designed in various ways - both using complementary and only single type (NFET or PFET) devices. The use of single type devices corresponds to high speed transceiver logic (HSTL) interface which is covered later in this section. Using a chain of inverters (complementary push-pull) as driver ensures that a full rail-to-rail swing is utilized, which allows to provide maximum drive strength to the output, meaning that large loads can be driven. Nevertheless, complementary circuitry has several drawbacks especially if planar technology is used. For instance, mobility differences in PFET and NFET transistors causes an approximate effective current ratio of 2 between PUN and PDN devices. Thereafter, either the speed and thus one of the signal transitions is compromised or PFET has to be made double the size of NFET to compensate for lack of mobility [24]. Luckily, strength deviation is practically absent in FinFET devices, hence most of PFET and NFET device characteristics are indistinguishable [25]. However, small deviations in effective current and output resistance accumulate in large area designs, causing the necessity to slightly differentiate sizing of PUN and PDN.

The most characteristic property of CTT is the symmetric termination, which sets the V_{cm} on the line to be half the termination voltage. Nevertheless, Thevenin configuration generates a direct current path between the power rails, consequently causing static power dissipation at all times. By increasing TERM resistance value the power dissipation can be reduced, however, it comes at a cost of increased signal reflections due to poor matching and charging delay at RX input due higher RC constant.

A particularly interesting characteristic caused by TERM upon signal switching from high-to-low or vice versa is non-constant slew rate on a single transition edge. This phenomenon is depicted in Figure 2.3. The split happens at half supply voltage due to termination first helping to bring the voltage up/down to the common mode level and later fighting the signal, trying to bring it back down/up.

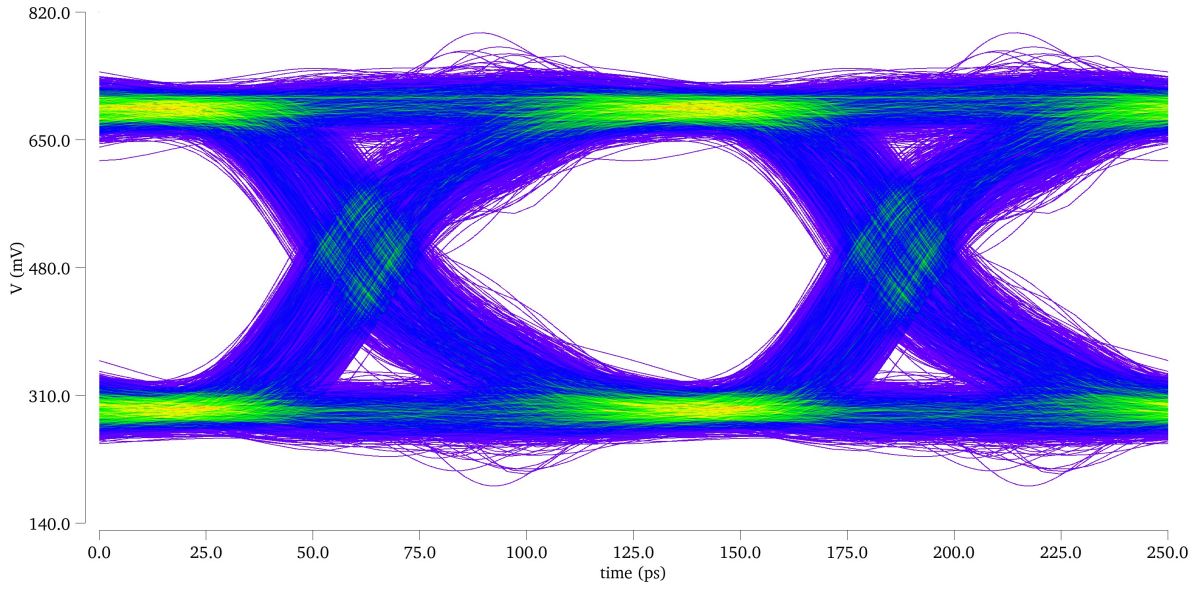


Figure 2.3: Slew rate non-uniformity at RX input voltage in single edge transition of CTT topology with following simulation conditions: Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

TERM can also be used for shifting V_{cm}. As TERM resistors act as a simple potential divider, using different resistor values for top and bottom leads to a change in center voltage. The shift might be required to compensate for common mode skew due to variation in PUN vs. PDN strength. It is

[3] URL `https://www.analog.com/en/design-center/glossary/push-pull.html` [cited on 10th of June 2023]

important to set V_{cm} to half power supply voltage in order to obtain symmetric signal properties for '0' and '1', especially if other skew correction methods are absent.

Alternatively, differences in PFET and NFET strength can be counteracted by using asymmetric inverters, however, it is far from the optimal strategy as the whole chain of inverters would have to be ideally skewed. The latter is infeasible for large, multi-stage designs since PVT influence on the system would demolish the meta-stability achieved in ideal conditions. Upon PVT variations, fluctuation of V_{cm} would be amplified with each subsequent stage and further enhanced by attenuation and reflections on the TL.

With all the CTT components defined, power consumption has to be determined. Observe, CTT power can be estimated in terms of TERM and voltage supply voltage value. The total power consists of two components: static and dynamic power. Static power is consumed continuously, when power supply has been activated, even if no input data is provided. In case no input is applied, the static power dissipation can be estimated by Equation 2.1, where V_{dd} represents the power supply voltage value. Otherwise, the computations are slightly more tedious and are shown after determination of dynamic power.

$$P_{static} = \frac{V_{dd}^2}{R_{top} + R_{bot}} \tag{2.1}$$

The dynamic power can be estimated to the first degree with taking certain assumptions. First, it is assumed that only the last stage of the inverter chain has a significant contribution to power dissipation. The assumption holds true as long as TL and PACK intrinsic capacitances are significantly larger than gate capacitance of last inverter. In such a case the current to charge the external circuitry has to be multiple times higher than self-loading current. Second, the output resistances of NFET and PFET transistors are assumed to be equal, implying that input voltage (V_{in}) equal to '0' or '1' provides equivalent loading case. The only difference is whether current gets sourced or sunk. Lastly, the transistor output resistances are assumed to be invariant, which can only be the case if ideal, zero rise time signals are present.

Ideally the dynamic power is negligible, which is the case when output resistance of active stage is infinite. This partially holds for low speed designs where TX area is small and transition of the signals happens rarely. However, as higher data rate requires more current to be provided by the transmitter to charge the load in shorter time period, the inverter area has to be increased. As a result, the internal resistance of the transistors is reduced. Taking a time instance where only NFET transistor is operational, the equivalent circuit model in terms of resistances can be seen in Figure 2.4. It can be noticed, that the NFET output resistance is effectively in parallel to pull-down termination resistance, thus reducing the overall rail-to-rail resistance. For calculations in this thesis, dynamic power is assumed to be of comparable magnitude to static power and thus has to be accounted.

With the former assumptions in place, the total power dissipation of CTT can be approximately determined as shown in Equation 2.2, where r_o corresponds to the output resistance of NFET. The dynamic power is approximated as $I_{OL}^2 r_o$, where I_{OL} represents the current value through the PDN transistor. Bear in mind, the power in the equation is given for case of turned-on NFET and turned-off PFET. Due to the aforementioned assumptions power in reverse operating case would be exactly the same, thus opposite scenario computations are omitted here.

$$P_{tot} = \frac{V_{dd}^2}{R_{top} + \left(\frac{1}{R_{bot}} + \frac{1}{r_o}\right)^{-1}} \tag{2.2}$$

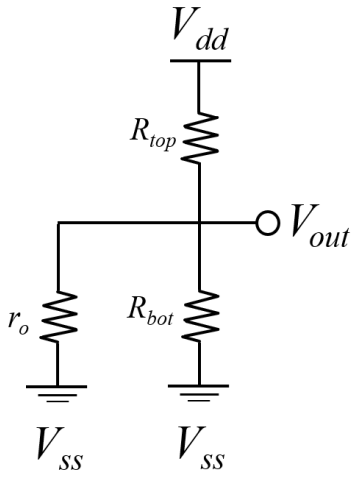


Figure 2.4: Equivalent CTT circuit model in terms of resistances, when only NFET is operational

To determine the individual components (static and dynamic) slightly more sophisticated computations have to be performed. For simplicity, only the dynamic power is determined as shown in Equation 2.3, from which static power can be expressed as $P_{static} = P_{tot} - P_{dynamic}$. V_{OL} is defined as threshold voltage of system's low state.

$$P_{dynamic} = r_o \left(\frac{V_{dd}}{R_{top} + \left(\frac{1}{R_{bot}} + \frac{1}{r_o}\right)^{-1}} \cdot \frac{\left(\frac{1}{R_{bot}} + \frac{1}{r_o}\right)^{-1}}{r_o} \right)^2 = \frac{V_{OL}^2}{r_o} \tag{2.3}$$

The inverter has a fully transparent signal propagation characteristic for an ideal input signal - either of driver's transistor is going to be 'on' if there is a V_{in} applied. Thereby, the total power given in Equation 2.2 represents the average power in time. In case of realistic signals, both the transistors could be operational simultaneously during V_{in} transition leading to higher dynamic power dissipation due to alternate (w.r.t TERM) current path to ground. Dynamic energy wise, each of the transistors would constitute for only a half of total dynamic energy as their duty cycles (on time compared to period) are 'on' average 50% long.

Saturated FET Driver

Saturated FET driver (SFD) is one of the earliest and simplest transmitter designs used in I/O systems [21] [26]. An example schematic of SFD topology can be seen in Figure 2.5. The simplest form of SFD includes only a single transistor operating in saturation region and one parallelly terminated load resistor. Saturation conditions for an NFET correspond to $V_{ds} > V_{gs} - V_{th}$, where V_{ds} is drain-to-source voltage, V_{th} is threshold voltage and V_{gs} is gate-to-source voltage [27]. To improve the drive strength and signal quality on the line another parallelly terminated resistance can be added to the source side. However, adding source side termination would come at the cost of nearly doubling the dynamic power consumption due to both terminations seen to be effectively in parallel - halving overall resistance.

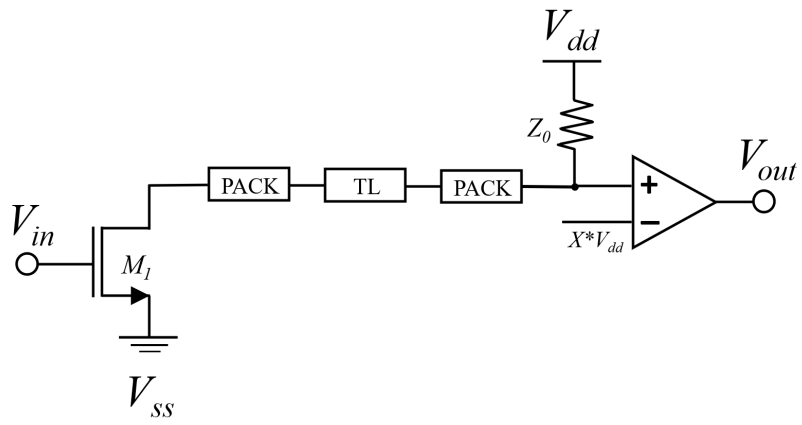


Figure 2.5: Schematic of Saturated FET Driver

The transmitter operates by being only partially transparent - the inverse of the signal is transmitted only when input is high, else, the line is pre-charged to termination voltage. During the evaluation phase, the NFET has to sink the current on the line to obtain the desired voltage drop across the termination resistor (see Figure 2.6). However, as termination is continuously going to resist the voltage drop, it can be concluded that the falling edge of signal seen at the input of RX is going to be significantly less steep than the rising edge. The lagging discharge could potentially cause bit errors, if long streams of 1 bits are transferred as interconnect's intrinsic capacitors would become fully charged and thus tought ot discharge. The errors can be prevented by increasing transistor's area, and as a consequence, enhancing the current sinking abilities and reducing TX output resistance.

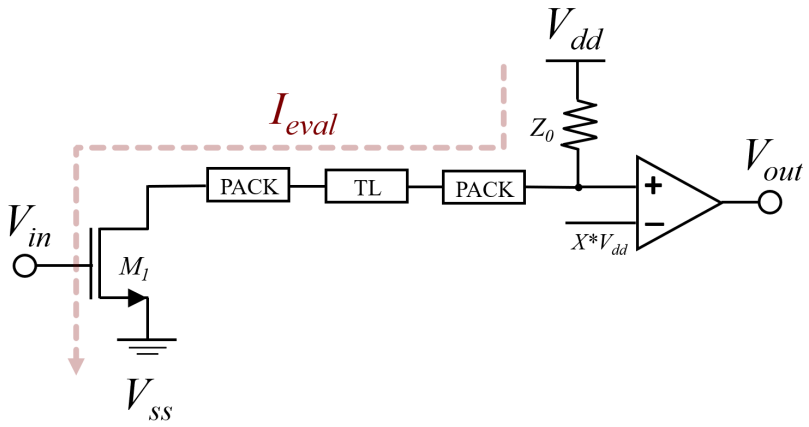


Figure 2.6: Schematic of Saturated FET Driver operation principle

Moreover, a single transistor configuration can lead to performance degradation of up to two times due to PVT variation. To stabilize the current sinking, a set of parallelly connected digitally controllable saturated current mirrors (CMs) can be added on the source side of the driver as shown in Figure 2.7. By switching various current mirrors 'on' and 'off' it is possible to set the current to a desired value - a reference has to be compared to the current in the circuit. Nevertheless, the extra circuitry leads to an immense increase in area as both finite state machine (FSM) for controls and idle CMs are required to ensure that correct current is provided. The PVT offset can also be partially rectified by having significantly larger output swing by directly increasing the saturated NFET size, if maximum swing is not utilized from the start. Note, however, direct increase of SFD main transistor could corrupt the signal if too much miss-match between TERM and *M1* output resistance arises upon sizing. [21]

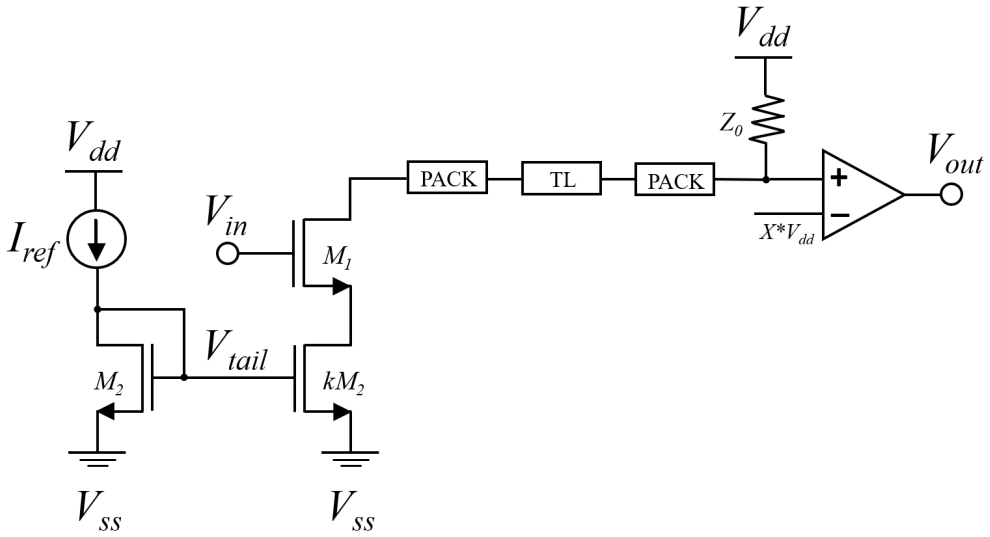


Figure 2.7: Schematic of Saturated FET Driver with a current source at its tail

Introduction of CMs effectively boosts the pull-down circuitry's output resistance by a factor equal to input transistor's transconductance (g_m) and current source effective r_o product. The former can be obtained from degenerated common source small signal analysis, for which the circuitry is shown in Figure 2.8 [27]. Increase in the pull-down resistance (R_{PD}) causes almost inversely proportional drop in the current pull-down circuitry can sink ($I_{sink} = \frac{V_+}{R_{PD}}$) and hence reducing the achievable drop on the TERM (refer to Figure 2.7). Notice, V_+ is defined as positive receiver terminal voltage. To achieve equivalent swing to the single transistor case, the area of cascoded topology would have to be significantly increased to reduce the overall resistance. Here, cascoded topology refers to the use of parallel transistor layout with respect to input-to-output path [24].

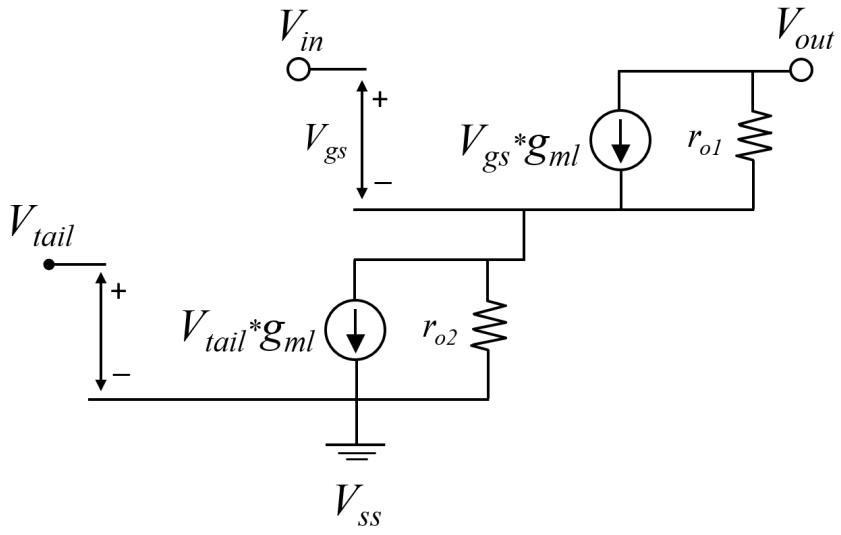


Figure 2.8: Small signal model of SFD with tail current source

Note, the gate voltage of the current mirror has to be barely above V_{th} of the device to keep it in saturation and ensure that drain voltage of the mirror is close to ground rail. Otherwise, the voltage swing provided by the input transistor would be compromised due the increased source voltage leading to lower current generation. Thereafter, CMs area is immense as device effective width has to compensate the almost quadratic voltage to current relation as given in long channel planar device saturation current expression (Equation 2.4 [27]). In the equation below I_D corresponds to transistor's

drain current, C_{ox} represents the device oxide capacitance, μ denotes the mobility of majority charge carriers, λ is channel length modulation parameter and W, L represent the width and length of the device respectively. When performing analysis to the first degree, λ can be neglected and assumed to be zero. Bear in mind, the equation is used just as an indication of the proportionality, rather than exact relation, as an exact equation for FinFET devices, including all second order effects would be unnecessary complex.

$$I_D = \frac{\mu C_{ox}}{2} \frac{W}{L} (V_{gs} - V_{th})^2 (1 + \lambda V_{ds}) \tag{2.4}$$

Lastly, the power consumed by SFD configuration has to be determined in a similar fashion as it was done for CTT. In this case the driver NFET is responsible only for the current amount passing through the circuit, thus controlling the swing - its intrinsic parameters can be omitted from the average power equation. Assuming ideal signals and infinite slew rate, the power consumption of SFD can be determined as follows:

$$P = \frac{0.5 V_{dd} \cdot V_{Z_0}}{Z_0} \tag{2.5}$$

, where Z_0 is the characteristic impedance of TL line, coinciding with TERM value at the particular case. Note, the current is proportional to the voltage drop across the termination resistance (V_{Z_0}) when NFET is 'on' (refer to Figure 2.6). A factor of one half is applied because the driver is opaque during half the cycle on average. In the other half pre-charege occurs and only leakage caused power dissipation is present, which is neglected for first hand analysis.

High Speed Transciever Logic

High speed transciever logic [28] usually consists of 2 series NFET devices instead of a complementary device pair and is suitable for a multitude of termination types. An example of HSTL with single parallel termination looks as depicted in Figure 2.9 - this configuration is also known as lower tapped termination (LTT) or near ground signalling (NGS) [29]. Using an NFET device instead of a PFET for the pull-up circuitry provides speed enhancement and area savings for planar devices. However, as the hole versus electron mobility ratio has been eliminated in FinFET device by increasing gate control over the channel, there is no significant benefit of using double NFET configuration in FinFET systems. Therefor, ability to use an inverter as driver for HSTL without a penalty leads to a similar configuration to CTT, with the main difference being the ability to change termination type. Nevertheless, using double NFET configuration leads to a reduction in power consumption due to swing limitations imposed by using NFET as PUN - V_{th} drop across PUN occurs [10]. If both input and supply voltage of TX can be varied at the same time, using CTT topology becomes more advantageous as similar power saving could be achieved at the cost of area.

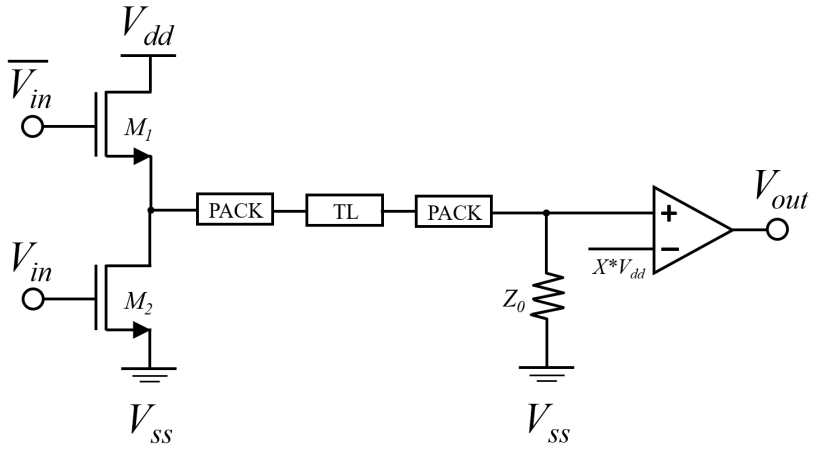


Figure 2.9: Example of HSTL topology

It can be noticed, that LTT is similar to SFD with the only change being full transparency with respect to the input. However, the issue of unbalanced rise and fall slew rates is still present in the design. The

main drawback of LTT compared to CTT is the reduced load bearing capacity as lower voltage signal gets propagated towards output. Moreover, RX of LTT has to be stronger than that of CTT due to weaker signal being propagated to RX input, which implies that RX generated current per unit area is going to be lower (refer to Equation 2.4).

The power for HSTL is not analysed here due to the uncertainty of termination type which would be selected for further analysis. However, the expression for LTT configuration is exactly the same as Equation 2.5 to first degree. With non-ideal signals the power consumed by LTT would be slightly higher than that of SFD where additional factor would be caused by PDN being operation for short period of time.

Note, using a TERM connected to V_{dd} rather than ground would lead to pseudo-open drain logic (PODL) design, which is effectively SFD with PFET PUN and parallel termination to ground. PODL has been shown to lag behind in achievable performance in comparison with LTT for low load designs [29]. Nonetheless, PODL can drive higher loads and achieve higher signal-to-noise ratio (SNR) if full rail-to-rail swing is used for output generation.

2.2.2. Differential Signalling Topologies

Differential signalling builds on the base of SES by simultaneously transmitting data and its complementary signal. Due to strong coupling between the complementary signals and both of them experiencing the same noise, such configuration leads to high electromagnetic interference (EMI) immunity and noise cancellation at receiver side, hence providing higher dynamic range. Additionally, DS usually dissipates less power than SES as power supply voltages can be diminished to achieve the same dynamic range and SNR. Commonly, DS also eliminates static power dissipation by getting rid of fully parallel TERM further boosting the power savings.

Even though DS is not the main data transferring method in state-of-the-art NAND memory applications, it has been used for transmission of supporting signals as *Data Strobe* and *Read Enable*. Both aforementioned signals require high slew rate, strong noise and EMI rejection and high purity, consequently making DS a perfect fit for them [13]. As I/O transmission speeds are bound to increase in near future, differential signalling will become a more popular design choice, as more challenges laid by crosstalk caused signal degradation will have to be overcome [23].

Pseudo-differential Signalling

Psuedo-differential signalling (PDS) operates by having a reference transmitted from TX to RX together with other single ended signals [21]. Having a set direct current (DC) reference voltage as comparison voltage is also the main difference between a differential and pseudo-differential signalling schemes. Therefor, PDS employs only some of the differential signalling benefits.

For instance, TX generated and propagated reference voltage experiences largely the same noise as signal lines, thereafter, the common mode noise can be almost entirely eliminated by differential RX. However, as DC voltage is effectively seen as ground by the high frequency signal lines, crosstalk induced signal variations are not coupled to the reference. Thereby, PDS does not provide almost any cancellation of crosstalk and EMI generated deviations. Additionally, if multiple signal lines are switching in the same direction simultaneously, severe switching noise would be observed on the reference causing improper comparison at RX amplifier. With crosstalk being the most detrimental degradation mechanism in high speed applications, it can be concluded that PDS is not significantly better than SES in the particular design scenario.

Even more, as single reference line can be shared between up to 4 data lines, the area is estimated to increase by at least 20% comparing to CTT [22]. Sharing the reference with more than 4 signal lines could lead to inaccurate reference voltage caused by too high fan-out and noise cancelling inefficiency due to physical distance between reference and farthest signal. Another downside of PDS is high power consumption inherent from single ended topology, augmented by power required to transmit the reference voltages. The exact power consumption depends on the selection of SES topology used as basis for PDS.

Scalable Low Voltage Signalling and Low-voltage Differential Signalling

Scalable low voltage signalling (SLVS) is an adjusted version of commonly used low-voltage differential signalling (LVDS) topology [30]. The current path and generation of differential voltage is shown in the images, where red corresponds to the case when $V_{in} =' 0'$ and green to $V_{in} =' 1'$ (non-inverting buffer

assumed). Observe, a buffer before the driver is placed which is not present in standard configuration. The buffer ensures that all incoming amplitude noise is converted into time noise (jitter), such that no unexpected oscillations or high frequency resonances occur on the TL and packaging interfaces. Else, in one stage design Miller effect creates a capacive link directly from input to output allowing certain amount of high frequency noise to pass straight to output [27].

By looking at Figure 2.10 one can notice the absence of current sources at inverter rails for SLVS topology, when compared to LVDS (Figure 2.11). Use of current source imposes a swing limitation on the driver transisotrs and TL, as a consequence, drive strength and power consumption of TX have to come at the cost of increased area.

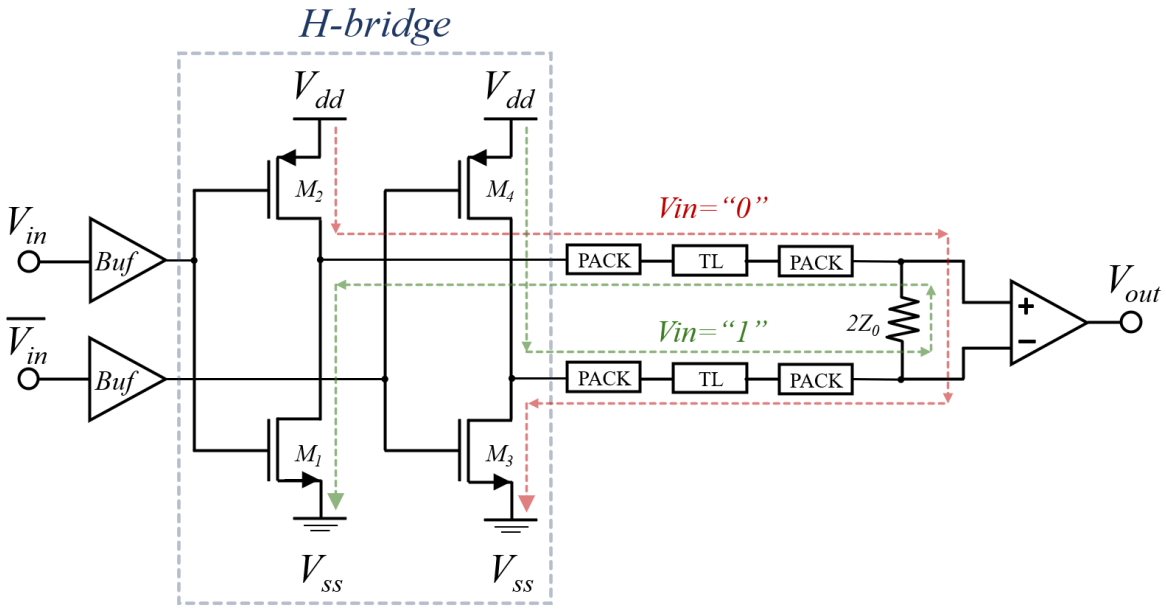


Figure 2.10: Scalable Low Voltage Signalling Schematic

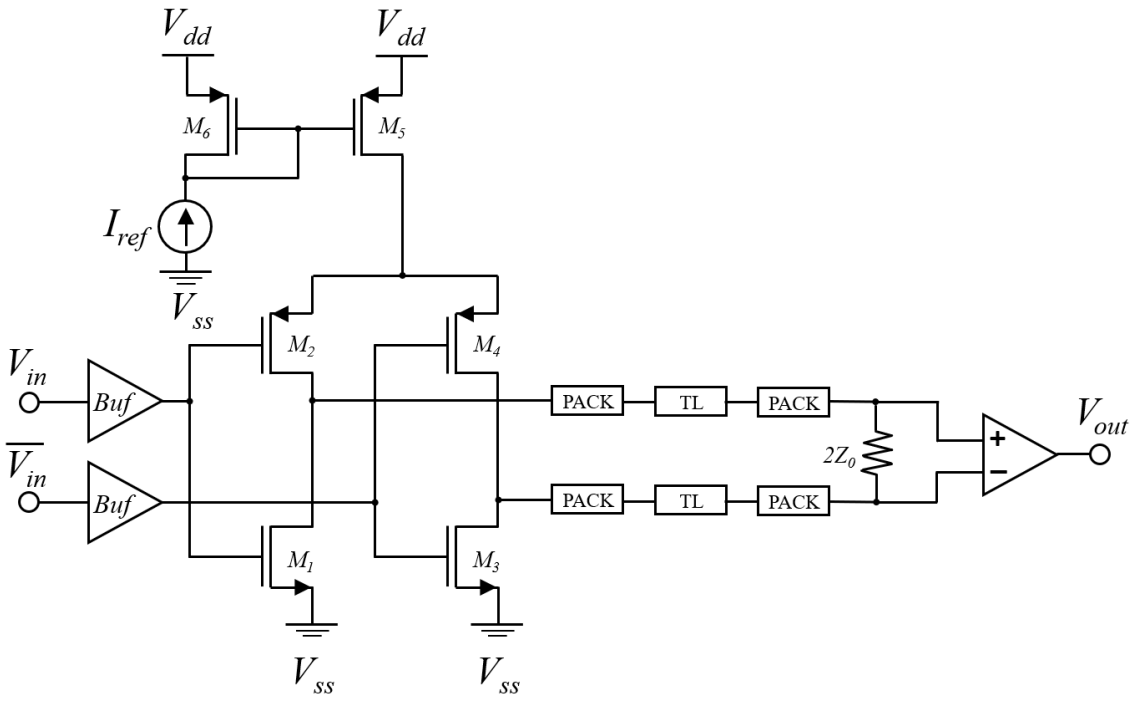


Figure 2.11: Low-Voltage Differential Signalling schematic

The main benefit the current source provides is a boost in overall circuits PVT immunity as current driven through the circuit is relatively stable across the various PVT conditions. However, for full stability, adjustable current source (similarly as discussed for SFD) would be required to ensure that constant current is sourced into the circuit.

Notice, CMs at LVDS rail provides improvements in immunity against voltage variations in a different manner than it is done in SFD. The driver transistors mainly operate in linear region, meaning that the current is highly drain-to-source voltage dependent, as seen in Equation 2.6 [27]. As current source is located in saturation regime, it is independent of V_{ds} to the first degree (refer to Equation 2.4). Thus, for voltage variations in the system, CMs V_{ds} varies such that close to equilibrium conditions of driver transistors are kept.

$$I_D = \frac{\mu C_{ox}}{2}\frac{W}{L}\left(2\left(V_{gs}-V_{th}\right)V_{ds}-V_{ds}^2\right)\left(1+\lambda V_{ds}\right) \tag{2.6}$$

Note, having only one current source in the circuit causes several deviations in the desired signal behaviour. For instance, the total output resistance seen by looking into the drain of pull-up transistor is increased by $gm \cdot r_o$ due to the source degeneration. The boost in output resistance of PUN causes a downwards shift of the V_{cm} seen when investigating voltages at positive and negative terminals of TERM at iso-current conditions. To balance the circuit, a current source at the ground rail of LVDS signalling could be used to equalize the swing on both sides. This would also ease the impedance matching and allow for higher value TERM resistance to be used without worrying about discontinuity caused reflections. However, use of two current sources leads to an immense increase in chip area, especially if they are set to be digitally controllable. Bear in mind, to ensure that both CMs generate the same current and do not cause deviations of signal over TERM a common mode feedback circuit is required [31].

Another option to balance V_{cm} of LVDS is skewing the design, making the PUN circuitry larger than the PDN to acquire approximately equal transistor output resistances. Such an approach, however, could cause a penalty on slew rate as intrinsic capacitors at iso-current case would be larger in PUN than PDN - more time to charge/discharge the devices is required. Additionally, impedance matching at the output would become a challenge, as with variations of process and temperature lack of symmetry in PUN and PDN impedances would cause deviations from the mean, leading to reflections. Using a shifted V_{cm} would require a stronger RX circuitry similarly as discussed for LTT (part of HSTL). [21]

It can be noticed that the main trade-off between SLVS and LVDS is PVT versus chip area. Using full rail-to-rail swing as is the case of SLVS would lead to higher current generated by driver transistors per unit area. Nevertheless, all the transistors would have to be significantly increased in size to compensate for signal degradation across PVT. Hence, the power consumption of SLVS could potentially be higher than LVDS, if increased nominal case swing is used to compensate for PVT. The power consumption in typical conditions (without compensation) can be determined as shown in Equation 2.7. Note, the expression is used only as power level identification since second order effects and design constraints will lead to additional power terms. The expression can be noticed to be exactly the same as given for SFD.

$$P_{tot} = \frac{V_{dd} \cdot V_{Z_0}}{2Z_0} \tag{2.7}$$

The main benefits of of SLVS and LVDS over every single ended topology discussed in Subsection 2.2.1 is low power consumption, increased noise resistance and superior slew rate performance [21]. The main power reduction mechanism lies in the fact that termination is effectively floating - does not create a current path directly between the power rails. Thereafter the circuit dissipates virtually no static power apart from leakage current generated path in transistors. Nevertheless, the drawback of using a floating termination is its inability to set the common mode voltage. Thereafter, if PUN and PDN strengths deviate from one another, a stronger '0' over '1' or vice versa could be acquired. The effect can be almost fully accounted for by accurately sizing both the buffer and the H-bridge.

Current Mode Logic

Current mode logic (CML), also called differential current steering (DCS), is a differential version of SFD, where complementary signal branch has been added (see Figure 2.12) [32]. The topology is a successor of bipolar junction transistor (BJT) based design extensively researched in the past named emitter-coupled logic (ECL) [33] [34]. Bear in mind, only a single termination at the load side is used - this will severely affect the signal quality due to poor matching throughout the circuit (reflections).

The benefit of using a differential pair with a current source at its tail is stable input transistor source voltage which allows for sharp turn-on transient due to intrinsic source capacitance pre-charge. Moreover, the slew rate of CML design is very steep as the device is always saturated, consequently peak currents can be reached in mid-swing of V_{in}. Lastly, the circuit draws almost constant current continuously (set by tail CMs) meaning that the circuit is less susceptible to power supply noise. [21]

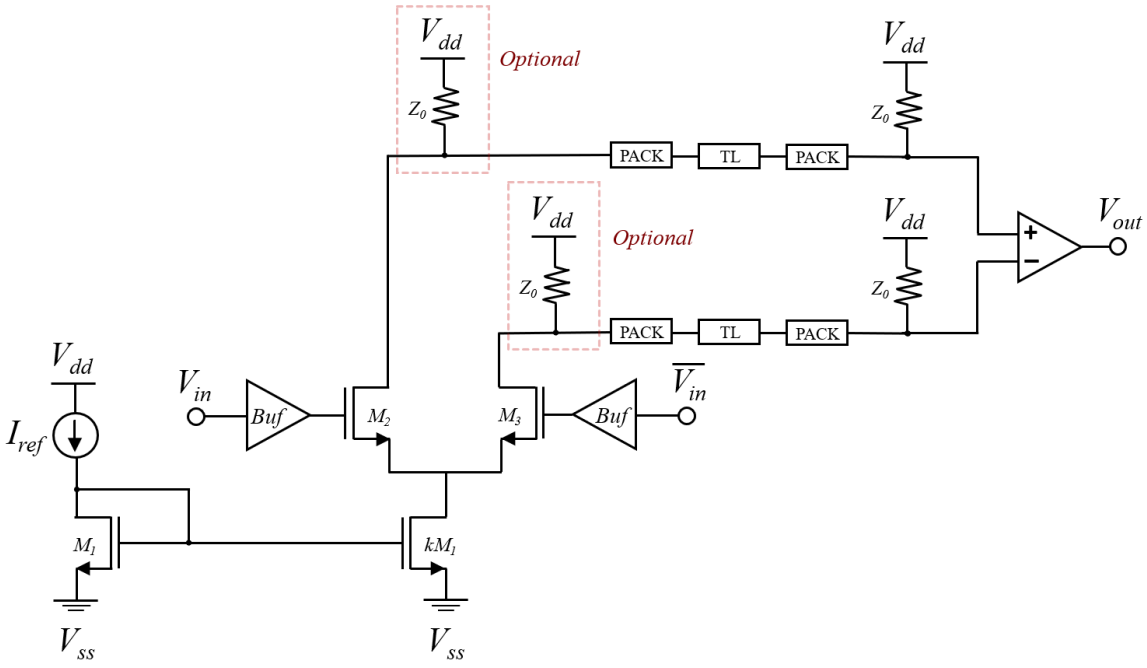


Figure 2.12: Current mode logic signalling schematic

The major benefit of using a CML topology is the possibility to achieve high bandwidth signals, if lower than unity gain is used - the gain-bandwidth product (GBWP) of a differential pair is relatively constant [24]. Additionally, CML is less liable to PVT variations than SLVS as all transistors operate in saturation rather than linear region (independent from V_{ds} to first order). Benefits of using CML over SES topologies are the same as for SLVS, e.g. better noise and EMI immunity, increased swing, apart from reduced power. Since the termination used in CML is half the TERM value of SLVS and it is inversely proportional to power (see Equation 2.7), the power consumption of CLM can be said to be at least double SLVS one for the same swing value. In a double termination configuration the power would be almost further doubled as the resistors are effectively seen to be parallel. TL and PACK resistances would account for an increase in load termination resistance value, however, the increase would not be too large. Notice, duty factor of each driver's NFET is ≈50%, hence the driver is always transparent, meaning that power expression does not require any additional terms.

There are more than the currently covered single ended and differential signalling topologies applicable for high speed applications. However, with CTT, SFD, HSTL, CML and SLVS serving as the main building blocks in high speed data transmission, the list covered in this thesis is not further expanded. With this in mind, the receiver topologies are covered next.

2.3. Receiver Options

For PU to be able to use signal incoming from NAND memory, the signal has to be first amplified, corrected for errors and sampled. All the aforementioned functions are governing characteristics of RX. Bear in mind, not all functions are strictly necessary as, for instance, asynchronous systems do not require sampling as signal gets continuously evaluated in such circuit topologies. Amplification of the signal is the first function of the RX as the incoming signal is usually heavily attenuated, where it could be starting to resemble noise, if sampled directly.

As signal amplitudes are generally small, even though the whole system can be assumed digital, the amplifier operates in analogue domain, where only small input voltage deviations are present. Amplifier has to be able to provide a high gain to ensure steep slew rate and good sensitivity to ensure sufficient, if not maximum swing signal which is then propagated to consequent RX stages. Take notice, to guarantee reliability of the chip, a headroom would have to be defined due to possibility of frequent overshoot. Headroom is defined as maximum swing value below power supply, which, if not exceeded, will not damage the system. [35]

The same way signal gets stronger when passed through an amplifier, so does the noise. As incoming signal is attenuated relatively close to the noise level, meaning that SNR is not too high, another desired property of an amplifier is noise suppression. For this purpose differential amplifiers exhibiting high common mode rejection ratio (CMRR) are used. Similarly, to prevent power supply noise from manifesting on the output, also the power supply rejection ratio (PSRR) should be high. Both of these considerations can be achieved by using differential amplifiers (partially available in SES configuration too). [35]

Amplifiers can be sorted in two major groups: static and dynamic circuits. The static amplifiers are continuously magnifying the incoming signal while dynamic circuits perform signal boosting only when an enable or clock signal is activating a transistor operating as switch on the current path from rail to output. Examples of static receivers are discussed in Subsection 2.3.1, while dynamic receivers are covered in Subsection 2.3.2. Note, dynamic amplifiers can also serve as sampling circuits at the same time if a latch has been added to its output.

Note, the amplifiers discussed further are single stage topologies only. In cases where a large output load has to be driven, cascade topology would be used, which consists of multiple stages, where each subsequent stage increases the overall amplification gain. The following stages can be built using the same amplifier topology as the previous one, as long as the inputs are compatible with the outputs - reuse of differential to single ended topology would be complicated and barely beneficial as 2nd stage requires a constant reference voltage. Bear in mind, use of multiple stages could cause stability issues due to increase in number of frequency poles along the path from input to output. Hence, multi stage amplifier would have to be designed with caution. [36]

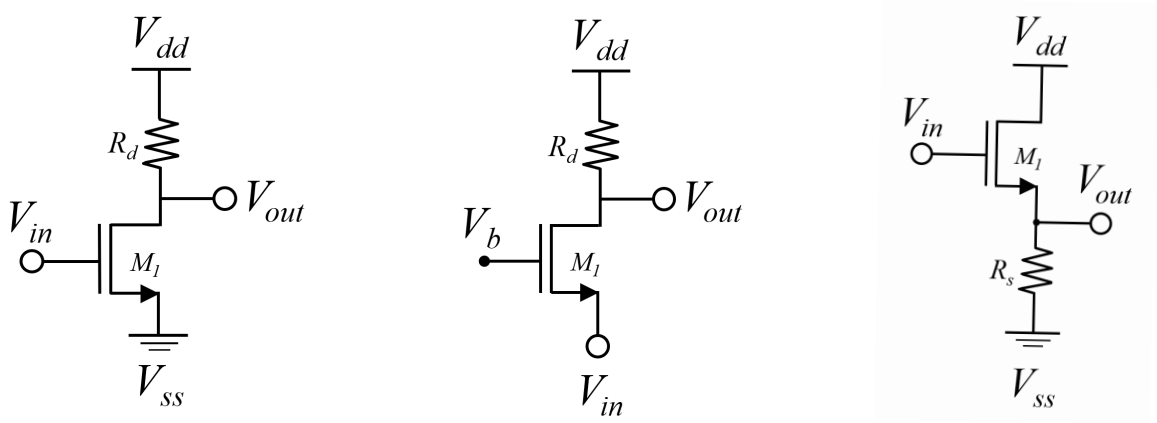
Similarly as for TX, in this section only the amplifier of RX is considered, with it being directly connected to the I/O pin of PU. Other RX components are neglected in this thesis assuming that the signal is error free and the sampling can be performed quicker than signal transmission. Note the former

assumptions are challenging to fulfill on their own, thereafter, it would be overwhelming to perform a complete system integration during the span of a single MSc thesis. Here only qualitative analysis of RX amplifiers is provided with more details provided in Chapter 3.

Notice, DRAM I/O speed per pin has been shown to reach 9 Gb/s margin [37], partially validating that sampling can be performed quicker than 8 Gb/s. However, as DRAM and NAND I/O interfaces substantially differ in modulation and device requirement schemes, adaptations to sampling and error correction circuits would be required.

2.3.1. Static Receivers

Static receivers are the most common form of amplifiers due to their simplicity, well known behaviour and low cost both area and signal requirement wise. The most basic static amplifiers, which serve as building blocks for other topologies are common source (CS), common gate (CG) and common drain (CD) (better known as source follower) configurations depicted in Figure 2.13a, Figure 2.13b and Figure 2.13c respectively. ”Common” in amplifier naming indicates transistors terminal, which is connected to static power supply, while the other two ports are connected to input or output. All of the aforementioned topologies in their most basic form consist of only a single transistor, which is complemented by a passive loading, such as resistor. Instead of the resistor, it is possible to use an active load as diode connected transistor to allow for easier in-silicon implementation. Additionally, drain termination resistance (R_d) can be replaced with a current source (DC biased transistor) to prevent output voltage swing fluctuations in case resistance value is altered for increased gain. [24]



(a) Common source amplifier **(b)** Common gate amplifier **(c)** Common drain amplifier

Figure 2.13: Simple amplifier topologies configured in a) common source b) common gate c) common drain

The most widely used topology from the aforementioned is the CS topology due to its characteristic high input impedance, relatively low output impedance (equal to R_d in parallel to r_o) and moderate gain at high frequencies. Ideally, the gain should be infinite, however, with technology node scaling to shorter gates lengths the intrnisic gain which a CS stage can achieve is reducing. The main culprit for this reduction is channel-length modulation (short channel effect) which starts to play a more dominant role in nanometer devices, leading to non-constant drain current vs. V_{ds} characteristics. [24]

CS operations can be described by current fluctuations on the load resistance. Note, CS looks exactly like SFD, implying that it mainly operates in saturation region, especially if output swing is low. Consequently, with small voltage variation at the input, transistor’s drain current starts to vary according to Equation 2.4 which is then projected on R_d. The current change is then transformed into output voltage change with the opposite polarity: increasing current reduces the output voltage.

CG configured amplifier provides almost identical characteristics (to first order) to those observed for CS stage. The main difference is that contrary to CS stage, CG provides a low input impedance which is attractive for low-power and high SNR amplifiers used in radio-frequency applications. The major benefit of having low input impedance is easy impedance matching, which is extremely useful in reflection prevention. Similarly to CS, CG amplifier suffers from gain vs. voltage headroom trade-off as both the gain and output voltage are proportional to R_d. Increasing the load resistance too much can cause the transistor to enter triode region, thus loosing most of the desired amplifier properties. [35]

Working principle of CG is slightly different than the one for CS - here, for small voltage change at the input, V_{gs} of the transistor is reduced by the same amount (direct proportionality to source voltage (V_s)), hence lowering I_D by g_m multiplied by input shift. Output voltage (V_{out}) on the other hand, is increased by rise in I_D times R_d. It can be observed that in contrast to CS, current gain of CG is equal to 1 as current from the input directly flows through the transistor and into the load resistance. Nevertheless, as already established, the same cannot be said about the voltage gain.

The only topology from the three basic configurations, which provides less or close to unity gain is CD configuration, making it a good option for a voltage buffer [38]. As drain voltage of the transistor is fixed, changing the gate voltage has to result in equivalent variation on the source voltage for V_{gs} to remain constant - $V_{gs} = V_{in} - V_{out} = V_g - V_s$, where V_g is transistor's gate voltage. The former is confirmed by small signal analysis, which returns a gain value of approximately 1. Notice, even though the V_{gs} remains constant, the drain current varies due to $I_D \propto V_{ds}$ caused by channel length modulation. [24]

Bear in mind, the aforementioned single stage amplifiers can be modified to differential topology. For instance, interconnecting sources of 2 CS stages will lead to most simple form of differential pair (DP), similar to what is used to drive CML. However, such configuration would not be useful in real life applications as output common mode voltage is not established, which leads to V_{in} dictating output V_{cm} properties. To stabilize output common mode voltage, a current mirror at DP tail can be added, which prevents undesired characteristics as clipping in case common mode on V_{in} is severely skewed towards either of the rails. Full differential pair would provide improved signal swings, common mode and power rejection, noise suppression and relatively high gain. [35]

With the three basic amplifier topologies covered, more sophisticated amplifier models can be discussed. The following topologies are merely a modification of the amplifiers explored in preceding paragraphs. Note, each modification provides different benefits while trading-off certain performance parameters - nothing in electronics comes for free.

Inverter as Amplifier

So far discussed topologies mainly aim to amplify analogue signals. By using an inverter one can immediately digitize the signal for further use. Topology wise it looks like a CS with resistive load changed to a pull-up active device, meaning that if properly sized, inverter becomes an inherently symmetric circuit. However, for it to generate equal strength '0' and '1', the common mode of the input signal should be well defined at half the rail voltage.

Inverter as an amplifier can be comparable to low-gain stage with inability to control V_{cm}. Thereby pre-requisites of V_{in} processed by inverter would be well defined common mode voltage, high SNR and large swing. Knowing that in high speed digital signalling such requirements are almost impossible to fulfill, especially for external interconnects [23], using inverter as the first stage of the amplifier is not feasible. Inverter as a second or further stage, however, could provide various benefits to the receiver. For instance, it would digitize the signal, prepare it for sampling and perform voltage level adaptations if different power supply is used for its rail voltage. [21]

Major drawback of inverter as amplifier is static power dissipation in high-gain mode due to both transistors being operational and likely in saturation at the same time. Also, due to the low-gain and thus low-sensitivity, high SNR of the incoming signal is required for inverter to be able to provide sufficient benefit to signal amplification.

For differential operation purposes an inverter would have to be added to each line, which are later passed through differential-to-single ended transforming stage. Note, the second stage would likely be amplifying as well.

Five Transistor Operational Transconductance Amplifier

Five transistor operational transconductance amplifier [35] (OTA) is composed of a differential pair, where the resistive termination has been replaced with a current mirror as shown in Figure 2.14. The main difference between OTA and DP is the generation of single ended output from a differential input, which is achieved by the current copying operation at the load. The circuit operates by amplifying differential signal through changes in branch currents. For instance, in case a small, positive differential V_{in} change is experienced, I_D of the left branch increases while the one on the right decreases. As current through right NFET falls, the output is pulled towards the upper rail. Assuming that PFET size ratio is 1:1, the current growth of left side is copied to the right further increasing the V_{out} voltage

towards the power supply rail. Thereafter, the overall effect is such that input output gets enhanced by both paths simultaneously.

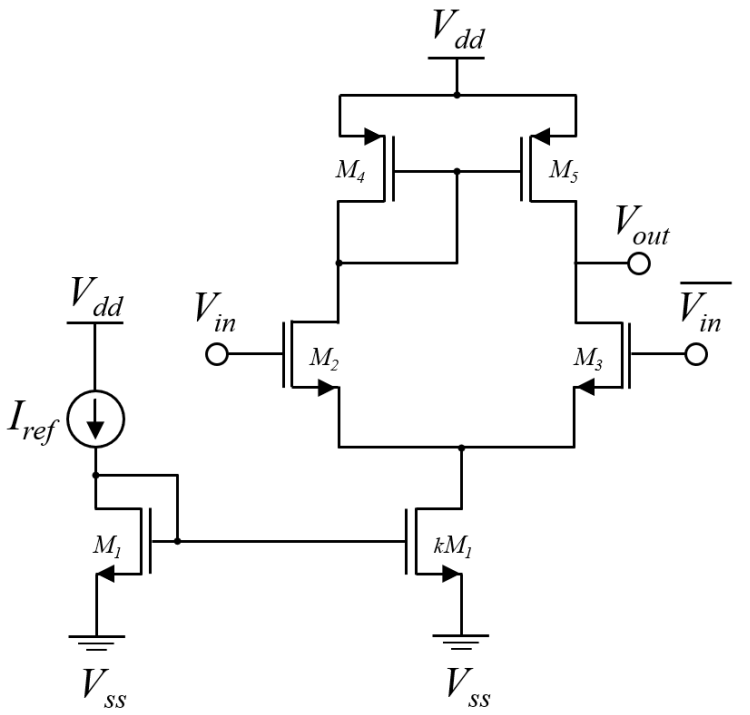


Figure 2.14: Five transistor OTA with NFET differential pair and PFET loading

One can observe that the circuit is asymmetric even when each type transistors have the same relative size. Asymmetry and conversion from differential to single ended signalling comes with certain drawbacks as low PSRR as voltage variation on rail almost directly manifest on the output. Also, CMRR is limited, implying that a fraction of V_{cm} changes at the input are propagated to the output. Process and temperature variations even further enhance asymmetry of OTA circuit, which can lead to undesired shifts in output common mode and current insufficiency in the circuit. Moreover, a frequency pole from the non-output side of the OTA is effectively mirrored to the output degrading gain's frequency response. [35]

One major benefit of OTA stage is that conversion to single ended output can be achieved with voltage gain equivalent to that of a simple differential pair. Even though theoretically the gain of OTA should be halved comparing to DP due to use of single ended output, use of active load devices boosts the gain to nominal value - aforementioned enhancement of branch currents. However, due to double pole at the output the gain sharply declines with increased frequency making this topology suitable for limited BW range applications. To achieve higher gain and thereby stronger amplification, amplifiers of telescopic or folded cascode can be used as discussed further.

Folded Cascode and Telescopic Amplifiers

Telescopic and folded cascode amplifiers build upon differential pair using similar ground rules - to increase gain, the effective output resistance of the circuit should be increased. For this purpose cascoding is performed by adding active loads acting as voltage controlled current sources in parallel to input transistor as many times as it is required. Nonetheless, increase in gain comes at a cost of reduced output voltage swing, implying that subsequent amplification stages would be required to achieve the desired levels. A two level example of each topology is represented in Figure 2.15 and Figure 2.16.

The basic structure of cascode amplifier is a CS stage which converts input voltage into current further fed into a CG stage, where it is forwarded into the active loading. Folded cascode builds on the concepts of telescopic amplifier by using a different device type for CS and CG stages which provides certain advantages discussed further [24]. Before that, common characteristics of both amplifier configurations are discussed.

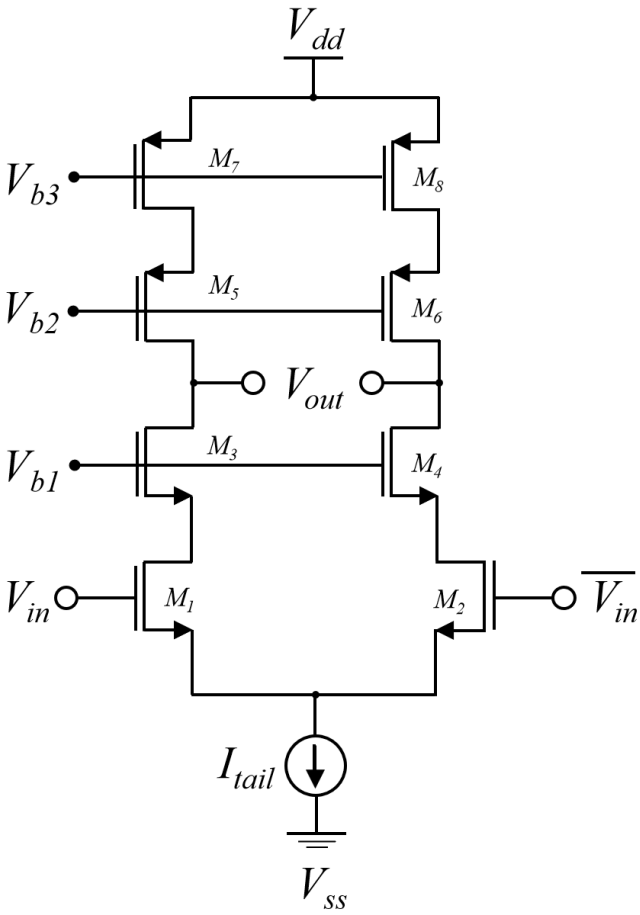


Figure 2.15: Two level telescopic amplifier

Both folded and telescopic amplifiers use higher number of active devices than simple amplifiers, thus generating more frequency poles due extra intrinsic capacitors on the signal path. As a consequence, gain-bandwidth response gets substantially affected causing steeper roll-off, especially if differential to single ended layout is used where double (mirrored) poles appear at the output. Single ended cascode amplifiers can potentially face stability issues due to large amount of poles, which can be located in vicinity of one another, leading to ever increasing oscillations and amplification of undesired frequency components. [24]

In order to provide maximum swing the telescopic amplifier is able to achieve every single input supporting transistor has to be biased just at the edge of saturation region. Thereby, budget for gate voltage controlling is very strict as significant deviations from desired properties could cause the amplifier to exhibit either immensely reduced swing or fully cut-off properties. To ease the burden of tight biasing conditions and at the same time increase the output swing of telescopic amplifier, folded cascode stage can be applied. Use of such configuration allows higher variation of V_{cm} at the input and regulation of common mode at the output with minor penalty on the voltage swing. Additionally, as input transistors are not a part of cascode structure, but are rather parallelly connected, contrary to telescopic amplifier case, the tail current source of differential pair does not impose any limitations on the voltage swing. However, an additional current source is now required to bias the cascode structure, implying that folded cascode power consumption is higher than that of telescopic amplifier. Additionally, as both, currents from DP and cascode structure are going to be sunk through common tail, the size of *M3* and *M4* would have to be large. Alternatively, the overdrive voltage ($V_{gs} - V_{th}$) of the transistors can be increased to amount for the same effect. Bear in mind, the latter would lead to higher swing limitations as V_{ds} of the common tail would be slightly enhanced. [35]

As for the gain of folded cascode vs. telescopic - from Figure 2.16 one can see that input transistors are in parallel with the tail NFETs of the cascode, implying a reduction in effective output resistance.

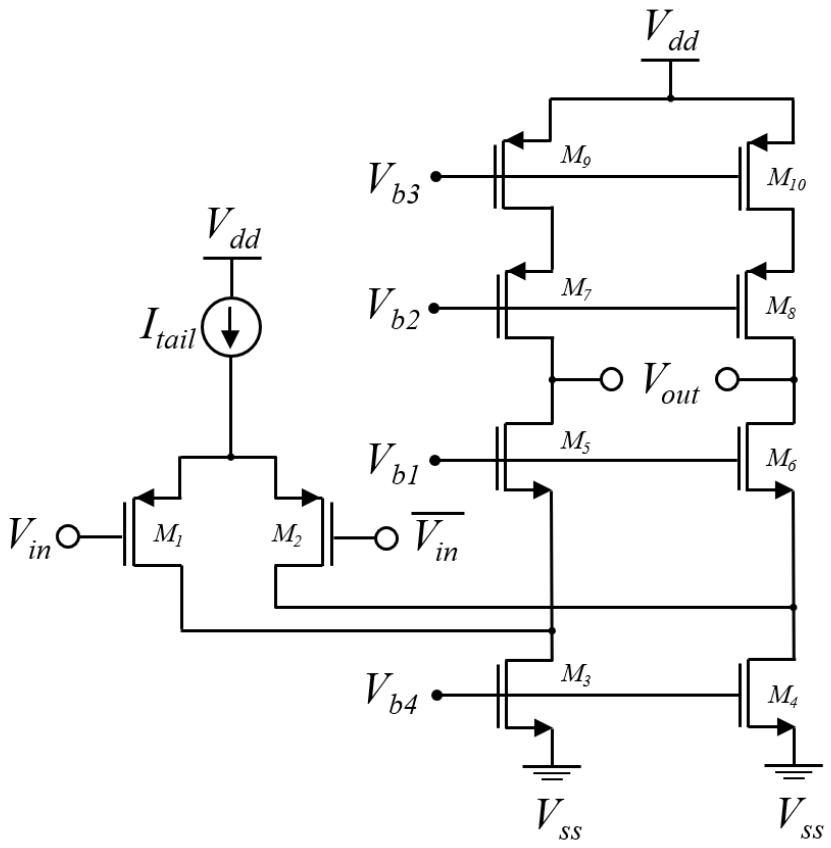


Figure 2.16: Two level folded cascode amplifier with PFET differential pair

Additionally, if transistors *M3* and *M4* from Figure 2.16 are made large for current compensation, the effective parallel resistance of common tail and input DP becomes fairly low comparing to the output resistance of telescopic topology. Lastly, extra challenge seen by planar devices is smaller g_m in PFET devices when compared to NFET leading to substantially reduced gain value of folded cascode in comparison to telescopic configuration. Using NFET devices for DP would enhance the gain but cause frequency pole shifting closer to DC frequency. [35]

Overall, in comparison with telescopic amplifier folded cascode provides a significant enhancement in V_{cm} control and a marginal benefit on voltage swing at the expense of increased power consumption, lower voltage gain and reduced pole frequency. The effects would get deteriorated even further in case a single ended topology of the amplifier would be used - voltage swing would be close to halved and a mirror pole would limit achievable BW. With this, general concepts and most well known static receivers are covered allowing to shift the attention to the dynamic RX subset.

2.3.2. Dynamic Receivers

Before delving into the concepts of dynamic receivers, some limitations to information discussed in this chapter have to be established. With the rapid advancements in data propagation between ICs use of dynamic receivers to directly amplify received burst I/O signal has dwindled over time mainly due to challenges posed by high frequency operations. The main reason for the operational delay is the high (ideally infinite) resistance characteristics of dynamic amplifiers, which is combined with high capacitance at the output node. The multiplication of both parameters is inversely proportional to BW which limits the maximum operating frequency of the system. Nonetheless, as output resistance of common dynamic amplifiers is high, they achieve great sensitivity with respect to their input signals, which is evidently one of the most appealing characteristics of such configuration.

Let's take a conventional sense amplifier shown in Figure 2.17 [39], where both direct and complementary outputs have been identified. It can be noticed, that output is set to have a positive feedback

loop due to cross-coupled inverters. Hence, the circuit is able to change the output state very fast, indicating high closed loop gain.

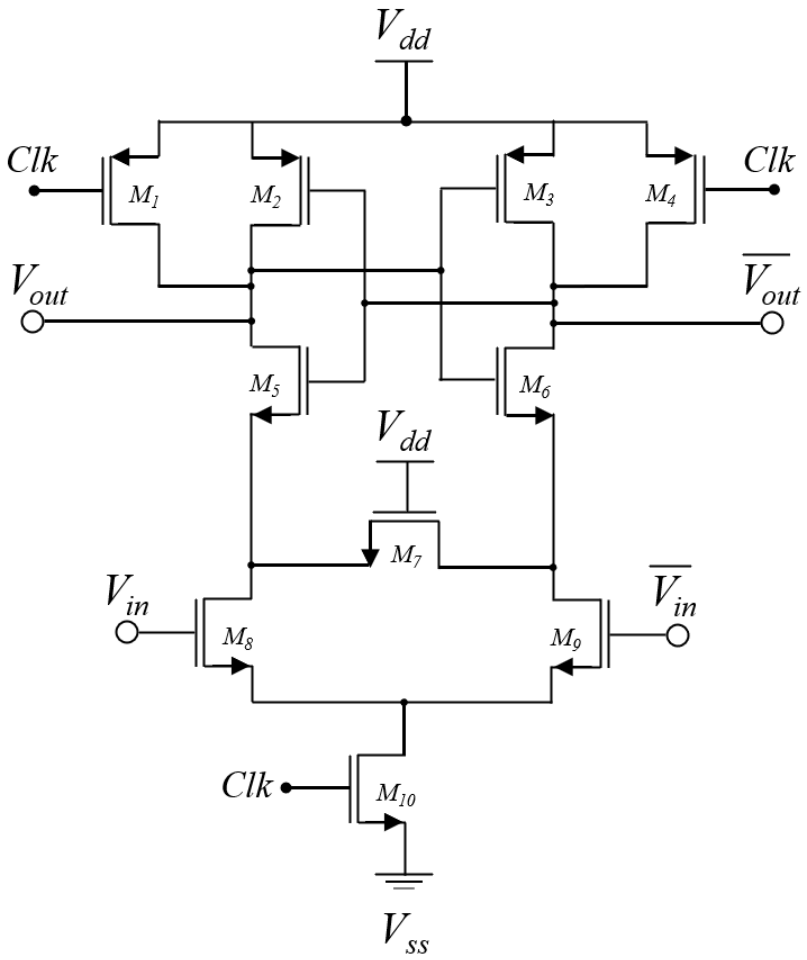


Figure 2.17: Conventional sense amplifier topology

When considering the open-loop gain of the amplifier, the output resistance of input to output has to be investigated for one branch. For this purpose, it is assumed that V_{in} is equal to '1' when the clock changes from low to high. One can notice that initially output resistance is equivalent to series resistance of 3 transistors (NFET at input, tail and cascode) in parallel to inverters PFET device. In case a second, stabilizing feedback loop between input transistors is used, the resistance is further reduced at initial stages due to a parallel current path. After certain settling time the output resistance can be observed to increase due to PFET transistor turning off, thus disabling parallelization of resistances. Having only the series resistance leads to a very high DC gain value even in comparison to high gain static telescopic amplifier. Note, the NFET feedback is used to avoid metastability of the circuit in case data shift happens at high clock stage [40].

However, gain is almost inversely proportional to bandwidth of an amplifier - constant GBWP to first degree - increased sensitivity and slew rate comes at the cost of more limited BW [21]. The other factor degrading bandwidth is the already mentioned intrinsic output capacitance. Investigating Figure 2.17, one can notice that capacitance of 4 devices is directly connected to the output node even when Miller effect is neglected - drain capacitance of direct inverter transistors and gate capacitance of cross-coupled inverter.

Alternatively, low BW can be explained by noticing that sense amplifier operates just like an integrating circuit. During pre-charge period (clock is low) output node and thus all device capacitors are charged to high voltage state which is close to the power supply. Former implies that all the intrinsic capacitors hold some finite value of charge, which can be assumed to be constant, if severe leakage

is not present. Thereby, it can be said that when input is applied, circuit does not react with self loading/discharging of capacitances if evaluation phase (high clock) has not kicked-off. During evaluation phase (high clock) voltage deviation at input is starting to apply current changes in the circuit, implying that the capacitances are slowly discharged. This can be better visualized using first order charge equation given in Equation 2.8, where charge deviation is denoted by ΔQ, C is system's capacitance and t represents time.

$$\Delta Q = C\Delta V = I\Delta t \tag{2.8}$$

Assuming all of the charge on input gets redirected to changing charge at the output one can notice, that longer applied input voltage leads to higher charge drop on the output. Since output capacitance is relatively large, it can be noticed that for short impulses almost no voltage change at the output is going to be present. As time is inversely proportional to frequency, it can be seen that frequency must be substantially lowered to allow for sufficient response and settling time. Conventionally, sense amplifiers rarely exceed 1 GHz frequency as then the high gain and thus high sensitivity characteristics are partially lost [41]. Even more, as evaluation of circuit happens only once during a clock stage, extra design considerations would be required to support double data rate (DDR) clock, where sampling is performed on both the rising and the falling edge. Other disadvantages include clock feed-through in case steep slew rate clock is used and charge leakage in case of improper timing conventions [27]. The former effect can cause signal overshoot beyond power rails causing reliability issues in long term, while the latter can lead to bit errors if noticeable amount of charge is lost before evaluation.

Even though BW of dynamic amplifiers is severely limited, for most operations their advantages outweigh the drawbacks. For instance, combining sampling and amplification in one step leads to major power reduction in comparison to static amplifiers. Here amplification is performed only at the instant of sampling, contrary to continuous boosting present in static amplifiers. Additionally, dynamic amplifiers provide lower timing uncertainty as delay caused by routing between flip-flop and an amplifier is eliminated. Lastly, dynamic amplifiers can distinguish lower voltage inputs due to their superior sensitivity extensively discussed before. [21]

To increase the operating speed of sense amplifiers, they can be implemented as demultiplexers - parallel stages operating on clocks delayed by data bit width. In such a way sufficient time can be given to each stage to evaluate its output even when data has switched due to the feedback implementation. Using parallelized configuration allows for combination of three actions in one step - amplification, sampling and deserializing. Thereafter, immense power and area savings can be achieved, even with using static latches right after sense amplifiers. [21]

Other topologies of clocked amplifiers include residue amplifiers, which amplify residual signal of, for example analogue-to-digital converters. A fully differential implementation of such amplifier based on inverter stage is provided by Akter et. al. [42]. Note, as BW of such amplifiers is significantly below target frequency no further discussion on alike amplifiers is continued. With this in mind, the next component of I/O interface can be investigated. With TL being one of, if not the biggest contributor to system parasitics, it is investigated next.

2.4. Transmission Line Overview

Conventional NAND flash memory employs an external interconnect to its designated PU. Such link is usually made on a printed circuit board (PCB) which allows for easy integration and bonding of components and at the same time attaining low overall system size. As memory and PU are located a finite distance between one another due to various regulations, one can conclude that transmission line analysis is required to determine the properties of interconnecting trace between them. The trace can said to follow the transmission line theory with certainty as the distance between I/O pins is significantly longer than the transmitted signal wavelength. This claim is validated in Subsection 2.4.1. Properties of conventional TL and their simplified models are discussed in Subsection 2.4.2.

2.4.1. Validity of Transmission Line Assumption

With the benchmark of the system being 8Gb/s, it can be easily shown that almost any wired link between two chips is going to act as a transmission line. For TL theory to hold, interconnection has to be longer than a fraction of product between input signal's rise time and velocity. Otherwise the circuit can be assumed to be lumped, as its response can accurately capture the quickest transient

component (rise or fall) of the signal. Distance the signal travels during its transition can be estimated using Equation 2.9, where t_r is the rise time, c is the speed of light in vacuum, v is propagation velocity of signal and ε_{eff} is the effective dielectric constant. The traveled length as given by the equation below corresponds to the maximum distance the trace can span in order to be modeled directly as lumped elements. To prevent under-designing of the system, the rise time - velocity product is divided by a factor of 6 [43]. Note, the scaling factor varies per source, thus obtained rise time - velocity product is taken only as a first hand approximation.

$$l_{prop} = \frac{v \cdot t_r}{6} = \frac{c}{6\sqrt{\varepsilon_{eff}\,(\varepsilon_r)}} \cdot t_r \tag{2.9}$$

ϵ_{eff} is assumed to be a function of relative dielectric constant (ϵ_r), as in multitude of TL cases the dielectric constant constitutes of a combination between two different insulating material dielectric constants. Moreover, fringing effects for narrow waveguides lead to an increase in the effective dielectric constant. [44]

Assuming an ϵ_{eff} to be roughly 3 and rise time equivalent to approximately 15% of the pulse width (125 ps), length from which TL theory starts to dominate is acquired to be around 0.55 mm. As all the interconnections on the PCB are going to be longer than 0.55 mm TL theory can be said to hold. The distance from any I/O pin in standard open NAND flash interface (ONFI) conventions exceeds half a centimeter just to the edge of the package alone. [13]

2.4.2. Distributed Transmission Line Approximation

Simple analytical model for determination of TL structure properties with satisfactory accuracy is the key in understanding the line implications on the transmitted signal response and quality. With the model and its impact determination intact, more sophisticated designs employing electro-magnetic (EM) solvers can be set up to acquire the high accuracy behavioral modeling required. The base model of the TL structure can be approximated with distributed lumped elements.

The distributed model consists of resistance (R) in series with inductance (L) and capacitance in parallel to conductance (G) connected between the signal and return paths as shown in Figure 2.18. The resistance and conductance components contribute to the signal attenuation, where R is made up of both alternating current (AC) and DC components with former accounting for skin effect and copper roughness. G represents direct current leakage through dielectric material from signal trace towards return path. The inductance is proportional to forward path self inductance of the trace for single ended case, while combination of self and mutual inductance is used for the differential case. The parasitic capacitance is generated due to the finite distance between signal and return paths, and/or the complementary line. The combination of L and C govern the behaviour of the signal across the TL by being the dominant influence on the voltage to current ratio of the line. This effect can be explained with L and C behaviour being strongly correlated to frequency contrary to that of R and G.

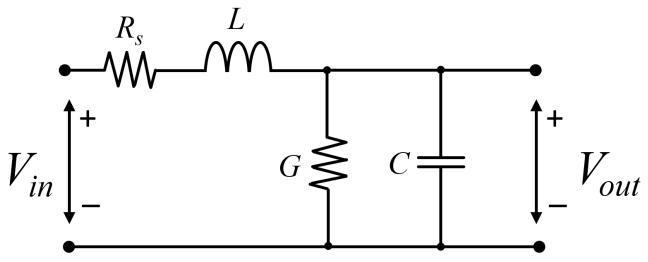


Figure 2.18: Distributed transmission line model

To properly capture the vast majority of transmitted signal characteristics, the length of each distributed instance (l_{TL}) has to be limited to a finite distance, which can be expressed as follows [45]:

$$l_{TL} = \frac{c}{20 \cdot f\sqrt{\varepsilon_{eff}\,(\varepsilon_r)}} \tag{2.10}$$

, where f is the dominant frequency of the signal. A factor of $\frac{1}{20}$ is incorporated to ensure that phase shifts along the line can be accurately detected. With the dominant frequency of the signal being 4 GHz and a first hand approximation of ϵ_{eff} being 3, l_{TL} is determined to be ≈2 mm. Thus, for simplified scaling of the model a unit length of the distributed element is assumed to be 1 mm for further distributed TL analysis.

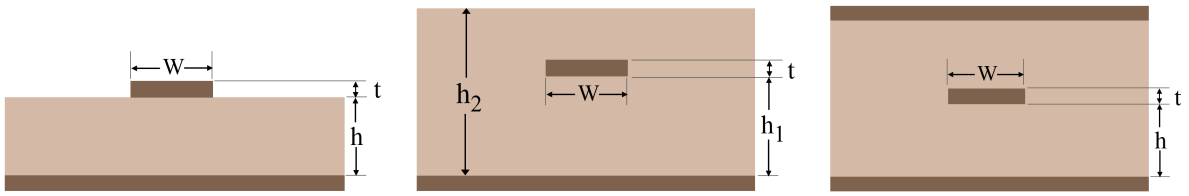
Characteristic Impedance

One of the most important factors of the TL is its characteristic impedance (Z_0), which provides an insight in voltage to current ratio across the line. The impedance can be expressed in RLCG parameters as given in Equation 2.11, where $\omega = 2\pi f$ is the angular frequency [43].

$$Z_0 = \sqrt{\frac{R + i\omega L}{G + i\omega C}} \tag{2.11}$$

The industry standards utilize a characteristic impedance of Z_0 =50 Ω, as it strikes a good balance between lowest attenuation and power consumption [4] in most designs, depending on the dielectric material selection. Thereafter, for design of any TL in the system a Z_0 value of 50 Ω is used as a rule of thumb. Nevertheless, if a stronger coupling between signal and return paths is required to retain quasi-TEM mode propagation, the characteristic impedance can be lowered. [46]

Knowing desired Z_0 value of TL allows to develop physical sizing strategies for various types and configurations of the interconnect. There is a high variety of different PCB trace types, from which most popular are microstrip (MS), embedded microstrip (EMS) and stripline (SL) shown in Figure 2.19a, Figure 2.19b and Figure 2.19c respectively. Notice, EMS provides better coupling to return path than MS as fringing field effect is reduced due to trace burying into dielectric material. The highest coupling and thus highest dielectric constant is provided by SL as two return paths are present in the design [47].



(a) Microstrip line Cross-section **(b)** Embedded microstrip line Cross-section **(c)** Stripline Cross-section

Figure 2.19: PCB trace type known as a) microstrip b) embedded microstrip c) stripline [46]

Each of the aforementioned line types have well established and verified empirically determined formulas which relate line's geometry to its impedance. For instance, by varying trace width (W) and dielectric substrate height (h), reference Z_0 value can be achieved with few iterations. Formulas for MS characteristic impedance [48] and EMS impedance [49] can be seen in Equation 2.12, where EMS Z_0 determination requires an extra step given in Equation 2.17. Note, a constant thickness (t) of 0.035 mm is assumed for the trace, which is the standard thickness of PCB copper layers including tolerances [5].

$$\varepsilon_{eff} = \frac{\varepsilon_r + 1.0}{2} + \frac{\varepsilon_r - 1.0}{2}\left[\frac{1}{\sqrt{1 + \frac{12h}{W}}} + 0.04\left(1 - \frac{W}{h}\right)^2\right] \tag{2.12}$$

$$Z_0 = \frac{60}{\sqrt{\varepsilon_{eff}}} \cdot \ln\left(\frac{8h}{W} + \frac{W}{4h}\right), \quad if\ \frac{W}{h} < 1 \tag{2.13}$$

$$\tag{2.14}$$

$$\varepsilon_{eff} = \frac{\varepsilon_r + 1.0}{2} + \frac{\varepsilon_r - 1.0}{2}\left[\frac{1}{\sqrt{1 + \frac{12h}{W}}}\right] \tag{2.15}$$

[4] URL https://www.digikey.com/en/blog/the-reasons-for-50-ohm-and-75-ohm-transmission-lines [cited on 6th of February 2023]

[5] URL https://www.eurocircuits.com/blog/tolerances-on-copper-thickness-on-a-pcb/ [cited on 15th of July 2023]

$$Z_0 = \frac{120\pi}{\sqrt{\varepsilon_{eff}}} \times \frac{1}{\left(\frac{W}{h} + 1.393 + 0.677 \ln\left(\frac{W}{h} + 1.444\right)\right)}, \quad if \ \frac{W}{h} > 1 \tag{2.16}$$

$$Z_{0_{embed}} = Z_0 \left[\frac{1}{\sqrt{e^{\frac{-2b}{h_1}} + \frac{\varepsilon_r}{\epsilon_{eff}} \left(1 - e^{\frac{-2b}{h_1}}\right)}} \right] \tag{2.17}$$

$$\tag{2.18}$$

$$b = h_2 - h_1 \tag{2.19}$$

Similar equations are defined for symmetric SL characteristic impedance [50] and can be found in Equation 2.20. Bear in mind, all the equations provided for characteristic impedance calculations of different line types are empirical, thereafter, they reach the highest accuracy around the reference point for measurements, which is set to coincide to Z_0 =50 Ω. Hence the accuracy of the equations employed in the calculations can be said to be better than 10% for the range of interest [46] and no crosstalk present. It has to be noted that the formulas are frequency independent - the approximation is performed assuming high frequency conditions, where L and C parameters are substantially larger than attenuation counterparts. Using these equations provides a good first order approximation of trace dimensions which can be directly fed into more sophisticated tools utilizing finite element methods (FEM).

$$Z_0 = \frac{60}{\sqrt{\varepsilon_r}} \ln\left(\frac{4b}{\pi D}\right), \quad if \ \frac{W}{b} < 0.35 \tag{2.20}$$

$$\tag{2.21}$$

$$b = 2h + t \tag{2.22}$$

$$\tag{2.23}$$

$$D = \frac{W}{2} \left\{ 1 + \frac{t}{\pi W} \left[1 + \ln\left(\frac{4\pi W}{t}\right) \right] + 0.551 \left(\frac{t}{W}\right)^2 \right\} \tag{2.24}$$

$$\tag{2.25}$$

$$Z_0 = \frac{94.15}{\left(\frac{W/b}{1 - t/b} + \frac{\theta}{\pi}\right)}, \quad if \ \frac{W}{b} > 0.35 \tag{2.26}$$

$$\tag{2.27}$$

$$\theta = \frac{2b}{b-1} \ln\left(\frac{2b-1}{b-1}\right) - \frac{t}{b-1} \ln\left(\frac{2bt - t^2}{(b-t)^2}\right) \tag{2.28}$$

Determination of R, L, C and G

To approximately determine the values of individual distributed line components, several assumptions simplifying the design are made. In order to determine L and C of the circuit, TL is assumed to be lossless. R has a complex component proportional to $\sqrt{\omega}$ (seen in Figure 2.20), which contributes to the effective inductance value [44]. However, it can be seen that ωL term becomes significantly larger (up to 2 magnitudes) than AC resistance (R_{AC}) at high frequencies validating the lossless assumption for L and C determination, causing maximum errors of 5%. The equation for lossless characteristic impedance Z_0 and phase velocity v simplifies to expressions given in Equation 2.29.

$$v = \frac{1}{\sqrt{LC}} \qquad Z_0 = \sqrt{\frac{L}{C}} \tag{2.29}$$

As phase velocity is known from Equation 2.9 and Z_0 is set to industry standard rule of thumb, L and C can be computed as a system of 2 equations with 2 unknowns. It has to be noted, that calculated L and C values are given in units per meter. Moreover, the calculated value for the inductance is appropriate for use only in high frequency computations as internal inductance is neglected due to

the skin effect onset prior to frequency used. Closer to DC the current penetrates into the material increasing inductance past the external (skin) inductance margin. [44]

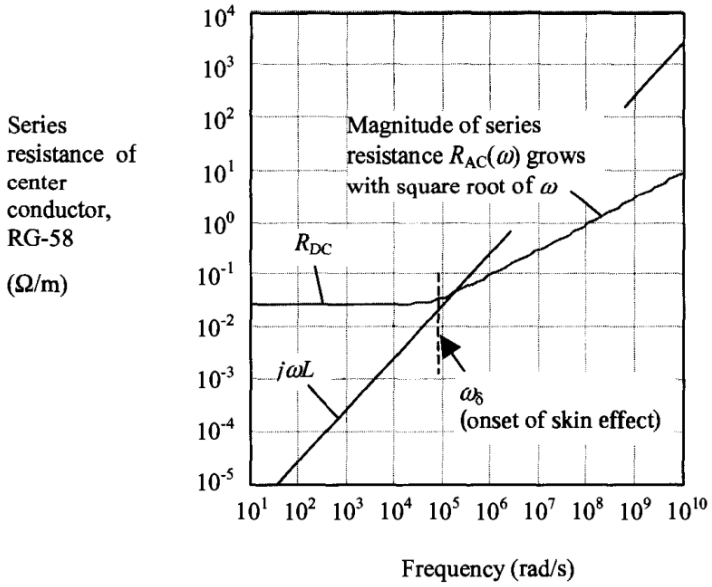
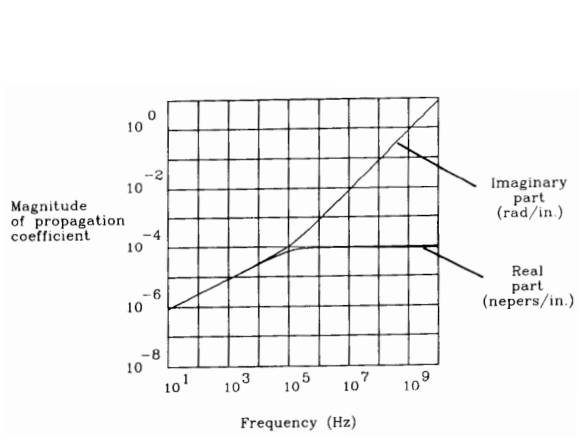
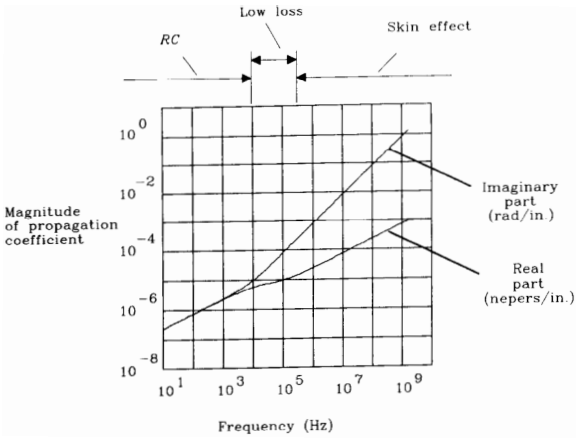


Figure 2.20: Resistance variation with frequency [44]

To verify validity of lossless assumption, graph giving frequency constant attenuation (Figure 2.21a) and graph depicting frequency varying attenuation (Figure 2.21b) [43] can be compared. It can be seen that the error caused in Z_0 computations assuming a lossless line is indeed small for frequencies in GHz range. Additionally, skin effect is less pronounced for low thickness conductors because the skin depth critical frequency is reached at a higher value.



(a) Frequency constant attenuation [43]



(b) Frequency varying attenuation [43]

Figure 2.21: Attenuation constant and wave number accounting for resistance with a) no skin effect b) skin effect [43]

Another form of verification can be performed by using empirical expressions directly determining L for rectangular conductors and C for asymmetric parallel plate capacitors. In general, inductance computations for traces significantly longer than their width provide an error of up to 2% [51], allowing for a quick and efficient verification tool prior to using field solvers. Similarly, microstrip line capacitance can be estimated with a multitude of different formulas, with the best one providing high accuracy results - in the range of 95+% [52].

With capacitance being calculated, G can be simply determined as follows: [21]

$$G = \tan\delta \cdot \omega C \tag{2.30}$$

where $\tan\delta$ is the loss tangent of dielectric material. There are alternative techniques to compute conductance, however, as G is negligible in majority of applications, it is reasonable to use the most simple estimation possible to avoid unnecessary effort in performing calculations.

The determination of R requires the most elaborate computations of all four distributed parameters. First, the DC resistance component per unit length can be computed as follows:

$$R_{DC} = \frac{\rho k_a}{A} \tag{2.31}$$

where ρ is the specific resistivity of used material (at room temperature), A is trace area and k_a is a correction factor considering return path resistance influence, which is in the bounds of [1;1.5] for microstrips. k_a is usually less than 2 as A_{return} is significantly larger than A of trace (e.g. the resistance of ground is comparably lower) [44]. Moreover current is non-uniformly distributed in the return path - concentrated in an area under signal path, with the main contributions located in a strip approximately 3x the width of signal trace. Thus, an accurate approximation of k_a is $\frac{4}{3}$ [6].

To compute AC components of R, current penetration depth w.r.t conductor surface (skin depth (δ)) has to be determined first. Skin depth can be estimated with a simple expressing given below [21]:

$$\delta = \sqrt{\frac{2\rho}{\omega\mu}} \tag{2.32}$$

where μ is material permeability (assumed to be μ_0 as copper $\mu_r \approx 1$). High frequency approximation of resistance, Equation 2.33 can be used to roughly determine the AC resistance past skin-effect onset [44].

$$R_0 = \frac{k_p k_r k_a}{p} \cdot \frac{\rho}{\delta} \tag{2.33}$$

where p is the cross-section perimeter, k_p is the proximity effect correction factor and k_r is the roughness factor. For simplicity the proximity factor between the signal and ground path is assumed to be 2 for both single and differential signalling as trace width is always at least twice greater than the dielectric height [44].

To determine whether the equation is applicable in the analysis, the skin effect onset frequency has to be determined as given below for a rectangular waveguide [44]:

$$\omega_\delta = \frac{2\rho}{\mu} \cdot \left(\frac{k_a p}{k_p A}\right)^2 \quad \rightarrow \quad \omega_\delta = \frac{8\rho k_a^2}{\mu k_p^2} \cdot \left(\frac{W+t}{Wt}\right)^2 \tag{2.34}$$

Note, the roughness coefficient is not taken into account for ω_δ computations as at transition frequency current is flowing throughout the conductor. In particular design case the operating frequency is always going to be larger than onset frequency of skin effect, thus Equation 2.33 holds.

To complete AC resistance calculations, roughness effect has to be determined. A simple k_r approximation is provided in Equation 2.35 [53].

$$k_r = 1 + \frac{2}{\pi} \arctan\left[1.4 \cdot \left(\frac{R_{avg}}{\delta}\right)^2\right] \tag{2.35}$$

R_{avg} is equivalent to average surface roughness of material (specified in μm) [7] caused by material etching and polishing [54].

The total frequency dependant resistance can be approximated to incorporate both DC and AC components as given below [44] providing the highest accuracy:

$$R_{tot} = (1+j) \cdot \sqrt{R_{DC}^2 + R_{AC}^2} \tag{2.36}$$

[6] URL https://www.protoexpress.com/blog/losses-in-pcb-transmission-lines/ [cited on 1st of December 2022]

[7] URL https://www.protoexpress.com/blog/losses-in-pcb-transmission-lines/ [cited on 7th of December 2022]

The total resistance is assumed to be composed from equivalent contributions of real and complex part, where the latter accounts for variations in internal inductance [43].

If the high frequency resistance is determined at a different point than the operating frequency, the transition to the required frequency can be performed as given below:

$$Re[R_{AC}] = R_0\sqrt{\frac{\omega}{\omega_0}} \tag{2.37}$$

, where ω_0 denotes the frequency at which R_0 is provided.

Note, the distributed element model is a useful tool for first order estimations of TL characteristics. Distributed element emulated TL is inferior to sophisticated EM models utilizing FEM, especially if more complex systems have to be implemented. For instance, modeling of cross-coupled TL using distributed models is a tedious process as each interaction has to be quantified precisely and independently. To obtain cumulative effect, superposition of these interactions have to be added to the model making it a time consuming and complex endeavour. Nevertheless, models using 2D solvers on FEM basis incorporate only slightly modified RLCG equations provided above to determine individual component contributions. Thereafter, for initial simulations distributed TL approximation is deemed an adequate tool.

With TL being defined and its important parameters analysed, packaging type suitable for high speed applications has to be investigated. Bear in mind, internal package traces resemble TL, thus, theory covered in this section is going to be useful also in Section 2.5

2.5. High-Speed Applicable Packaging

One of the limiting factors to achieve high speed for digital interconnects is the packaging. As package contribution to circuit is purely parasitic, one has to design/choose a package which does not detrimentally suffocate the signal. For instance, package pins and footprint bondpads cause a significant parasitic inductance to appear in the signal path, thus increasing time constant of the circuit and reducing slew rate of the system. Thereafter, conventional large-pin packages as Quad flat no-lead (QFN), Quad flat (QFP) or silicon on insulator (SOI) are not the best choices for the particular design case. [55]

To alleviate the pin limitations, flip-chip (FC) technology employing solder ball attachments to the board (Ball grid array - BGA) is proposed. As a low parasitics alternative direct chip-to-chip bonding which allows to almost completely bypass the external interconnect (PCB trace) could be used. Omission of transmission line would significantly enhance the achievable bandwidth, delay time and signal quality properties assuming iso-area conditions since effective load capacitance seen by transmitter would reduce. Nevertheless, the latter option is regarded as future work due to the current commercial products utilizing separate memory and PU, meaning that each device is likely supplied by different manufacturers. Investigating a bonded configuration is a sophisticated task as immense amount of additional information on memory and PU technology compatibility for bonding purposes has to be found. It was deemed impossible to perform as part of a large scale MSc project of 9 month duration.

Even though BGA package reduces the parasitics in the signal path, its manufacturing costs are significantly higher than for conventional packages mentioned earlier leading to an increased product cost. Thereafter, flip-chip package would be used for a fully operational commercial product, while for testing purposes, one would settle for a cheaper, but more parasitic option - chip-on-board (CoB) as an example. This implies that the commercial product's performance (throughput) is ought to be better than tested value/magnitude, if iso-area tests are performed comparing to the simulations. Assuming that design process would include both FC and CoB packaging types, analysis on former can be found in Subsection 2.5.1, while the latter is discussed in Subsection 2.5.2.

2.5.1. Flip-chip Package

A possible model for flip-chip package can be seen in Figure 2.22[8]. Note, the connection between the silicon die and copper interior of the package is realized with an additional set of smaller BGA. For lower speed applications a bondwire could be used for the same interface to cut on the manufacturing costs.

[8] URL `https://www.pcmag.com/encyclopedia/term/flip-chip` [cited on 14th of December 2022]

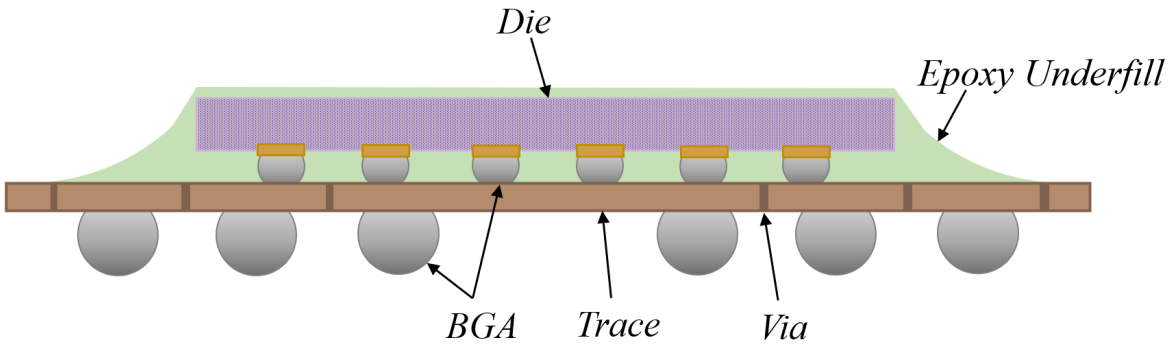


Figure 2.22: Flip-chip packaging model

A simplified model of the package can be seen in Figure 2.23, with 3 distinct identifiable components. The components are analysed in the sequence of how parasitic they are deemed to be, starting with the metallic trace and finishing with solder ball of the BGA.

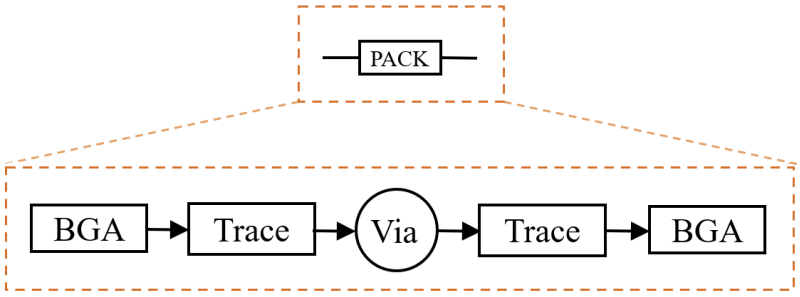


Figure 2.23: Simplified flip-chip package

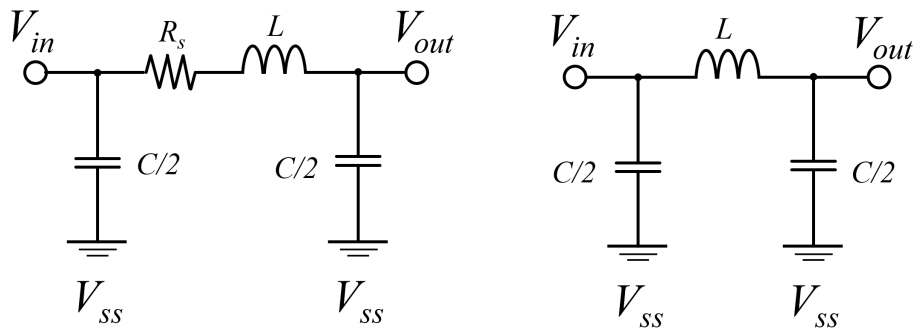
Metallic Trace

When inspecting Figure 2.22 one can notice that the metallic trace resembles a TL like structure as it is a copper lead set between two dielectrics: package insulator and underfill or the PCB core. Thereafter, the elaborate analysis provided in Section 2.4 is reused for package's trace instance. The only difference is the choice for set of formulas required - air is now always replaced with a second dielectric, meaning that conventional MS cannot be applied. For simplicity, it is assumed that both underfill and PCB core share the same dielectric properties. Notice, package traces have to be more compact in size compared to the external transmission links to ensure that the standard commercial encasing dimensions are met.

Via

For a packaging case depicted in Figure 2.22 a via is required to connect the copper levels adjacent to the two distinct ball grid arrays. The via in majority of the cases is short, thus, the distortion contribution of its parasitics to total circuitry is comparably low. Nevertheless, to encompass all the packaging effects, via is implemented in the design analysis as a lumped RLC model. With via being located in between two identical characteristic traces, it can be said to be symmetrical - have the same number of pads at each of its tips [56]. For via to exhibit symmetrical properties to the highest extent, a balanced π-bridge model is via's portrayal in simulation environment [57]. An example of simplified via model can be found in Figure 2.24a.

Via component values are determined sequentially, starting with inductance and finishing with resistance. Via inductance is determined as an average of two empirical relations given in Equation 2.38 [58] and Equation 2.39 [59]. The mean of the two relations is taken to ensure that the value is neither underestimated [59] nor overestimated [58] as such trends are moderately followed when each expression is used separately.



(a) π-bridge model of the via **(b)** π-bridge model of the BGA

Figure 2.24: Lumped element model of a π-bridge for a) via b) BGA

$$L_{via_1} = \frac{\mu_0}{2\pi} h_{via} \left\{ \ln \left[\frac{h_{via}}{r_{via}} + \sqrt{\left(1 + \left(\frac{h_{via}}{r_{via}}\right)^2\right)} \right] - \sqrt{\left(1 + \left(\frac{r_{via}}{h_{via}}\right)^2\right)} + \frac{r_{via}}{h_{via}} \right\} \tag{2.38}$$

$$L_{via_2} = \frac{\mu_0}{2\pi} \left[h_{via} \cdot \ln \left(\frac{h_{via} + \sqrt{r_{via}^2 + h_{via}^2}}{r_{via}} \right) + \frac{3}{2} \left(r_{via} - \sqrt{r_{via}^2 + h_{via}^2} \right) \right] \tag{2.39}$$

In equations above, h_{via} is defined as via height and r_{via} as via radius. Note, the impact of mutual inductance caused by ground vias generating a return path is neglected for simplicity. Thus, the grounds are assumed to be separated from signal via by a considerable distance, leading to overestimation of total loop inductance.

Via capacitance can be determined using a simplified closed loop approximation as given in Equation 2.40 [56] where C_{vb} is via body capacitance, while C_{vp} denotes pad capacitance. N describes the amount of connection pads to the via. For signal transmission in the package, the number of pads is set to be 2, assuming that microstrip-to-microstrip communication takes place.

$$C_{via} = C_{vb} + NC_{vp} \tag{2.40}$$

Via body capacitance between two metallic plates can be estimated by splitting the height of the via into small equidistant instances and summing the respective capacitive contributions together as given in Equation 2.41 [56].

$$C_{vb} = \frac{2\pi\varepsilon_{eff}}{\ln\left(\frac{r_{ap}}{r_{via}}\right)} \times \sum_{k=1}^{m} \frac{(h_{via} + t)(r_{ap} - r_{via})}{m\sqrt{\left[\frac{h_{via}}{2} - (k-1)\left(\frac{h_{via}}{2m}\right)\right]^2 + (r_{ap} - r_{via})^2}} \tag{2.41}$$

In Equation 2.41 r_{ap} represents the anti-pad radius, t the metallic layer thickness and m is the number of capacitive increments (Amount of layers between via pad-to-pad). Assuming a 3 layer stack-up where the middle layer is a shared ground between the signal planes, the full body via capacitance can be acquired by multiplying Equation 2.41 by 2.

To acquire via pad capacitance more sophisticated equations have to be used, for which only the final expression is presented here. For full derivation, refer to the source paper [56]. C_{vp} can be determined as shown in the following expression:

$$C_{vp} = \frac{2\varepsilon_r\varepsilon_0}{r_p} \left[\frac{r_{ap} - r_p}{\ln\left(\frac{r_{ap}}{r_p}\right)} \right]^2 \left\{ \frac{\pi^2}{0.985 r_p} \frac{(r_{ap} - r_{via})\,\Lambda}{\left(\frac{r_{ap}}{r_p} - 1\right)^2} - 4\ln[\sin(\psi)] - \left[\frac{1}{\psi} \sin(2\psi) \right]^2 \right\} \tag{2.42}$$

, where ψ is given as:

$$\psi = \frac{\pi}{2} \frac{r_{ap} - r_p}{r_{ap} - r_{via}} \tag{2.43}$$

and Λ is given as:

$$\Lambda = \frac{\left[J_0\left(\chi_{01}\right) N_0\left(\chi_{01}\frac{r_p}{r_{via}}\right) - N_0\left(\chi_{01}\right) J_0\left(\chi_{01}\frac{r_p}{r_{via}}\right)\right]^2}{\left[J_0^2\left(\chi_{01}\right)/J_0^2\left(\chi_{01}\frac{r_{ap}}{r_{via}}\right)\right] - 1} \tag{2.44}$$

In Equation 2.44 Bessel's function of first kind of order zero (J_0) and Bessel's function of second kind of order zero (N_0, also called Neumann's function [9]) are computed using the first root χ_{01} defined as given in Equation 2.45 for $\frac{r_{ap}}{r_{via}} \in$ [2.5, 3.5].

$$\chi_{01} \approx \frac{3.1}{\frac{r_{ap}}{r_via} - 1} \tag{2.45}$$

Combining Equation 2.42 and Equation 2.41 as shown in Equation 2.40 leads to an estimated value of the via capacitance.

Lastly, the resistance of via is computed using both AC and DC components. DC resistance is determined assuming a hollow cylinder with the equation as given in Equation 2.46.

$$R_{DC} = \frac{\rho h_{via}}{\pi(r_p^2 - r_{via}^2)} \tag{2.46}$$

AC resistance can be determined as shown in Equation 2.47, where δ is the effective skin thickness [44]. Note, to simplify the calculations, AC contribution is determined assuming a solid conductor without a drill hole at its centre.

$$R_{AC} = \frac{k_a k_r \rho h_{via}}{\pi r_{via} \delta} \tag{2.47}$$

The total resistance can then be expressed as the square average of both contributions as given further:

$$R_{tot} = \sqrt{R_{DC}^2 + R_{AC}^2} \tag{2.48}$$

It can be seen that for determination of Via's resistance the same base formulas as given for TL are used, with slight alterations.

Ball Grid Array (BGA)

As last of the packaging elements, both the internal and external BGA are analysed. For simplicity, it is assumed that the pitch to ball radius ratio scales proportionally from inner to outer BGA, such that a unified model with equal values can be used for both parasitic contributions. Additionally, as interconnect is of the same width (if not greater) as the distance it spans, the resistance of the model is neglected, as can be seen in Figure 2.24b [60]. Similarly, as done for via, a π-bridge configuration is used as a model of BGA structure.

Knowing that the solder ball parasitic contributions to the system are minuscule compared to metallic traces and vias, the values for its inductance and capacitance are determined from literature rather than with the help of analytical expressions. An a rule of thumb, an inductance value of 0.01 nH and capacitance value of 0.1 pF are chose to be representative for the BGAs [61]. Note, the effect of determining and using quasi-exact values for BGA would have close to no effect on the design response thus, it was decided to shift the time focus on more detrimental effect (e.g crosstalk, TL impedance missmatch, etc.) modelling.

2.5.2. Chip-on-board Package Model

Chip-on-board is one of the most simple and cheap packages as it includes only a handful of components. Most frequently the package consists of a bondwire between two pads as depicted in Figure 2.25. The manufacturing process of CoB is rather straightforward - bare chip die is directly mounted onto a PCB using a conductive adhesive [10] and a bondwire is used to connect chip pad to PCB pad. The

[9] URL `https://encyclopediaofmath.org/wiki/Neumann_function` [cited on 17th of March 2023]

whole structure is engulfed in protective structure as epoxy moulding compound to protect the chip and bondwires from external damage.

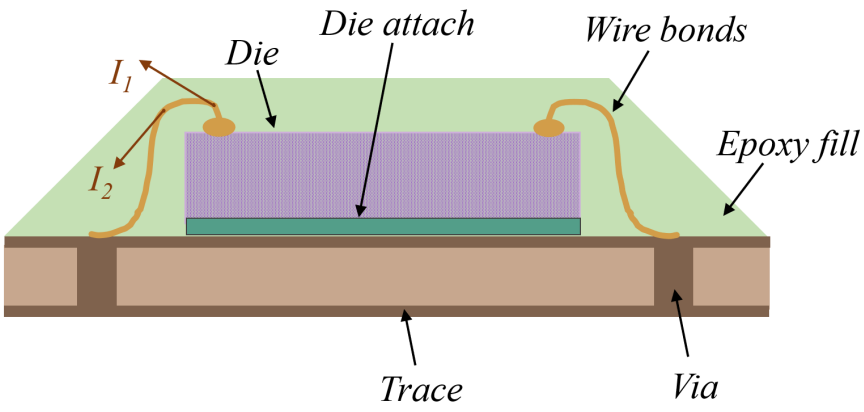


Figure 2.25: Chip-on-board package cross-section view

Simplifying the package similarly as it was done for FC, a symmetric model consisting of pad and bondwire can be seen in Figure 2.26. For analysis round and rectangular conductor conductor analysis can be utilized. Note, Figure 2.26 excludes chip pad - it can be noticed that signal through this component flows laterally to the largest cross-sectional area, thus its influence on signal degradation can be effectively neglected.

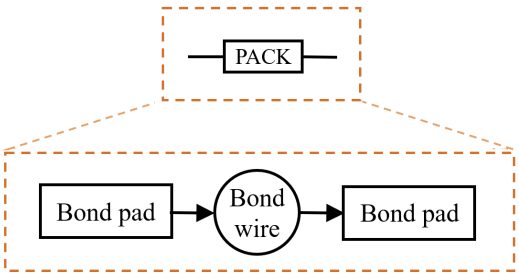


Figure 2.26: Simplified chip-on-board package

PCB pad can be assumed to act as a very short transmission line, thus TL analysis can be applied to directly compute its properties. Thus, no supplementary formulas and determination strategies for pad RLC components are provided in this section.

The wire, on the other hand, requires a change of analysis because first it is located well above PCB return path and second - the cross-section of the bondwire is a circle rather than a rectangle. Assuming that signal bondwire is separated from its return path by a finite but large distance, the capacitance it generates can be neglected. Thus, a bondwire can be approximated simply by series RL lumped element model. The resistance can once again be computed as it was done in Section 2.4 by changing A calculations to incorporate formulas for a circle.

Self inductance for a round conductor can be determined as shown in Equation 2.49 [62], where l_{bw} indicates bondwire length and r_{bw} the radius. Bear in mind, Equation 2.49 is derived for a straight wire, hence actual bondwire inductance would deviate from L_{wire} proportionally to curvature. The actual L_{self} could be marginally lower as currents at different points are in angle relative to one another - vertical axis projections would have opposite polarity (see I_1 and I_2 in Figure 2.25). As slight overestimation of degradation components is beneficial, no alterations to Equation 2.49 are made.

[10]URL `https://www.caplinq.com/semiconductors/advanced-packaging/chip-on-board-cob/` [cited on 17th of July 2023]

$$L_{self} = 2\left[l_{bw}\ln\left(\frac{l_{bw}+\sqrt{l_{bw}^2+r_{bw}^2}}{r_{bw}}\right) - \sqrt{l_{bw}^2+r_{bw}^2} + \frac{l_{bw}}{4} + r_{bw}\right] \tag{2.49}$$

To determine total bondwire L, mutual inductance between adjacent wires has to be accounted for. Assuming that two bondwires are parallel with respect to one another, mutual inductance can be determined as follows [63]:

$$L_{mut} = 2\left[l_{bw}\ln\left(\frac{l_{bw}+\sqrt{l_{bw}^2+d^2}}{d}\right) - \sqrt{l_{bw}^2+d^2} + d\right] \tag{2.50}$$

, where d is the separation between the wires. Mutual inductance contribution of adjacent wires has to be determined with each trace separately as the distances are varying. Assuming only one signal and one return path, total inductance of signal loop can be expressed as $2(L_{self} - L_{mut})$. Note, total inductance of each wire separately is approximated simply as difference between self and mutual inductance. In industry, rule of thumb for bondwire inductance per 1 mm is 1 nH [61].

With packaging types established and possible models analysed, termination as last subsystem of I/O can be discussed. TERM has to be sized when PACK and TL models are defined to ensure that high impedance matching is attained, such that reflections are minimized as much as possible.

2.6. Termination

As briefly mentioned in Section 2.2, termination's main task is to define a voltage variation at the output, which is captured and further interpreted by RX amplifier. In certain topologies as CTT (refer to Subsection 2.2.1), termination can also be responsible of defining V_{cm} at the RX input. Moreover, having a matched termination - impedance of TL is set equal to TERM value - leads to minimization of signal reflections caused by propagating wave interactions with discontinuities. Reflections are partially responsible for inter-symbol interference (ISI) - interaction between a preceding and subsequent wave leading to signal partial overlap and hence degraded signal voltage response in time. ISI is mainly caused by reflections, incomplete settling of circuitry components (under/over-charged, noise deviations) and crosstalk [64]. Thereafter, TERM value not only has to be set for ideal V_{cm} and swing properties, but also to reduce reflections caused ISI.

With former considerations in mind, TERM type, implementation and location in the system can be discussed. Note, most of the topologies discussed in Section 2.2 have pre-defined load-side (far-end) TERM with the freedom to choose weather to use a source (near-end) TERM. Using a source termination allows to absorb wave of incoming back-propagating reflections. Otherwise, TX would reflect the signal back once again, however, this time it will be propagating in the forward direction causing direct degradation of incident signal. However, adding source resistance causes an increase in power consumption due higher voltage drop across the line and another point for discontinuities to arise with large variations of PVT. Thereafter it can be omitted in almost perfectly matched, low parasitic systems which do not have severe initial reflections. Bear in mind, not having source TERM limits the design to point-to-point interface use only as signal at the near-end of the spectrum can be unusable. [65]

The termination can be placed at two locations in the system: right before receiving IC package and on the chip - referred to as on-die termination (ODT). Terminating TL externally will cause a significant amount of noise being introduced into the signal due to path discontinuity caused by packaging. PACK effectively modifies TERM to be a complex impedance due to its inherent RLC contributions (refer to Section 2.5). An example of outside termination can be seen in Figure 2.27, where signal eye diagram at the RX input is shown. Eye diagram is a graphical interpretation of entire transient signal response, which is achieved by cutting required waveform in intervals equal to data period and overlapping them all. More elaboration on eye diagram properties are provided in Subsection 3.2.1.

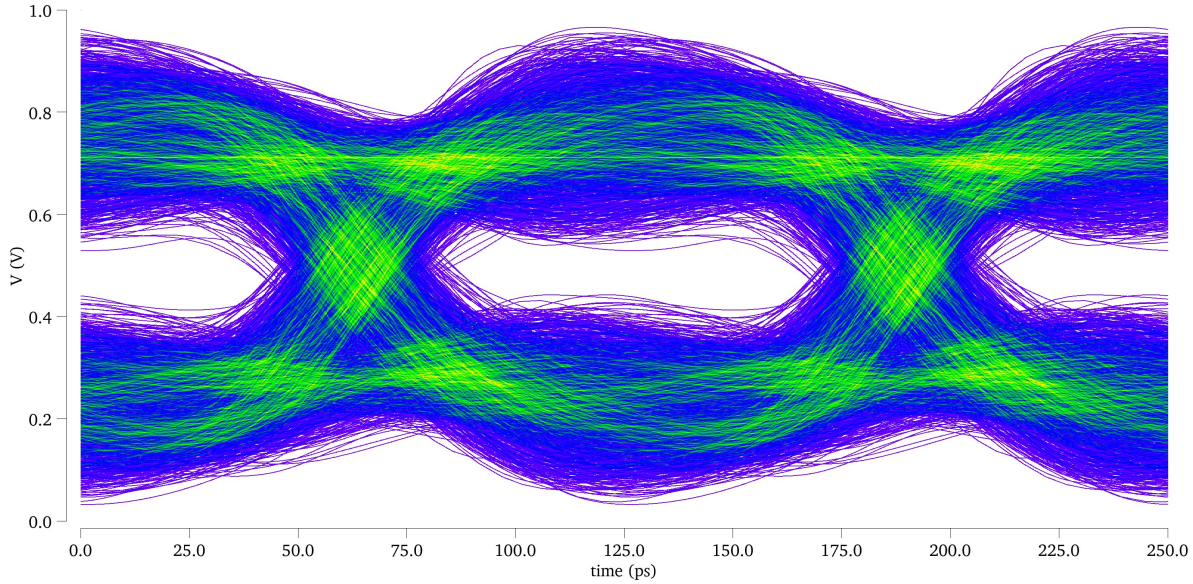


Figure 2.27: Example of transient signal eye when termination is located outside of the chip. Simulation conditions (refer to Section 4.1): CTT topology, Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

When TERM is placed on the chip as close to the RX input as possible, eye diagram shown in Figure 2.28 is obtained. Note, signal experiences both amplitude and period fluctuations due to change in TERM location only. When comparing the figures, one can conclude that having ODT is significantly more beneficial and should be done if possible.

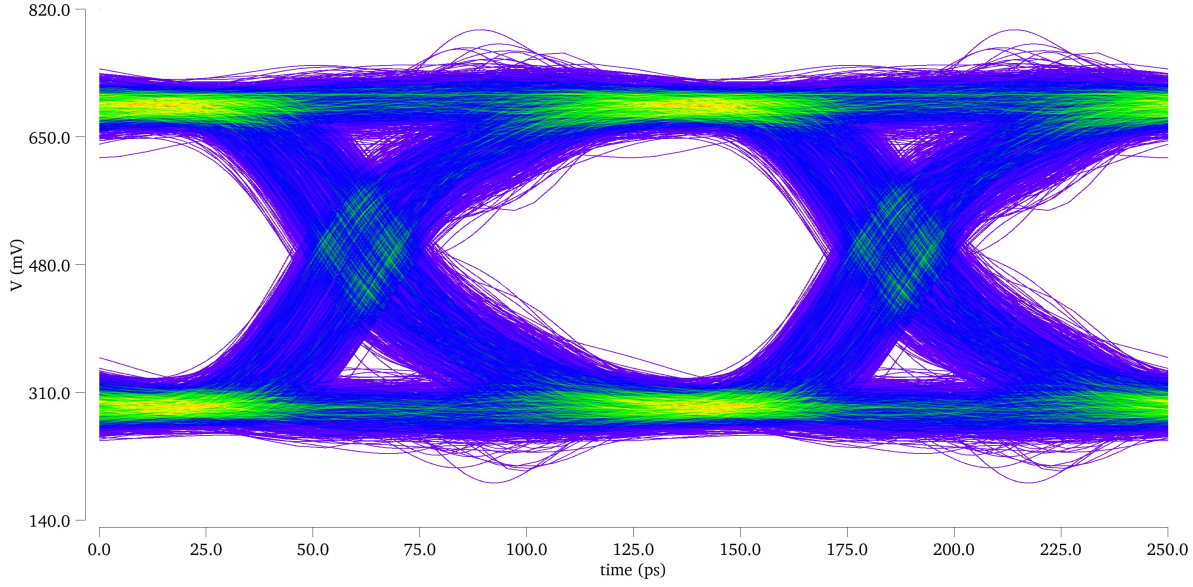


Figure 2.28: Example of transient signal eye when termination is located on-chip. Simulation conditions (refer to Section 4.1): CTT topology, Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

2.6.1. Passive Termination

Passive termination utilizes simple resistors and in cases where AC coupling is needed also capacitances. Passive components have the great advantage of being linear with temperature variations, which is beneficial for system's stability. This is also true to a certain degree when resistors are implemented on the metallic layers of the chip. There, the linearity is achievable in the temperature range of

interest, which is commonly set to automotive standards [-40, 150] °C. For larger deviations the resistance starts to show slight non-linearity, however, the slope changes can be assumed to be negligible.

The main drawbacks of using passive devices are large area in comparison to active devices and more complicated adaptability or resistance control. Large area comes from the fact that metal layer thickness and usually also the width is fixed, leaving only 1 degree of freedom to influence resistance - the length. As width is commonly predefined, unit resistance per square can be determined allowing for easy determination of resistance per unit length. To obtain the exact resistance required, it might be necessary to put several devices in parallel, if length is limited to stride-wise changes only.

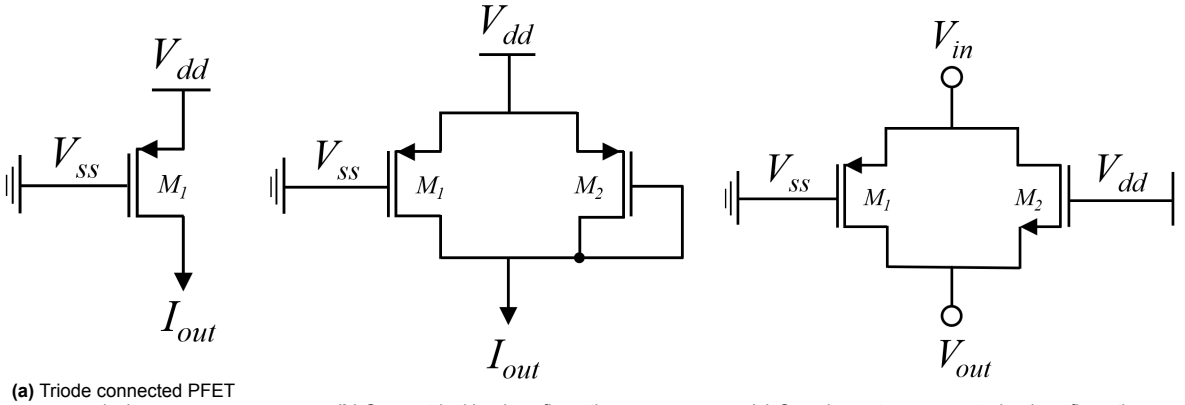
An additional minor disadvantage coming directly from use of large area - in case temperature distribution in chip is non-uniform, resistor could experience two different temperatures at its edges, causing the R values to shift non-linearly w.r.t temperature. Nevertheless, the latter effect can be neglected in the first order analysis.

In high crosstalk conditions TL characteristic impedance might change continuously [64], which would cause significant reflections on the line if termination resistance is kept constant. To achieve TERM controllability or trimming [24], parallel branches of idle resistances would have to be implemented and activated on command. This would lead to a substantial increase in area not only because of the area required by resistance itself, but also active devices responsible for controlling the on/off switching of resistances and control unit evaluating required and current state condition in-compatibility.

Passive TERM downsides covered above can be rectified by using active devices in load setup to generate required resistance. Nevertheless, for small adjustments to obtain exact resistance required, passive elements could be used as complementary structure to enhance the performance.

2.6.2. Active Termination

To use an active device as a TERM resistance, it is desirable to bias it in triode region (see Figure 2.29a), where current is almost linear w.r.t V_{ds} voltage, hence resembling passive resistance characteristics. The most effective way to bias a transistor in linear region is to have maximum achievable voltage supplied at the gate (or minimum for PFET). Additionally, using minimum channel length enhances channel-length modulation effect, leading to steeper slope of I_D vs. V_{ds}. Pushing L value close to the minimum leads to triode curve being close to equivalent to saturation one, providing ideal resistance characteristics. As a benefit, using lowest possible length value results in smallest active device area. [21]



(a) Triode connected PFET device

(b) Symmetrical load configuration

(c) Complementary pass gate load configuration

Figure 2.29: Active load device in a) triode configuration b) symmetric configuration c) complementary pass gate configuration

To extend linearity region beyond triode, a parallel transistor connected in diode configuration can be used as shown in Figure 2.29b. It is observed, however, that the provided characteristics are not fully linear, but rather slightly deviating from trend-line with symmetry at the middle point as shown in Figure 2.30. The latter happens because at the middle point the triode and diode characteristics are not fully balanced - either of the devices provides a slightly higher current than the other. To acquire the currents to be alike in the middle point ($V_{gs} = \frac{V_{dd}}{2}$), V_{th} should be as low as possible. In such a case, triode and saturation regime currents become equivalent at $V_s = \frac{V_{dd}}{2}$, implying that full linearity is going to be acquired.

If design allows for tolerances, one should opt for having close to linear resistance in full rail-to-rail swing to minimize average reflection impact. Additionally, symmetric load provides a good PSRR, thus limiting amplitude noise being converted into jitter. [66]

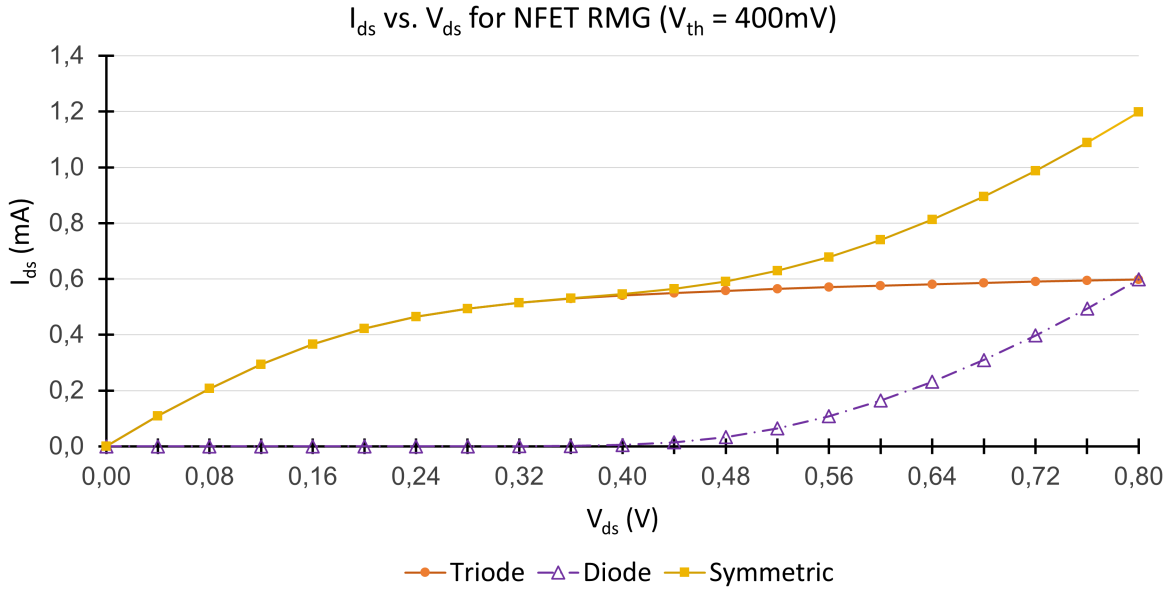


Figure 2.30: Ids vs. V_{ds} characteristic for symmetric load configuration. Simulations performed with Imec FinFET 14 nm tech. RMG and V_{th} value of 400 mV

To implement a resistance supplied by any voltages for which V_{cm} is preferably defined at $\frac{V_{dd}}{2}$, complementary pass gate structure can be used as shown in Figure 2.29c. Pass gate configuration operates mainly by balancing resistance value with 'on' strength of complementary transistors at different voltages. For instance, assuming that V_{cm} is set to $\frac{V_{dd}}{2}$, both NFET and PFET will contribute almost equally to output resistance of the structure (if equal conduction capabilities are set) for any differential swing inside power rails. Shifting common mode voltage closer to either of the rails will lead to one of the devices being more superior and thus slightly skewing equivalent output resistance. Bear in mind, with symmetric design, common mode variation from $\frac{V_{dd}}{2}$ in either rail direction should amount to the same difference in equivalent resistance as identical change towards opposite rail. It can be concluded that pass gate configuration's resistance values would have a bell curve dependence on V_{cm}. With respect to differential voltage, on the other hand, changes in resistance can be assumed negligible, especially if low swing fluctuations are present in the system. [21]

PVT variations can cause transistor's nominal parameter deviation by up to two times in both best and worst case scenarios. Large differences in TERM value over time could lead to low reliability of system output due to degradation mechanisms as crosstalk induced TL changes and thus severe reflections damaging signal beyond repair. To achieve best TERM stability with respect to PVT, termination has to be made controllable. To digitally trim TERM value, several load structures with varying output resistances would have to be put in parallel and turned on/off depending on required conditions. To simplify scaling, a base value of resistance can be predefined using a static load structure - this restricts achievable TERM range, but provides more precision to equivalent resistance. As parallelizing resistances leads to lower overall R, static TERM module has to be set to highest predicted resistance under best performance conditions (lowest R per unit area). For lowest predicted R similar approach is taken - all branches are turned on and target has to be reached under worst performance conditions (highest R per unit area). [21]

Digital TERM trimming main downside is large area it requires. The control unit would consist of comparator comparing instantaneous TERM value with a stable reference, decision performing logic and logic gates to enable required modules. With the amount of extra components, it can be seen that supporting circuitry could exceed load modules in size. Moreover, if all the inputs would be switched simultaneously, high amounts of switching noise would appear on the output. To prevent noise spikes,

either switching slew rate has to be reduced or turn on signals would have to have a finite delay w.r.t one another. It is also possible to trim TERM using analogue control. Nevertheless, it is not the preferred method of control because deviating V_{gs} from rail voltages leads to higher device base resistance (less steep I_D vs. V_{ds} curve) and thus lower available V_{ds} swing. Furthermore, if control voltage is shared between several load devices, voltage level losses over interconnect could cause significant deviations in transistor biasing conditions and thus the effective output resistance. [65]

With TERM now defined, the I/O system has reached completion. Initial simulations for various designs could be made, however, to validate the claims, sensitivity analysis has to be performed. Prior to sensitivity analysis, predictions on the results have to be made, to ensure that the results match expectations related to theory. Hence, scope of sensitivity analysis is provided in the next section.

2.7. Scope of Sensitivity Analysis

Design sensitivity analysis (SA) is an important, iterative intermediate design step, which helps to determine whether a system is able to provide sufficient performance upon design parameter/condition change. For instance, insight on how exactly system responds to power supply variation allows to implement necessary counter measures early-on, guaranteeing that circuit will operate as intended. One can perform SA as soon as the system has been established and initial simulations made, which confirm that most of the requirements are met. The main benefit of performing a small scale SA not long into the project allows to identify whether small alterations of design parameters lead to extreme performance changes. If that is the case, already made decisions have to be back-tracked and verified - potentially a completely different approach/design has to be selected to attain higher design stability. Otherwise, sooner than later a real life system will fail due to limitations posed by simulation being only a mere approximation of reality, where all the effects cannot be predicted and accounted for.

Sensitivity analysis can be performed on multitude of design parameters and in countless different ways - one could look only at performance variation due to design parameter change, or immediately look at how much other parameters have to be adapted to compensate for performance degradation. For example, power supply adaptation leads to reduction in effective current transistors provide, and thus degraded performance which can in most cases be measured as output voltage drop. Alternatively, it is possible to determine how much transistor area would need to be increased to achieve close to iso-performance conditions at the output node. Note, not always straightforward compensation can be used, as performance metrics are usually not a single parameter, but rather several parameters inter-dependant of one another - improving one parameter might cause losses on the other.

SA is an iterative process because commonly countermeasures taken to rectify a worst case performance scenario are taken as the new nominal condition design. From here other parameters which were less influential have to be checked again with the new sizing/design and the new worst case state becomes new standard case. The process has to be repeated until acceptable performance variation tolerances are obtained.

In further sections parameters for which design SA was performed in this thesis can be found, starting with DR variation. Observe, accompanying explanation on predictions of variation impact and guesstimated first order trends are added for each parameter considered. Note, however, clarifications on performance variation mechanisms are merely hypothesis based on the theoretical and practical knowledge of the author.

Lastly, whenever performance is mentioned, it refers to throughput of data with imposed minimum swing and slew rate requirements. The requirements used in this thesis are elaborated on in Subsection 3.2.1

2.7.1. Variation of Data Rate and its Impact

Reaching a particular throughput is the first objective of the thesis, making design response to DR one of the most interesting parameters for investigation. Nevertheless, change of operating frequency is one of the most complex parameters to be explored, as most of the circuit operations are frequency dependent. For example, gain value of an amplifier starts to substantially reduce after reaching corner (dominant pole) frequency, EMI caused degradation rapidly grows as it is proportional to slew rate steepness ($\Delta V = L\frac{dI}{dt}$), power dissipation increases in complementary FET circuits due to more frequent charging/discharging of intrinsic device capacitances, etc. Thereafter, net effect of output signal quality cannot be accurately determined, but it can be indicated that to acquire iso-performance conditions a

stronger driver circuitry and thus larger overall circuit area would be required. Non-linear performance degradation for SES with increase in frequency has been confirmed by SA found in literature [23]. Notice, analysis done in Yong et. al. uses a stripline TL type, which has the highest crosstalk rejection compared to other TL types due to its high coupling to return paths [47].

Nevertheless, results provided by Yong et. al. have to be used with caution - each design has its own unique response w.r.t frequency. For instance, SES designs are affected more severely by EMI and crosstalk than any DS topology. Good coupling between two complementary signals used by DS reduces the effective crosstalk as both lines experience close to similar amounts of interference. Thus, it can be said with high confidence that SES requires significantly higher area compensation than DS to achieve sufficient signal quality when going to higher operational frequencies. Additionally, maximum frequency at which SES would operate properly for infinitely large driver size is going to be lower than that of DS - at high frequencies crosstalk would start to dominate the signal, effectively reducing output swing to zero [67]. Lastly, it is predicted, that DS area will grow approximately linear for short range of DR variation, if good coupling between the signal lines is going to be established, as strong EMI rejection will be obtained.

2.7.2. Variation of Transmission Line Length and its Impact

Another important parameter to be explored is TL length. Determining TL length influence provides designers with knowledge of how much freedom regarding component placement and routing the system permits. Yong et. al. has also investigated SES line dependence on TL length, which is used as a reference [23]. Note, however, system used in Yong et. al. is non-terminated for majority of investigations with the exception when reflection contribution was disabled. Seeing that topologies covered in Section 2.2 are all load terminated, it can be assumed that severe under- and overshoot present in observations for short TL would not be present in an actual design [68]. Such severe amplitude changes at periodically repeating time instance resemble standing waves which can be generated inside a PCB TL in case a termination is lacking or it is non-matched [11].

Assuming that no standing waves would be present in the design (close to no reflections), the most signal degrading mechanism would become crosstalk. In such a case, the longer the TL, the more coupling is present between two lines running in parallel. Thereafter, it is predicted, that increased TL length leads to first-degree linear reduction in performance or equivalently linear increase in area. The estimate is performed by investigating the amount of additional bits present on a longer TL. By splitting a TL in discrete, same size intervals, one can see that each piece will simultaneously contain approximately equivalent amount of bits compared to other sections. As probability of bit being a '1' or '0' is approximately 50%, sections can also be said to accommodate the same amount of signal transitions. Thereafter, crosstalk can be said to increase linearly to first degree.

Alternatively, increasing TL length by a finite fraction leads to increase in effective line inductance and also the mutual line inductance ($\Delta V = L_{mut} * \frac{dI}{dt}$), which is directly proportional to voltage swing variation at signal transitions (switching noise) [69]. Here, the voltage change induced current variations due to mutual capacitance ($\Delta I = C_{mut} \frac{dV}{dt}$) are neglected. $\frac{L_{mut}}{L_{self}}$ is commonly higher than $\frac{C_{mut}}{C_{self}}$ [64] - the latter in combination with almost instantaneous current change w.r.t time comparing to gradual voltage fluctuations in time makes mutual inductance the governing crosstalk mechanism. The mutual capacitance accounts for some trend non-linearity which might be observed in both performance vs. TL length/DR variation as it degrades/enhances signal slew rate depending on whether even/odd coupling occurs. When slew rate becomes the critical compatibility condition, L_{mut} and C_{mut} effects can be said to be comparable.

Note, it has to be assumed that slew rate does not deviate from signal to signal and thus can be assumed a constant. Bear in mind, linearity would hold only for limited range of TL length similarly as it is shown by Yong et. al. [23]. For differential circuits where strong coupling between signal and its complementary are present, it is expected that this range will be significantly larger than for SES.

2.7.3. Variation of Guard Line Spacing and its Impact

Second TL parameter of high importance is the separation between two adjacent signal lines. Increasing the gap between two nearby signal lines leads to EMI coupling reduction due to lowered mutual

[11] URL `https://resources.pcb.cadence.com/blog/2022-understanding-standing-wave-patterns-on-interconnects-and-antennas` [cited on 19th of July 2022]

inductance and capacitance. As guard lines (grounded lines) are placed between any two signal traces, only closest neighbouring signal paths have to be taken into account during this analysis [70].

Similarly as for TL length variation, guard shifting effect can be imagined using $\Delta V = L_{mut} * \frac{dI}{dt}$. Spreading out traces on the PCB causes crosstalk induced voltage variations to decrease as mutual inductance between the lines drops. The change in inductance is expected to be linear for a relatively large range of guard spacing values. However, at some point spacing is so large that almost ideal coupling between signal line and return path is obtained - at this point L_{mut} is approximately zero. Thereafter, the reduction of area and power saturates converging to a finite value. Here, C_{mut} has similar considerations as for TL length variations, thereafter it is neglected to first degree. Note, relaxation of area is predicted to be lower for DS design than SES as former is inherently more immune to crosstalk than the later, implying that DS L_{mut} value reaches zero earlier.

Note, in case guard spacing gets too narrow, mutual parasitic components could become similar to self-generated ones, leading to complete annihilation of data signal. In such a case, quadratic, if not exponential increase in active area would be observed, rather than linear trend discussed above.

2.7.4. Variation of Power Supply Voltage and its Impact

Voltage on the power rails is a design parameter which is going to vary in every system regardless of the countermeasures taken. Severe power supply noise can lead to extreme over-voltage which causes faster gate oxide wear-out and hence lowered reliability of the circuitry. Drastic under-voltage on the other hand causes more data errors due to delayed charging of devices. The latter is amplified for active technology suffering from strong channel length modulation, implying high I_D dependency on V_{ds} ($\propto V_{dd}$). FinFET and long channel devices are more tolerant against power supply voltage fluctuations. [71]

In digital systems amplitude noise on the power supply can be represented as variation in timing conventions - delayed charging of intrinsic capacitors leads to data skew (noise) in time known as jitter. If jitter becomes too high causing huge miss-match between sampling clock and the signal, data identified at the output could be wrong. Thereafter, power supply noise can directly limit the maximum operating frequency of a circuit [72]. The severity of power supply noise is dependant on how well different current loops interact with each other. To reduce coupling between high impedance lines decoupling capaictors can be added to interrupt voltage variations across mutual inductance components [73]. Alternatively, it can be said that the capacitors are placed to establish charge balance in the system - to provide missing or absorb the excessive charge [74]. Bear in mind, a capacitor in real life system has an effective series resistance and inductance, implying that it behaves similarly to a band-reject filter [75]. Thereafter, high frequency components corresponding to noise do not get the necessary attenuation. Combining several capacitors in parallel can alleviate filter incompatibility issue, providing reduction in noise for required frequency range [76]. Note, decoupling capacitors effectively increase C_{mut} contribution, making it the dominant effect and thus suffocating any undesired voltage variations at DC nodes.

Voltage variation in the system depends on how power delivery is performed. If more than one power delivery modules is present, system could experience different voltage noise at distinct locations, causing signal discrepancies and signal skewing due to unbalanced voltage at transistor terminals. For instance, if input voltage of an inverter would vary in a range of [0, 0.8] V, while its rails would see continuous 1 V DC voltage and perfect ground, the PFET drive strength would be superior over NFET, causing stronger '1' than '0'. In case the voltage scenario of V_{in} and V_{dd} would be reversed ($V_{in} \in$ [0, 1] V, V_{dd} = 0.8 V), drive strength would be skewed in the opposite direction - stronger '0' than '1'. Using a single power supply could lead to similar issues if long resistive traces are used for interconnection of power supply to the system as each parallel segment would experience a voltage drop proportional to trace R. From this analysis 3 different variation cases can be identified. First, only the circuits input voltage undergoes variations while power supply remains constant. Second, only the power supply voltage varies while input remains constant. Lastly, both V_{dd} and V_{in} vary at the same time, which can be assimilated as technology voltage scaling.

In first and second case output voltage skewing is expected as discussed before. In case more than one stage is used, subsequent stages are going to partially correct strength miss-match. This would happen because gate voltage range of further stages would coincide with power rail voltages. The main difference between aforementioned case one and case two lies in circuits power consumption - changing V_{in} does not affect power almost at all (if pre-circuitry is neglected), while V_{dd} variations are

linearly proportional to power dissipation ($P = IV_{dd}$). The latter is true assuming that output current is the same for all designs in a set DR system for single topology - in case of area compensation to reduced V_{dd} power would reduce linearly. The assumption holds in case load capacitance is significantly larger than intrinsic device capacitance. The third case should lead to lowest distortions as no signal strength skewing is present in the system, however, power consumption should match the one observed in second case.

The performance degradation for each case has to be investigated independently. When V_{in} variations are small (0-100 mV), not much change should be observed in multistage systems as last stage sees almost nominal case conditions, apart from unbalanced drive strength. For larger V_{in} swings the V_{cm} of the first stage is substantially altered, thus causing strong biasing towards '1' with slew rate lag of falling edge. Inferior slew rate gets propagated throughout stages resulting in asymmetric output. Skewing of first stage can be visualized using saturation current equation shown in Equation 2.4. Reducing V_{gs} leads to quadratic reduction in drain current which is proportional to slew rate to the first degree as shown in Equation 2.51 [27]. In higher order terms the output capacitance cannot be assumed to be constant as intrinsic capacitances vary with transistor's biasing conditions. Thereafter, the overall performance degradation trend for first case can be assumed to be a mix between linear and quadratic relations. Note, reduced slew rate leads to shorter or no hold time - stable, settled output voltage level. Also, the current of subsequent stages would not change, thus, power consumption would not vary.

$$Slew\ rate = \frac{\Delta V}{\Delta t} = \frac{Q}{C_{out}\Delta t} = \frac{I_d \Delta t}{C_{out}\Delta t} = \frac{I_d}{C_{out}} \tag{2.51}$$

It is estimated that changing only V_{dd} as in the second case would lead to a fully quadratic drop in system performance with respect to the changes. Note, the variation could be slightly altered for high power supply voltage variation due to skewing of first input towards '0' with lagging rising edge. Alternatively, area has to be increased quadratically to ensure iso-performance conditions. Exactly the same considerations apply to the third case, however, here the trend should not be altered even at large voltage variation case, except if cut-off or moderate inversion (just in between 'on' and 'off') boundary has been reached.

2.7.5. Variation of Temperature and its Impact

Performance changes upon temperature variation are usually hard to predict exactly as temperature influences multiple parameters simultaneously. For instance, carrier mobility is inversely proportional to a temperature exponent [77] and threshold voltage is generally reducing with temperature [78]. Even more, the exponent in carrier mobility is usually not constant with process and temperature variations[78]. Hence, exact net effect of temperature increase on I_d cannot be directly determined theoretically as reducing μ tends to reduce it, while V_{th} reduction tends to increase it according to Equation 2.4. When both effects are balanced, a point where I_d is temperature invariant emerges - zero temperature coefficient (ZTC) point has been determined [79]. Above this point current vs. temperature is usually decreasing and dominated by μ reduction effect, while below this point V_{th} reduction is the dominating mechanism as shown in Figure 2.31. Thereafter, measuring temperature dependence on performance is highly dependent on gate voltage biasing point and no speculations on expected trend are made at the moment, but elaborated when full SA setup is defined.

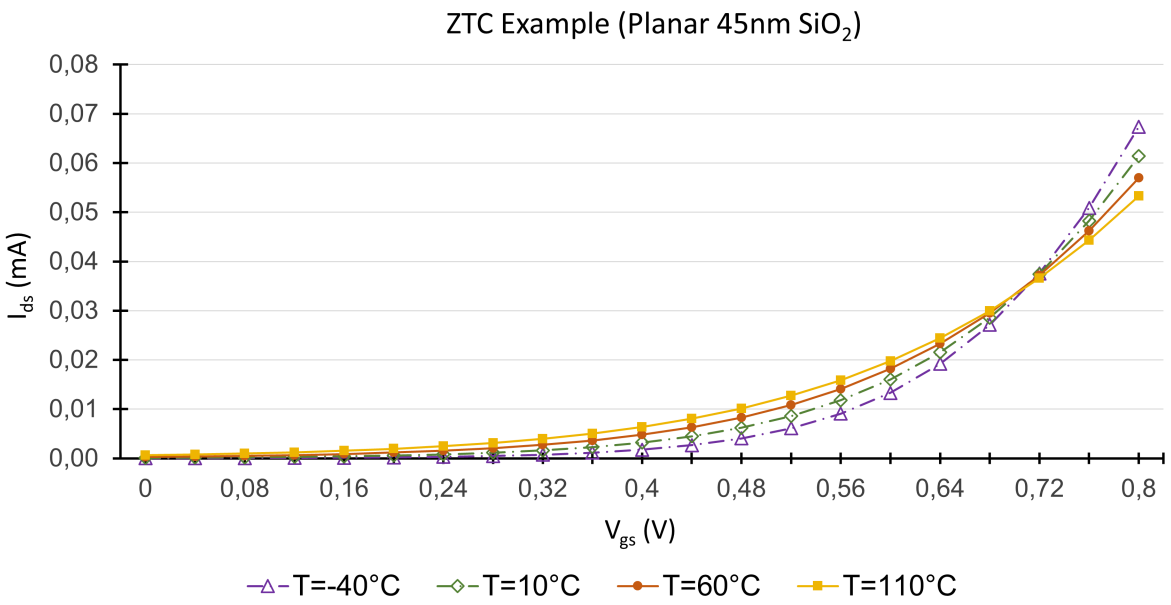


Figure 2.31: I_D vs. applied gate voltage for various system temperatures generating a ZTC. Simulations performed for isolated planar NFET using Imec planar Silicon gate oxide technology

Note, ZTC point can also be influenced by other parameters than temperature. For example, process variations of the system can shift V_{th} value [80] leading to different optimum between μ and V_{th} dominated effects. On that account, performing SA on temperature is advised to be done as one of the last steps when other effects have been taken into account and less variability of ZTC can be assured.

2.7.6. Variation of Process and its Impact

Process variation can be predicted only in general terms - fast-fast (FF) corner provides more current per area and also quicker response than typical-typical (TT) and thus higher and stronger throughput, while slow-slow (SS) achieves the opposite. The uncertainty is caused by lack of complete overview of how the process variation is implemented in the compact model used for simulations. [81]

Bear in mind, both corners are investigated even though sampling circuitry is not used. For latch and flip-flop designs both process variation modes have to be assumed critical, as FF can lead to data punch through if new data arrives prior to time (violation of hold time), while SS can lead to delay of one or more clock cycles (violation of setup time). [82]

2.7.7. Variation of Threshold Voltage and its Impact

Lowering V_{th} of the device leads to lower controllability margin of the system and increased 'off' power dissipation, but elevated current provision per unit area in 'on' state [83]. At V_{th} value of 0 V the device is always on and thus operates as a constant current source if gate and supply voltage is applied. Thus, it can be seen that for high throughput or heavy computation applications, where speed is more crucial than power consumption, low V_{th} devices would be used. To determine whether V_{th} shift provides substantial benefit, performance limiting factor has to be identified first.

In most cases insufficient slew rate is the main driver for design changes, meaning that saturation current has to be optimized. In such a case, system throughput performance degrades quadratically with increase of V_{th} as per Equation 2.4. However, if voltage swing values are insufficient, in most cases linear region current has to be improved - applied gate voltage is commonly larger than V_{ds} in settled devices for digital circuits. As a consequence, linear degradation of performance is expected with V_{th} according to Equation 2.6 when output swing imposed constraints are present. Overall, if V_{th} is low and only small threshold voltage changes are observed, the latter case is assumed to be critical as velocity current is reached early on and device is fully saturated. When V_{th} is high, velocity saturation might not be reached and thus first of the two conditions would most likely apply.

Lastly, it has to be noted that while power supply variation and V_{th} shift is analogous, they are not exactly the same. Increasing V_{th} reduces I_D generation, but at the same time, it lowers the V_{ds} value

required to enter saturation. If TERM can be varied, V_{ds} voltage can be made constant. Thereafter, drain current does not see as large of a current drop as is the case in reduced power supply directly.

2.7.8. Variation of Jitter and its Impact

Lastly, jitter effect on signal quality has to be investigated. Jitter can be approximated as noise in time which leads to bit width fluctuations and thus change in available charge/discharge time of the system for one data pulse. Thereby, it can be said that pulse width alterations are propagated throughout system and cannot be reduced in any way. For real signals, which do not have infinite slew rate, jitter can be partially remedied by high gain stages, which slightly increase the bit period with their high sensitivity of low voltages. For instance, if amplifier begins its operations significantly prior to signal reaching $\frac{V_{dd}}{2}$ point, bit width at the output becomes equivalent to $T_{bit} + \Delta t$ as shown in Figure 2.32 assuming infinite slew rate. Bear in mind, amplifier itself has an amplitude noise contribution to the system, thus in reality causing additional jitter to arise. [84]

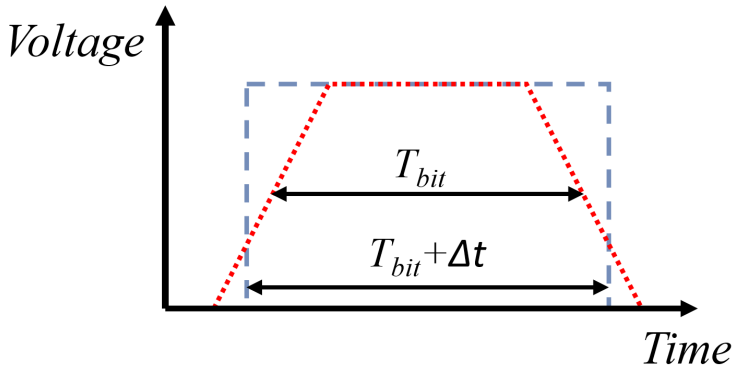


Figure 2.32: Bit width changes due to immensely high gain stage

It is hard to speculate on performance degradation proportionality to jitter as it mainly affects the duration of voltage being applied to the gate terminal. Thereafter, current characteristics are not affected - in case substantial headroom is given for output signal hold time, jitter will have limited effect on the performance. With increasing jitter, bit width is getting more narrow, hence in majority of cases critical point is reached only when settled output's hold time is violated. With the aforementioned in mind, no estimates on performance vs. jitter are provided here - one can only speculate that system employing DS will see lower performance drop than SES. The latter is true due to higher noise immunity of differential signalling in comparison to single ended one. [85]

With both, design and sensitivity analysis theory covered, choice of signalling topology and initial simulations can be performed. Prior to this, however, signal throughput validity conditions are defined and limitations imposed on the system are discussed. Hence, target case results and their discussion for various device technologies can be found in Chapter 3 after aforementioned considerations have taken place.

3. Selection of Topology and Design Case Simulations

In previous chapter broad list of options for possible I/O configurations and their component design was provided. However, as it is unfeasible to perform an investigation for every combination of the given parameters, design space has to be narrowed. With sensitivity analysis being a large part of the project, it is decided to select only one SES and one DS for implementation in simulation environment. Else there is a significant risk of not reaching design goals in allocated time for MSc thesis. Additionally, to not impose harsh restrictions for full system designers using a commercial memory chip, overestimation of signal disruptive elements has to take place. For instance, most lossy TL type for more than required interconnect length is used to ensure that in majority of cases system significantly outperforms minimum signal quality requirements.

With former in mind, general considerations for any I/O link used in this thesis are provided in Section 3.1. Further, performance metrics and system constraints are discussed (see Section 3.2). Design choice for a SES and DS is elaborated in Section 3.3, while TL, TERM and PACK design is given in Section 3.5. Brief discussion on various device manufacturing flows and their inherit properties are covered in Section 3.6 Finally, discussion on simulation results for design case can be found in Section 3.7.

3.1. Top-level Implemented I/O Interface Characteristics

Prior to analysing processing unit and memory interconnection, the governing interface characteristics for all designs covered in this thesis have to be established. Predetermined feature set allows to properly set up simulation environment and limit the design to desired complexity. For instance, the interface is assumed to incorporate only point-to-point interconnects. As large pin-out NAND flash memories tend to have more than one independent memory channel, the need for using a multidrop on the same data line is diminished, except for fail-safe (redundant component) design. Each data line of a 8-bit bus is going to have a unique connection to a processing element. [13]

Note, address line is likely shared across multiple components simultaneously, thus, for specific signals multidrop configuration is inevitable. However, for simplicity, only the data I/O is investigated in this thesis. Additional investigation on multidrop imposed design constraints as impedance miss-match due to simultaneously active outputs, more severe reflections, additional loading due to leakage in idle circuitry among others [64] is left as future exercise.

Another decided upon governing characteristic of the design is having an external interconnect of 4 cm between memory and PU using PCB trace, which is terminated at the receiver and/or source side internally. The exact length is defined as memory I/O (DQ for large pin-out chips) pin distance to longest chip side (1 cm) doubled and taken twice, assuming a mirrored conditions for receiving IC. Bear in mind, chips are commonly put almost next to one another, thus spread of 2 cm between two chips is a substantial overestimation. Nevertheless, this allows designer the freedom of using long serpentine traces for trace length matching [86]. Consequently, to prevent inter-trace delay, equidistant line spacing and equivalent route length for all traces is defined. TL considerations are further elaborated in Subsection 3.5.1.

Moreover, the power supply voltage is set to be 1 V as it is the characterisation voltage value of Imec FinFET 14 nm technology. The nominal value of power supply for FinFETs is 0.8 V implying that the used V_{dd} leads to a significant over-voltage and hence would result in low reliability designs. Nevertheless, it is assumed that almost constant area scaling factor between 0.8 V and 1 V exists irrespective of system conditions. Thus, the main supply voltage used is assumed irrelevant for analysis in this thesis as direct scaling can be used. However, the former claim has to be verified as is partially done during SA (Section 4.6). Additionally, using 1 V for V_{dd} simplifies power and area comparison between FinFET and planar (45 nm node) technologies, as nominal voltage of 45 nm node is 1 V.

Lastly, to simplify analysis, only the last stage of TX and the first stage of RX are analysed. As gate(-s) of the input transistor(-s) effectively isolate (pseudo-infinite impedance) majority of prior circuitry RLC contributions, only noise and jitter of pre-TX stages are assumed to influence signal integrity. It has to be noted that such an approach overestimates system's reliability and response quality [35]. To load the RX, a duplicate of the receiver is added at the output to ensure a non-zero fan-out of the system.

3.2. Simulation Constraints and Performance Metrics

In Section 1.2 main goal of the thesis is defined to be reaching 8 Gb/s of data transfer for 3D NAND I/O using 14 nm Imec in-house developed FinFET technology. From the above one can conclude that system throughput is the main characteristic to be measured and evaluated with respect to certain requirements. Hence, timing conventions as slew rate and output signal hold time are used to assess whether necessary DR is achieved. If either of signal properties is violated, consequent active circuit/system stage might miss-interpret incoming signal leading to bit errors - system does not have enough time to settle [85]. Thereafter, design is said to reach 8 Gb/s DR only when it is able to transmit signal with limited number of mistakes for a specified time period. For this purpose a signal eye (see Subsection 3.2.1) can be constructed and inspected.

3.2.1. Quality Requirements

An eye diagram is a graph that represents entire time response of any system node - full transient simulation is cut into period long instances which are overlapped creating eye-like shape when centered. When constructing an eye, data period is more commonly referred to as unit interval (UI). An example of an eye can be seen in Figure 3.1. Notice, a hexogon is added in the centre of the plot which is commonly referred to as compatibility eye mask (CEM). CEM allows to visually represent minimal signal conditions which have to be met at a particular system node to ensure that signal can be properly evaluated. Here, the most important properties are the minimum height, width and slew rate. Note, the height is proportional to signal swing, while width indicates minimum remainder of bit width when jitter is present. In this thesis, CEM is used as signal quality evaluation tool. [85]

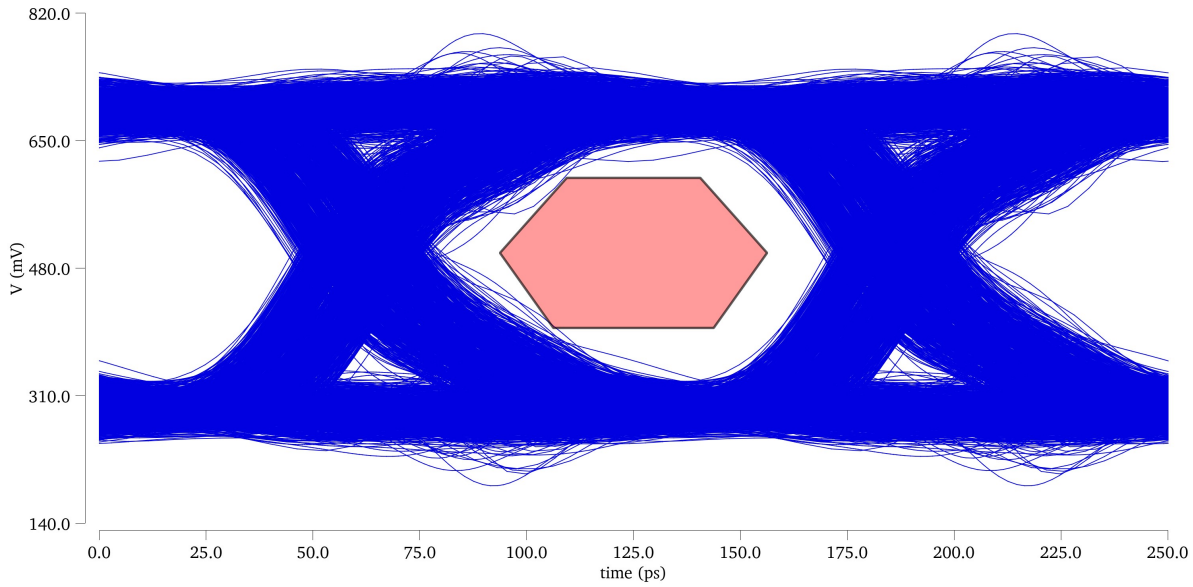


Figure 3.1: An eye diagram including compatibility eye mask for CTT topology with following simulation conditions: Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

For conventional architecture where memory and PU are independent entities, CEM is usually defined at RX input to avoid limiting designers on choosing any memory-PU compatible IC pair themselves. Thereafter, already predefined NAND I/O CEM is used from this point onward, where weakest available PU is assumed, implying highest TX drive strength is required. The conditions are as follows: eye opening has to be at least 200 mV in a centered continuous time interval of 30% UI and eye width has to be larger than $\frac{1}{2}$UI. Note, signal center is assumed to be at half ideal bit width. Slew rate conditions for both rising and falling edges are taken to be symmetrical, thereby minimum slew conditions applied can be directly determined from the above. [13]

Note, to optimize the design to minimum area, signal eye has to be almost perfectly matched with CEM. Thus, size optimization technique is rather simple - if either slew rate or amplitude requirements are violated, either driver area is increased or TERM value shifted. If driver consists of several stages,

size ratio between stages upon area scaling is kept close to the same, meaning that all TX devices are proportionally reshaped by the same factor. The former strategy is most accurate when external interconnect's effective parasitic capacitance is substantially larger than TX intrinsic output capacitance. In such conditions active device self loading is playing very minor role in system's signal behaviour. Nevertheless, when area is increased, TX output impedance shifts which can lead to signal deterioration if high miss-match conditions are present and no source termination has been added. In case undesired system behaviour is observed, sizing is performed on stage-by-stage basis.

Simulation length is decided to be 400 ns for all design cases, which correspond to 3200 bits for 8 Gb/s DR. The first 200 bits are effectively ignored to allow for sufficient system setup time. For verification purposes used bit count is significantly too low as bit error rate (BER) of 10^{-9} cannot be checked and it is unclear whether any of pre-defined critical bit transitions are present in the transient analysis [12]. Using longer simulation time significantly slows down optimization loop, thus, limiting breadth of analysis, however it provides considerably higher accuracy of system behaviour and confidence in device area estimations. Assumed system response can be said to be under-determined, implying that actual area has to be larger than one obtained via simulations - required scale-up factor investigation is left as future exploration. Applying the same transient time limitation for all designs of different topologies ensures that precise comparison in-between designs is acquired - scale up factor is of approximately the same magnitude.

With signal quality requirements known, signal degradation mechanisms and limitations are looked into. The more factors reducing signal quality are introduced in simulation environment, the better approximation of real life component is acquired. Thereafter, most severe degradation mechanisms to first degree are included in simulations - discussion of them can be found in the next subsection.

3.2.2. Signal Limitations and Non-idealities

Signal integrity and reliability of any electronics system is challenged by both internal and external disturbances as noise, crosstalk, path losses and component non-idealities. Signal loss and degradation mechanisms have to be understood to allow for attenuation prevention strategy establishment and enable efficient high-performance system design. With majority of signal loss elements implemented, accurate sizing of the system can be obtained giving high reliability that a manufactured product will operate as desired. Thus, signal degradation factors as amplitude noise, jitter and crosstalk are analysed and explained in this section.

Noise and Random Jitter

System configuration used in this thesis can be defined as digitally driven and digitally receiving interconnect, meaning that majority of the noise is going to manifest on the output as bit-width variations or jitter. For instance, different rise and fall transition slew rates of gate voltage cause a one-sided delay in device turn-on/off characteristics leading to a variation in bit width. Amplitude noise can be seen to possess similar effect - the higher average positive/negative deviation from power supply voltage due random noise is observed, the faster/slower the signal response is. The main cause for such behaviour is the self loading of interconnect capacitance - stored charge requires more/less time to be discharged ($Q = Idt$) in case current through system reduces/increases.

Jitter and noise directly propagate throughout the system, thus, their application point does not matter per se. However, to capture circuit response at any node, random noise and jitter is superimposed on input voltage. The noise source is set to represent noise coming from memory cells directly and routing prior to I/O circuitry. However, as a physical NAND I/O component is not accessible for measurements, noise amplitude and frequency pairs are merely assumptions estimating worst case contribution. Note, such noise and jitter are random, with semi-deterministic range - guesstimated without any component testing.

Jitter is added to V_{in} by randomly varying rise and fall time fractions of a square wave as can be seen in Figure 3.2. In simulation the slew rate variations are implemented by using verilogA block, which generates quasi-uniformly distributed rise/fall times as fraction of a single bit width. To acquire uniformity, trigonometrical functions are employed as shown in Figure 3.3, resulting in rise/fall time being a value in range of [0, 0.25] T_{bit}. Notice, as random numbers generated are converted into radians when used by a trigonometric function, rise and fall time fractions can represent any value in the range above with equivalent likelihood. Observe, using 2 random value factors leads to different

[12] URL `https://www.viavisolutions.com/en-us/products/bit-error-rate-test-bert` [cited on 17th of March 2023]

rise and fall times (see Figure 3.2). For differential signalling the same instantaneous slew factors are assumed for complementary signals, to keep the mirror-symmetry intact.

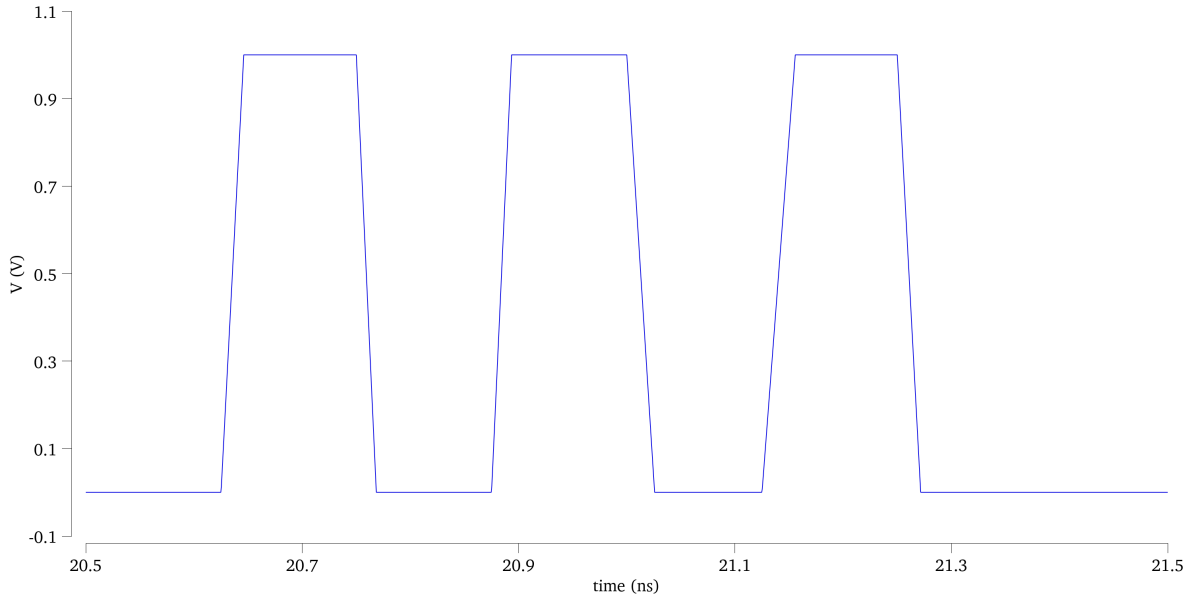


Figure 3.2: Jitter effect on an ideal square wave signal

```
analog begin
        $bound_step(period);
        x = $random;
        y = $random;
        fall_time_shift = abs(sin(y)*sin(y)*cos(y)*cos(y));
        rise_time_shift = abs(sin(x)*sin(x)*cos(x)*cos(x));
        V(Out) <+ transition(V(In), 0, rise_time_shift*period, fall_time_shift*period);
end
```

Figure 3.3: Varying slew rate implementation using VerilogA

Input signal after slew randomization is passed through an exponential horn - 4 inverters in chain with an increase factor of 1.7 per stage. Determination of scaling factor is performed using linear delay theory of optimum stages and scaling, which is also elaborated in Subsection 3.3.4 [87]. The inverter chain is used to convert the randomized slew rate into jitter and to obtain a finite drive strength capabilities of V_{in}. Scaling factor of 1.7 can be directly applied to planar device circuitry, while FinFET devices require rounding to the nearest integer due to FinFET sizing discreteness. The difference in signal eye with and without jitter can be seen in Figure 3.4.

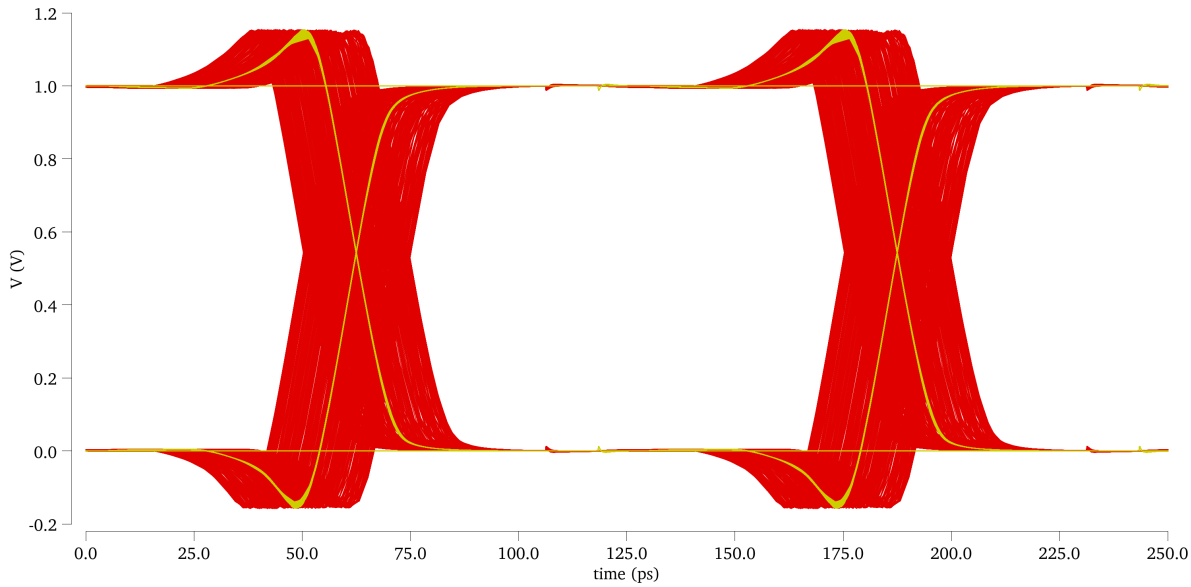


Figure 3.4: Difference in signal eye at inverter output with (red) and without (yellow) jitter

The quasi-ideal signal at the output of exponential horn is further superimposed with amplitude noise to distort both, signals quickest components (transition edges) and deviate voltage levels away from the rails (V_{dd} and ground (GND)). The former effect is attained by applying noise of lower than signal frequency while the latter effect is achieved by adding higher than signal frequency noise. Signal distortion due to amplitude noise can be seen by comparing quasi ideal (post exponential horn) signal in Figure 3.5 with noisy signal in Figure 3.6. Note, slight bit width variation can already be noticed in Quasi-ideal signal, with 2nd bit being $\approx$10% narrower than 1st and 3rd bit. The noise is set such, that the average rail value differs up to 10% from V_{dd} or GND and instantaneous noise varies up to 30% of V_{dd}. Implementation of amplitude noise is achieved via VerilogA in a similar fashion as done for the jitter shown previously, thus, code snippet is omitted here.

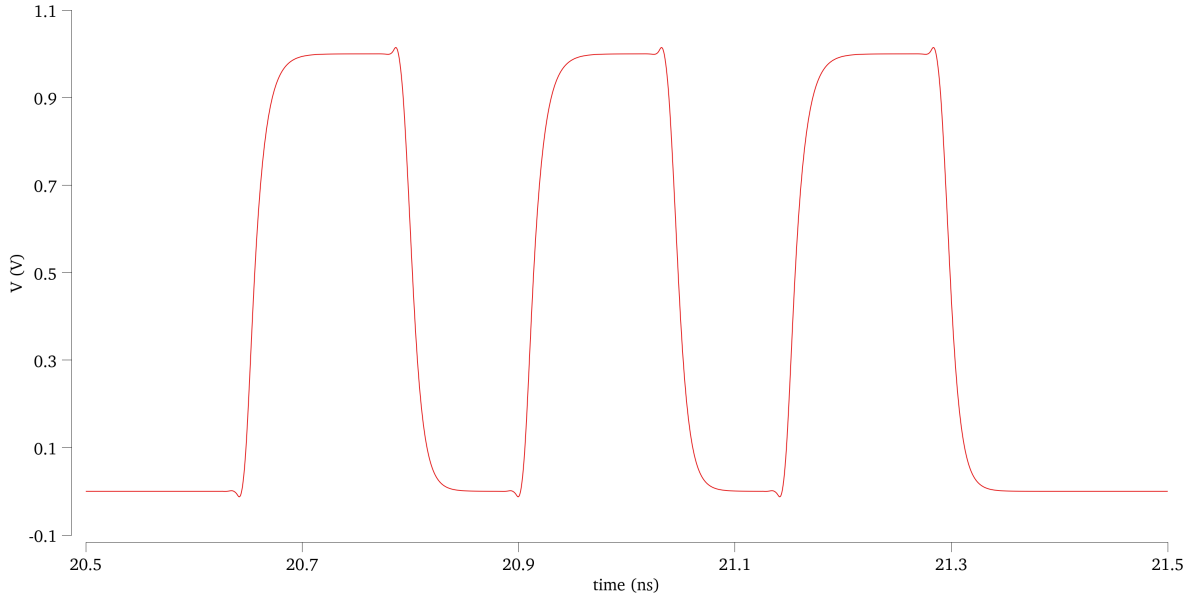


Figure 3.5: The quasi-ideal signal at the output of the inverter chain

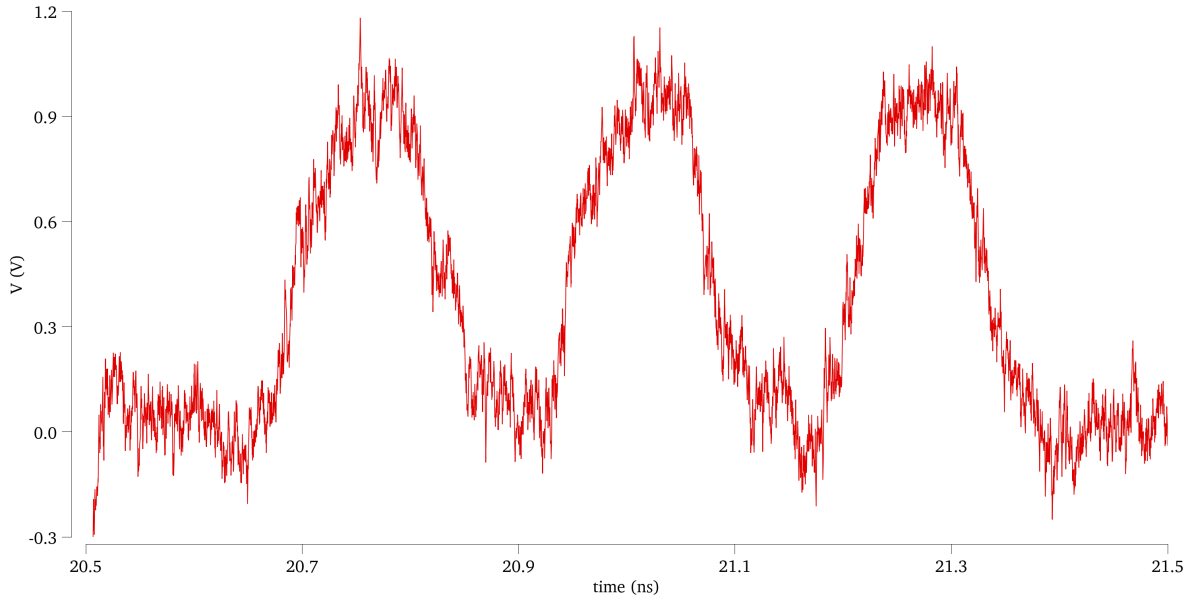


Figure 3.6: Quasi-ideal signal with the addition of noise

It has to be noted that imposed noise is discontinuous - the amplitude is applied at discrete frequencies such that overall simulation time is not substantially increased, but enough distortion effect is present in the system. Noise frequencies are selected such that they are not an integer multiple of one another and severe resonance on TL is prevented. Main culprits for simulation time growth are high frequency components as they require a finer step to be properly approximated.

After applying noise, ringing on TL and the input of RX for DS was observed. This has to do with the fact that only a single stage of transistor circuitry is located between V_{in} and TL for conventional driver, preventing the full conversion of amplitude noise into jitter (direct noise propagation due to Miller effect). Thereafter, to minimize the oscillatory effects, a low-area buffer prior to TX was added as was shortly mentioned in Subsection 2.2.2. With noise and jitter implementation covered, system response to crosstalk presence can be looked into.

Crosstalk

Memory-to-chip interconnect is inevitably composed of several bitlines running in parallel, whether it is data or various enabling signals. Thereafter, regardless of how optimally the lines are spread out, there is always some degree of electromagnetic coupling present between any two lines. Routing is predicted to become even more dense in future due to increasing number of data lines and area miniaturization at the same time [88]. Thereafter, analysing crosstalk is especially crucial for high speed data transfers, since crosstalk becomes one of if not the most detrimental signal degradation mechanism [23]. Crosstalk gets especially elevated in high-speed signals because of more often transitions from low-to-high and high-to-low in shorter periods of time. More frequent shifting requires steeper slew rate, which is proportional to coupled noise: $V_{noise} = L_{mut}\frac{dI}{dt}$ & $I_{noise} = C_{mut}\frac{dV}{dt}$ [64].

Even though crosstalk is present throughout the system - also in on-die FETs [89] and routing [90] - only external (on PCB) crosstalk is implemented in this thesis. The main reasons for such decision is the complexity of internal crosstalk modelling and the immense simulation time increase it leads to. Additionally, on chip crosstalk can be minimized by performing careful chip layout design, which is assumed to take place here. Note, as modelling and investigating all the disturbances in a NAND I/O is not the ultimate goal of this project, various major simplifications for crosstalk implementation are made as discussed further in this section.

First, in built Cadence® Virtuoso® crosstalk modelling components as *ncline* and *stackup* are going to be used rather than distributed TL elements with mutual inductance factors. Thus, crosstalk is assumed to be present only for a certain distance on the PCB where traces are perfectly parallel to one another. Bear in mind, most of the chips nowadays use non-parallel traces implying reduced EMI, however, analysis and implementation of such circuitry is more resource and time consuming [91].

Using parallel lines is not entirely realistic as trace length matching techniques would be required to ensure close to no inter-signal delay. For simplicity, it is assumed that length matching is performed by employing serpentine tracing [86] close to the transmitting and receiving end, where coupling is neglected. Also self-coupling due the serpentine pattern presence is ignored. To partially compensate the lack of crosstalk caused by neglecting non-parallel traces, it is decided to artificially increase the overlapping coupled line length causing higher signal degradation. Thereafter, overestimation of degradation is present, which is deemed beneficial for this particular design exercise.

To limit the size of the system and thus the time it takes to run a simulation, only the adjacent signal lines are assumed to be affecting any chosen trace [70]. To ensure validity of the assumption, guard lines - ideal ground/power supply traces - are added in between two neighboring lines, increasing each line's coupling to the return path [86]. Nevertheless, using guard lines as a crosstalk preventive measure lead to other potential issues as line length miss-match to be compensated due to increased PCB area and thus longer routing length. Note, there is no functional difference in using V_{dd} or ground lines as guard traces, since both signals are DC voltages, thus seen as AC grounds.

Worst case degradation due to crosstalk happens when simultaneous switching is present [92]. Due to implementation of jitter, the average transition take-off time values have shifted, and thus worst case point has to be determined anew. Assuming that aggressor traces are not ideally length-matched, close to worst case degradation is found if 7% and 11% bit width delay is present in the neighboring lines, equating to 1.5 and 2 mm of length deviation for 8 Gb/s case. Note, signals on each line are set to have a different bit pattern and thus simultaneous same edge switching is not guaranteed.

Delay in combination with unrelated signals strikes a good balance between even and odd mode coupling. Even mode corresponds to adjacent lines having the same magnitude and transitioning in the same direction, while odd mode implies exactly the opposite (opposite magnitude and different direction). Even mode coupling causes overshoot of the signal, potentially leading to reduced long-term reliability and device lifespan due to increased voltage stress on the device [71]. Contrarily, odd mode coupling causes an undershoot leading to increased likelihood of crossing CEM, which in turn infers a high chance of bit error [64].

To reduce corsstalk generation, various prevention mechanisms could be used. However, to ensure that a system would operate in high stress conditions, performance analysis is limited to worst case scenario. Note, an exception is guard line spacing - even though spacing between lines can be reduced with lower DR to obtain worst case scenario, spacing is kept constant for all simulations. Else, comparison between different designs would be meaningless as general conditions would differ.

Notice, to reduce crosstalk induced signal variations, circuit sensitivity has to be reduced. To do so, output capacitance of TX has to be increased without increasing the current. Slew rate of the system is negatively altered (see Equation 2.51) and thus crosstalk is reduced. In case swing amplitude is the limiting quality requirement, small artificial capacitance can be introduced. Directly up-scaling TX will increase both current and output capacitance retaining almost constant, if not increased slew rate.

Voltage Source Limitations

Limiting power supply effectiveness in simulations is a necessity to verify design's likeliness to perform well with non-ideal drive strength at its terminals. The first realistic consideration is assuming an external power supply, meaning that a complete current path from source to chip is required. For simplicity, the link used for chip-to-chip interconnect is established as a baseline for supply-to-chip connection, only with a shorter TL length (2 cm).

To keep routing distances short, it is assumed that there are 3 independent power supplies - one for memory module, one for TX and one for RX. When shared power supply is used, interconnect from power supply to furthest IC could experience significant drop on rail voltage, leading to immense reduction in component performance. Furthermore voltage source is assumed to be non-ideal and having 2 Ω internal resistance [13]. For highly parallelized power grid the overall circuitry resistance is rather low causing a significant loss of voltage swing on the devices due to potential division. The former serves as an additional factor of imposing separate voltage supplies or using a higher voltage shared supply to avoid too large of a voltage drop across components.

Moreover, to limit power supply variations on chip level, decoupling capacitors close to every component of supply-to-chip connection have to be added. As already mentioned in Subsection 2.7.4 lack

[13] URL `https://www.telerex-europe.com/content/files/pdfs/productPdfs/TR/tsr1sm.pdf` [cited on 30th of August 2023]

of decoupling capacitors would lead to undesired oscillations on voltage rails as high frequency components would not be filtered out. With the aforementioned considerations in mind, a model of a single power supply used for the simulations can be seen in Figure 3.7.

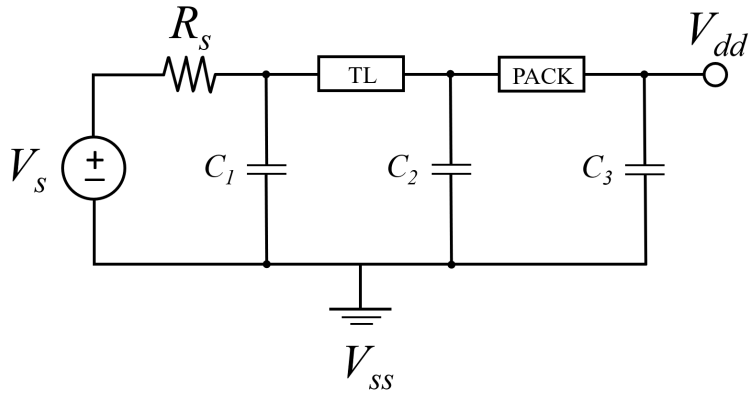


Figure 3.7: Non-ideal voltage source simplified schematic

Amplitude noise can be superimposed on the power supply in exactly the same manner as it was done for V_{in}. Nevertheless, simulation tests showed that almost all the noise gets rectified by decoupling capacitors, leaving negligible amount of high frequency components on the supply voltage. Thus, noise sources are removed from power supply to reduce simulation time.

Miscellaneous Performance Degradation Mechanisms

Not every degradation mechanism can be integrated into simulation environment, as that would lead to unfeasible computation time and too high design complexity. In this section signal degradation factors not or partially implemented in simulation environment are discussed. First, the device model files were not altered - with exception of gate work function for V_{th} variation. Thus higher order device effects are limited to whatever compact model file includes. For instance, during SA it was determined that temperature variation of device is incomplete as it is not matching to expected trend obtained from literature (elaborated in Section 4.8).

Second, all passive components are assumed to be ideal. It is well known that capacitors have both an inductance and resistance contributions, making it a band reject filter [76]. This also applies to only partial accounting of on-chip routing parasitics (resistance and capacitance) which can be detrimental in high sensitivity nets as gate voltage of barely saturated tail current source. Even a relatively light drop of gate voltage can lead a current source transistor into cut-off region reducing its generated current drastically. Hence, improper routing as using long minimum width interconnects has to be avoided as on-chip metallic layers are highly resistive. The only routing which is accounted for in all designs is metallic interconnect prior to TX driver, being located on 3rd metallic layer.

Thereby, assuming trace cross-section of 36x36 nm and length of ≈1.5 μm DC resistance can be approximated to be ≈25 Ω ($R_{DC} = \frac{\rho l}{A}$). Note, AC resistance is negligible as skin effect does not penetrate nano-meter thick components at 4 GHz frequency. Capacitance of the interconnect can be approximated as twice the parallel plate capacitance of 2 adjacent lines ($C = \frac{\epsilon_0 \epsilon_r A}{d}$). Assuming spacing between the lines (d) to be equivalent to 72 nm and ϵ_r of 11.7 (pure silicon), parasitic capacitance can be determined to be 0.3 fF as shown in Equation 3.1. Note, the determined capacitance value is negligible - it can be assumed that other parasitic contributions discussed further are going to have a more detrimental impact on capacitance.

$$C = 2\epsilon_r \epsilon_0 \frac{A}{d} = 2 \cdot 8.85 \cdot 10^{-12} [m^{-3} kg^{-1} s^4 A^2] \cdot 11.7 \cdot \frac{1.5[\mu m] \cdot 36[nm]}{72[nm]} = 0.3[fF] \tag{3.1}$$

Using ideal components directly leads to omission of another signal degradation mechanism known as simultaneous switching noise or ground bounce. Realistic ground and traces can be modelled as an RLC network having finite values for each component. One can notice that upon signal switching, voltage on GND inductance varies per $\Delta V = L\frac{dI}{dt}$. A single trace does not lead to a significant variation

on the supply rail, however, in large chips with many signal transitions happening concurrently, the overall inductance can be seen to add up. Moreover, with increasing data rate and limited further power supply scaling, steeper slew rate is required to meet signal quality standards, thus increasing $\frac{dI}{dt}$ component of noise voltage. [44]

Note, varying supply voltage directly affects the drive strength of active devices as V_{gs} is reduced, implying that less I_d is produced per unit area. This directly leads to longer data propagation times or less reliable system in case over-voltage is present [27]. Even though ground bounce can be a severe signal degradation mechanism, its implementation is left for future design investigation. At the current stage developing an accurate RLC model for a non-ideal ground is deemed unfeasible due to lack of knowledge on application specific details as PCB layout and ground plane implementation. Nevertheless, it is assumed that all ground bounce preventive measures would take place and hence the effect can be ignored in first order analysis.

The cheapest and most simple ground bounce prevention technique is implementing more ground and V_{dd} bonding pads on the chip [93]. By doing so the overall inductance of current return paths would be minimized and capacitance maximized, thus reducing noise amplitude and eliminating high frequency components. Another relatively simple approach is adding different resistances at the gates of same stage parallel devices (fingers) to artificially impose a slight delay between transitions. Note, however, if high number of fingers is used in a single stage, even a slight delay between each consequent stage could cause severe implications on tight timing budgets of high speed systems [22]. Thereafter, the former approach is viable only when long bit periods are used, where same stage inter-delay can be neglected in comparison to bit width.

Alternatively, if possible, output voltage swing or load capacitance can be reduced, as then less time is required to load the output to required swing value in iso-area conditions. Additionally, slew rate should be adjusted to minimum allowable value to guarantee the lowest switching noise. Lastly, as mentioned in Subsection 2.7.4 using decoupling capacitors reduce the noise on the power supply lines by filtering the AC voltage components [94].

The last signal altering component not implemented in the design is electrostatic discharge (ESD) protection. ESD protection has to be added at every chip node that has a connection with external environment - in case other statically charged entities (PCB, probe, human hand) are brought in the vicinity of chip pin, air between both components gets ionized and high voltage discharge occurs [21]. The high voltages generate currents several magnitudes higher than those in normal operations for a short period, which can lead to gate-oxide insulator breakdown. ESD is especially damaging for nanometer transistor technologies as gate oxide becomes thinner and physical dimensions as width and length are continuously scaled down. [95]

Protection against ESD can be achieved by generating a low impedance path which accumulates and discharges static charge directly to the ground bypassing on-chip circuitry. To withstand high voltage generated stress, protection circuit has to be large in size and quick enough to manage a complete discharge before chip internals are harmed. Using large area devices considerably increases the node capacitance restricting maximum achievable bandwidth. Thereby the ESD protection circuit should not be over-designed, such that just enough protection is provided with no penalty on the system speed. To not disturb system behaviour, ESD protection should remain almost idle during normal chip operations, meaning that shifting from low to high impedance state has to be present during regular operational conditions. Moreover, to provide satisfactory protection, voltage at chip pins has to be capped slightly below oxide breakdown voltage. [22]

To limit the current and thus voltage throughout the system, a resistor along the signal path can be utilized. However, using a passive resistance can cause heat generation on the chip which has to be dissipated using active or passive means - else temperature related reliability issues might arise. Moreover, resistance introduced as any other component generates noise which causes SNR to drop. [95]

Contrary to series resistance, parallel devices providing a low impedance path are usually realized using active devices. One option is to use forward biased polysilicon diodes connected parallely to signal path, as such devices can tolerate high currents. Nevertheless, the impedance of forward biased diodes is strongly related to the applied voltage, meaning that impedance miss-match can appear even during normal operations causing unwanted reflections. Also, the per unit area resistance of diodes is high. Thus, significant area would be required to ensure that overall resistance is low and right voltage at the input port is clamped. [96]

Instead of using diodes, it is possible to use a reverse biased off-state FET device. For instance, NFET device with gate and source terminals grounded and drain experiencing ESD voltages can generate high enough electric field across the device causing breaking of covalent bonds and hence creation of free charge carriers. The free charge carriers are further accelerated leading to process repeating itself - this is better known as avalanche breakdown. Using such devices can be more beneficial than diodes as during regular chip operations the transistors contribute to power dissipation only by small leakage current. However, similarly to diodes, also FET devices introduce a large capacitance at the node of application, thereby making impedance matching a difficult task. [95]

With the former in mind, a conventional ESD structure looks as depicted in Figure 3.8. Notice, two primary ESD structures are added - one before and one after current limiting resistance. The second stage is implemented to assure that all the ESD generated current is discharged in case the first stage is not enough even when fully employed. With former in mind, using a three stage ESD protection makes it a fail-safe system even with high applied voltages. Bear in mind, for ESD protection to be most effective, it has to be placed as close as possible to external pins while as far as possible from other chip components. [22]

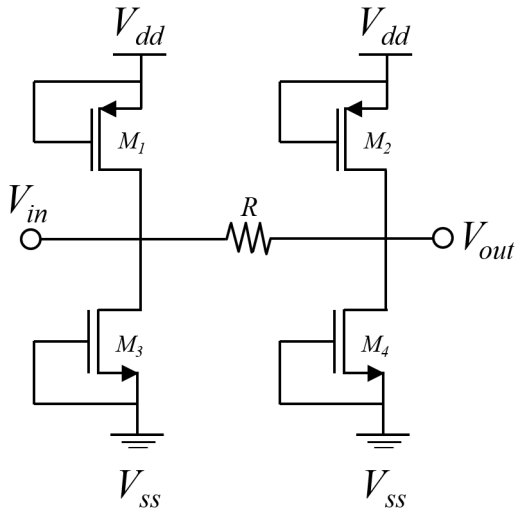


Figure 3.8: Basic ESD protection circuitry

With all the limitations and performance metrics defined, it is the right time to choose topologies from each signalling mode for which the simulation environment implementation is performed. Sizing and power estimations of previously discussed topologies is discussed in the following section using the newly set performance metrics.

3.3. Selection of Design Topology for Simulation Implementation

Power estimations for requirements specified in Subsection 3.2.1 are derived and first hand sizing is performed. Additionally, compatibility of the design to the particular design case rules are investigated - whether signal swing can be centered at $\frac{V_{dd}}{2}$, output high and low are of approximately equivalent strength and design can drive a high load output.

Due to the time limitations of this project, it is practically impossible to investigate all topologies in depth, especially with design's sensitivity analysis being one of the thesis objectives. Thereafter, it is decided to carry on with only one design per signal class to ensure a study of broad range of SA parameters as mentioned at the beginning of this chapter. Otherwise, re-optimization of design during SA for an extra design would lead to significantly increased time consumption of one parameter (estimate of 25-35%), causing lower remaining time for diversification of criteria. With the objective of design narrowing placed, each topology is analysed in more detail and then the choice of SES and DS topolgy is made as described in the following subsections.

3.3.1. Sizing and Power Analysis of CTT

Power analysis covered in Subsection 2.2.1, can be directly applied here with requirement defined values plugged into equations. Ideally, the off-state static power should lie around 5 mW for a 1 V power supply - a current of 5 mA is required to cause a full voltage drop of 1 V on the output termination (series 200 Ω). However, if termination is biased towards any of the rails by increasing either TERM value, the off-state power reduces. Note, to further reduce static power dissipation, controllable TERM can be used, which can be made idle when no input data is received.

The total power consumption can be determined using on-state conditions of the driver and assuming ideally matched network. Nevertheless, when ideally matched V_{out} can be noticed to have a swing of 333 mV - parallel network depicted in Figure 2.4 would result in equivalent resistance of 150 Ω with one of the resistors being 100 Ω. Hence, the power consumption can be determined as 6.67 mW. To reduce the swing to 200 mV such that minimum circuit size is acquired, the area of FET devices can be simply reduced. The area has to be set, such that PDN/PUN equivalent resistance results in 66.7 Ω, which corresponds to required voltage swing of 200 mV. The resistance value is obtained by setting the circuit transition voltages of high state to 600 mV and low state to 400 mV. With this, total power consumption of CTT can be determined as shown in Equation 3.2. By reverse engineering r_o value is found to be 200 Ω.

$$P_{tot} = \frac{V_{dd}^2}{R_{top} + \left(\frac{1}{R_{bot}} + \frac{1}{r_o}\right)^{-1}} = \frac{(1[V])^2}{100[\Omega] + 66.7[\Omega]} = 6[mW] \tag{3.2}$$

Note, 200 mV opening does not account for output amplitude variations of up to 70 mV due the non-ideal charging of the circuit, which is highly dependent on incoming data pattern. The actual average swing is thus closer to [200, 300] mV - with higher swing for higher data rate. Thereafter, average swing for power determinations is assumed to be 250 mV for all topologies, implying that CTT total power is approximately 6.25 mW (Equivalent resistance of 160 Ω).

From the above, dynamic power fraction using PDN can be simply determined as $\frac{V_{OL}^2}{r_o}$. With V_{OL} value equal to 375 mV $\left(V_{cm} - \frac{V_{diff}}{2}\right)$ and r_o of 150 Ω, dynamic power dissipation can be determined to be approximately 0.94 mW. Note, if TERM value skewing is present, instantaneous dynamic power fraction is going to increase for circuit network in parallel to increased TERM resistance - equivalent resistance has to be kept constant, thus r_{o_1} has to be lowered. On the other hand, instantaneous dynamic power of the complementary circuit is going to reduce as to keep V_{OL} and V_{OH} (high state voltage threshold) ratio to be the same - now the reference TERM value is higher than 100 Ω. Thus, to keep the same potential division properties, also the complementary equivalent resistance has to be increased, implying increased r_{o_2}. Observe, indices 1 and 2 are used merely to distinct whether PUN or PDN is discussed - neither 1 or 2 indicates a specific network type as TERM skewing direction is undefined. Even with the variations in individual dynamic power, the average P_{dyn} should remain largely unchanged.

For higher data rates the average dynamic power increases because of rise/fall times taking up a higher fraction of bit width. More frequent and relatively longer (w.r.t period) rise and fall times lead to both transistors of the inverter being operational at the same time more often. The alternative current path does not contribute towards generation of V_{out} swing, rather power is purely dissipated due to complementary transistors opposing each other during transition. Other factors as crosstalk and attenuation are going to cause swing variations, which to the first order are assumed to be implemented in 250 mV average swing, thus compensated in full effect.

With CTT power determined, first order size estimation can be performed. In this thesis a chain of semi-exponentially sized inverters are selected for TX driver, with them being the most common and simple I/O driver.

To determine the size of the inverter chain (exponential horn [21]), linear delay model can be used as a first-order estimation tool [87]. For linear delay theory to hold, it is applied only at the early stages of the design before introduction of disturbances as crosstalk and noise. Thereafter, the model can be said to be accurate solely in approximation of cascaded inverter stages. Exponential sizing is used as initial guideline, with sizes being adapted to meet performance requirements. Note, optimization of exponential horn stages is used primarily to minimize the propagation delay - even relatively large deviation from optimum scaling factor does not cause rapid increase in time delay [87].

The optimal number of stages can be acquired by using Equation 3.3, where N is the number of stages, C_L and C_{in} are the load and input capacitances and α is the scaling factor between two consecutive stages. [21]

$$N = \frac{\ln (C_L/C_{\text{in}})}{\ln(\alpha)} \tag{3.3}$$

By looking at Equation 3.3, it can be derived that first step in the determination of N is estimating the load and input capacitance ratio (electrical effort). However, here the first problems arise as TX is the last stage of I/O transmitter, thus, the input capacitance can be merely guesstimated. The load on the other hand, can be seen to consist of all the capacitive elements of the packages, transmission line and receiver input gate capacitances. The aforementioned capacitors have to be charged for RX input voltage to be at a stable level for majority of the equivalent bit width.

Transmission line contributes to the largest portion of system capacitance, equaling to a total of staggering ≈5 pF effective capacitance (refer to Subsection 3.5.1 for values). Total parasitic capacitance of the packages reaches around ≈2 pF (refer to Subsection 3.5.2 for values). The RX capacitance can be said to be negligible even with significant circuitry size as intrinsic FinFET capacitance per μm is only in fF range [97].

Determination of input capacitance is slightly more complicated process - it requires 1st inverters intrinsic capacitance determination and prior circuitry capacitance estimation. Since FinFET transistors are used in designing the interconnect, inverter's PFET and NFET capacitances are assumed to be approximately equal. Thus, both input and output parasitic contribution can be simplified to two times NFET respective capacitances [97]. To avoid over complication of calculations, the in-Cadence® Virtuoso® obtainable DC operating point values are used for calculations. To obtain transistors DC values, a simulation of isolated transistor biased in linear region is performed. Triode region is selected merely because capacitance value in this region is the largest and TX is biased in triode region during nominal circuit conditions [98].

For approximation of effective NFET input and output capacitances, the following expressions are used, with C_{gs} being gate-to-source capacitance, C_{gd} gate-to-drain capacitance and C_{ds} being drain-to-source capacitance:

$$C_{in} = C_{gs} + C_{gd} \qquad | \qquad C_{out} = C_{gd} + C_{ds} \tag{3.4}$$

Equation 3.4 provides only a crude estimation, thereafter it should be used with care. Reading the values from DC simulations leads to C_{in} =2.1 fF and C_{out} =1.2 fF, for a 4 fin transistor (set to technology characterisation size). The optimum sizing factor is iteratively determined to be 3 using Equation 3.5, where $\eta = \frac{C_{in}}{C_{out}}$.

$$\alpha = e^{\left(1+\frac{1}{\eta\alpha}\right)} \tag{3.5}$$

Further routing and prior circuitry (inverter chain for drive strength limitations) leads to a tenfold increase in capacitance at the input of TX driver - the total input capacitance is estimated to be 21 fF. As the delay is not the main concern of the design at the current project stage, it is decided to set the α factor to the range of 5-6, being slightly outside the optimal zone [21]. High α ensures a rapid current increase per-stage vs. area usage, allowing to reduce number of transistors used and thus, potentially reducing overall area (less clearance - lower unused silicon). Prior strategy ensures that optimal cascaded stage count of 3 is retained. Note, the number of cascaded devices is unchanged even if the size of the first inverter is altered (changing C_{in}).

As next, the sizing of inverters is adapted to match the slew rate and amplitude requirements elaborated in Subsection 3.2.1. The inverters are interchangeably up-scaled/down-scaled in size depending on which requirement has to be satisfied. To satisfy RX input swing requirement, the last inverter is boosted to a larger than α =6 factor ($\frac{W_{inv_3}}{W_{inv_2}}$ ≈10 for DR$\in [5, 8]$). Then, to improve the slew rate and also hold time, 2nd and 1st inverters are adapted consequentially, equalizing the charging strength distribution. Further elaborating: each consequent stage sees a unevenly increasing load, thus, leading to sizing structure as $\frac{W_{inv_2}}{W_{inv_1}} \approx \alpha$ and $\frac{W_{inv_3}}{W_{inv_2}} \approx 2\alpha$, where $\alpha \in [3, 4]$, with the exception of low DR designs.

Note, the size of the last inverter has to be chosen such that it can support instantaneous current of $\frac{0.375\ [V]}{160\ [\Omega]}$ = 2.35 [mA] and thus an average current of 1.18 mA for NFET or PFET with duty cycle of

50%. Note, for a full inverter, the average current is going to be close to 2.35 mA if switching times are neglected. Assuming that gate voltage of TX drivers for all topologies is exactly the same, expected design area is estimated in terms of average current of last TX driver stage. Note, this assumption holds only to the first degree as short channel effects and triode region currents indicate a direct dependence on V_{ds} voltage. Thereby, average currents have to be assumed to be characterised by saturation region equations (see Equation 2.4). Notice, leakage currents are excluded from the power analysis due to uncertainty of exact chip area - off-state current generation per μm in devices used is known, and is deemed small for a single TX driver. For a complete chip, power dissipation due to leakage would accumulate to a considerable contribution and thus would have to be accounted for when developing power delivery network.

Alternatively, in case of non-divergent V_{cm} designs, r_o of the devices can be compared, as it is directly proportional to active device area for triode biased transistors. Note, however, FET resistance values are closer to being constant in saturation region, thus, r_o comparison has its limitations when comparing saturated vs. triode biased transistors [27].

Summary: power dissipation is 6.25 mW, r_o is 160 Ω and average last stage current is 2.35 mA.

3.3.2. Sizing and Power Analysis of SFD

Similarly as for CTT, SFD power analysis defined in Subsection 2.2.1 is applied here. With the average eye mask opening of 250 mV, the current passing through the system is equivalent to 5 mA, assuming single 50 Ω termination. Thereafter, with transistor in on-state and V_{dd} =1 V average power consumption is determined to be 2.5 mW (using Equation 2.5), of which 25% is dissipated across the resistor. r_o value cannot be determined for the particular case as saturated driver is used.. For cases the transistor is biased close to transition region due to low V_{th} value of 200 mV, the equivalent NFET resistance can be determined to be $\frac{750\ [mV]}{5\ [mA]} = 150\ \Omega$. Note, the power consumption is likely to significantly increase for high speed designs to compensate for low falling edge slew rates as briefly mentioned in Subsection 2.2.1.

To reduce the power consumption of the design, the termination voltage can be reduced. However, this comes with the consequence of increased transmitter area as current sinking abilities are diminished due lowered drain-source voltage. Moreover, the signal would become more susceptible to noise due the reduced rail voltages and thus currents in the receiver. Also, the RX would have to be increased in size due to the newly reduced V_{gs} at its input. In case V_+ would be reduced to 0.625 V to ensure V_{cm} of 0.5 V, dynamic power dissipation would become 1.56 mW on average.

The area value can be estimated from the average current flowing through the NFET device. With the on-state current reaching 5 mA, implying 2.5 mA average driver current, the area of SFD can be observed to be approximately equal to CTT last inverter size (NFET + PFET) to first degree. Notice, the r_o value of SFD is equivalent to a single NFET device (150 Ω), however, as V_{cm} value significantly differs between CTT and SFD, r_o cannot be used for comparison. In case V_{cm} is set to $\frac{1}{2}$ of nominal V_{dd} value, r_o becomes equivalent to $\frac{375\ [mV]}{5\ [mA]} = 75\ \Omega$, which corresponds to the size of CTT inverter. When paralleled, 150 Ω would result in 75 Ω. Note SFD would also enter linear region, implying that its base operations would be altered.

In case a current mirror is used at the tail, the area can be observed to be at least double that of CTT. To keep tail source barely saturated, its V_{gs} has to be set only slightly above V_{th} value - the quadratic V_{gs} relation to I_D has to be compensated by making mirror area large. Thereby, area-power product of SFD would become higher than that of CTT.

Even though area-power product of SFD without tail CMs is lower than that of CTT, it does not come with several drawbacks. First, the noise of power supply is directly projected at the input of RX, as during pre-charge stage, TERM is in a floating state. Moreover, the design is prone to incoherent data transmission, caused by the lagging slew rates - the transistor has to continuously fight the termination. Additionally, in case a fixed reference voltage of $\frac{1}{2}V_{dd}$ is provided to RX without possibility to change it, RX might exhibit lack of drive strength on one of the transitions due to incompatibility between signal V_{cm} and reference. Lastly, since TERM resistance value and the output impedance of TX driver substantially differ in value, high reflections could be caused on TL. Hence, a parallel source termination might be required to mitigate reflection caused signal degeneration. The price of an additional resistance would be almost doubled power consumption and consequently doubled active area. Then r_o would become 75 Ω large, while effective parallel termination would result in approximately 25 Ω - 50 Ω lower gap

between transistor and TERM impedance.

Nominally SFD V_{cm} is close to supply voltage, meaning that RX branch current difference will be larger due to operations with higher applied V_{gs}. Thus the signal opening of SFD RX output is going to be wider compared to CTT output for iso-area conditions. The improved amplification of the receiver implies that voltage swing at the RX input could be reduced to lower than 250 mV leading to a decrease in power consumption. However, to be able to benchmark topology performance against a common metric the minimum compliance eye mask is kept to be 250 mV as stated previously. Note, casocded SFD topology is not strictly required, nevertheless, CMs are suggested. Single NFET SFD is not only vulnerable to PVT, but also more susceptible to crosstalk due to reduced intrinsic capacitances and opaqueness at low input voltage - current changes would directly affect power supply voltage.

Summary: power dissipation is 2.5 mW, r_o is 150 Ω and average last stage current is 2.5 mA.

3.3.3. Sizing and Power Analysis of HSTL

An example of HSTL topology is the so called LTT as mentioned in Subsection 2.2.1. Comparing to CTT, LTT achieves a power saving of ≈50%, which is achieved by almost completely eliminating static power dissipation - pull-up TERM is commonly removed from the design. Thereafter, signal swing is usually centered close to ground voltage, making LTT an almost inverted version of SFD. Here, however, the PDN network is not removed leading to full transparency during input signal switch and a very strong '0', if both NFETs are of the same size [29]. Bear in mind, power reduction comes at the cost of lower signal integrity and noise margin limiting topology's load bearing capabilities [10]. Also, lower V_{cm} voltage means that RX area has to be increased to compensate gate voltage drop at its input transistors.

While total power of LTT design is only a $\frac{1}{2}$ of CTT, peak current for LTT can be larger - it is strongly dependent on selected TERM value. If termination is selected to be Z_0, 5 mA of peak current are required to generate 250 mV swing, similarly as for SFD. Here additional current is dissipated by PDN, however, this value is relatively small. Thus, overall LTT TX driver can be said to be slightly larger than CTT last stage inverter. The ratio is usually less than 2 because PDN is commonly small. Moreover, NFET device leads to an inherent voltage drop across PUN, implying that PUN size has to be increased as full swing is not used for current generation (reduced V_{ds}), making LTT design larger than CTT even in FinFET technology.

Notice, PDN provides only partial transparency as it is operational during transition from high-to-low state after which it conducts only leakage current coming from the PUN. Thereafter, LTT shares majority of the drawbacks associated with SFD. Nonetheless, not using pull-up TERM rectifies power supply noise manifestation at the output continuously due to lack of direct connection between V_{dd} and V_{out}. Thus, LTT is less prone to power supply variations compared to SFD.

Summary: power dissipation is 2.5+ mW, peak r_o is 150 Ω and average last stage current is 2.5+ mA.

3.3.4. Choice of a Single Ended Signalling Topology

The most obvious choice of topology for which exploration is carried on is the system providing lowest area-power product. However, as accuracy of area determination is deemed low due to uncertainty of exact signal response to disturbances as noise, crosstalk and PVT, best all-around performing topology is selected. First, it has to be noticed that TX driver load is equivalent to TL intrinsic capacitance which is relatively high comparing to conventional test standards [13], implying that LTT topology can be eliminated from the analysis. Further, SFD topology is more beneficial than CTT only in ideal conditions when power supply noise and crosstalk are not present. SFD's inferior pull-down slew rate will significantly suffer upon disturbances making necessity for larger area. In such a case, SFD loses its appeal. All in all, it is decided to use CTT topology for further analysis of SES topology applicable for NAND I/O due its strong and symmetric slew rates and higher load bearing capabilities compared to other topologies.

It has to be noted, that in spite of higher signal degradation at 8 Gb/s data rate, SES driver can still provide sufficient quality signal to RX input (e.g. properly interpretable). Nevertheless, signal quality comes at the price of increased area and power consumption, as larger swing is required to counteract the crosstalk and reflection caused negative effects on signal's slew rate and amplitude. Thereafter, even when analysing point-to-point transmission without extra circuitry, single ended topologies are not going to provide a factor of 2 reduction in area with respect to differential topologies, making DS highly

appealing for investigation. Hence, in the next sections SLVS and CML power and size analysis is performed.

3.3.5. Sizing and Power Analysis of SLVS and LVDS

First, a choice has to be made on whether SLVS or LVDS is going to be used for further investigation of the design. As the main goal of the project is to define a system able to reach DR of 8 Gb/s at typical conditions, PVT is not a significant concern at the current stage. Thereafter, it is assumed that if PVT variations would be present, SLVS design results in a smaller chip area even after defect compensation by directly increasing input transistor size. The assumption can be said to hold as input transistors are always operating in triode region regardless whether current sources of LVDS are present, meaning, that PVT immunity of LVDS is not crucially better than that of SLVS, if CMs are barely saturated. Thereafter, it is decided, that SLVS is more applicable for NAND I/O interface if area minimization is the secondary goal in mind.

Thereafter, with a 250 mV voltage drop on TERM resistance, circuit current flow can be estimated by $\frac{\Delta V}{2Z_0}$, resulting in 2.5 mA in case 100 Ω is used. Thus, power can be directly computed as current voltage supply product shown in Equation 2.7, resulting in 2.5 mW consumption [99]. The power slightly increases due to alternative current path as the data rate increases - recurrent transitions with non-ideal slew rate cause all transistors to be operational simultaneously during a fraction of rise and fall of input voltage.

Note, to slightly reduce TX power consumption or better yet enhance V_{out} swing, TERM resistance can be increased for iso-area conditions. Larger TERM value is proportional to I_D reduction for constant output voltage swing ($\frac{\Delta V}{2Z_0}$). As H-bridge transistors mainly operate in linear region, where generated current is linearly dependent on V_{ds}, swing cannot remain constant. Thereafter, as current is reduced also V_{ds} drops, increasing voltage swing across the resistance. This, however, increases I_D - contradictory behaviour is observed. From the former, one can derive that the circuit will settle to a stable optimum for which the overall current reduces and the output swing slightly increases. There are limitations to TERM increase as at a certain point severe reflections are caused which lead to diminishing returns.

Also, larger than 100 Ω TERM value allows to diminish crosstalk induced effects. As odd mode signalling generated undershoot is the limiting factor in signal quality requirements, increased swing and overall impedance can reduce these disturbance effects. For instance, when comparing RX input eye of conventional SLVS topology with 100 Ω (see Figure 3.9) to 135 Ω (see Figure 3.10) TERM case, one can notice a cleaner - less undershoots - and more open V_{out} eye in the latter scenario. The same pseudo-random bit stream has been used in bth cases, thus, system is subjected to exactly the same conditions apart from TERM value.

There is a drawback of changing the resistance, however. When comparing the images (Figure 3.9 vs. Figure 3.10) more jitter and higher spread in output voltage levels can be observed. The growth of both jitter and V_{out} inconsistency is caused by increased reflections and higher current variance in combination to altered RC characteristics. For longer simulation times, such variations can generate higher levels of ISI due partial charge propagation from one stage to another as two consecutive signals can see larger voltage difference. Moreover, as devices are pushed into a deeper triode operating region to provide the wider opening, higher sensitivity to PVT could be observed.

Furthermore, as TERM is a floating resistance between two circuit nodes, it does not allow shifting of common mode voltage levels as is the case for CTT. Thereafter, in case there is a small deviation in effective current generation of NFET vs. PFET, PUN and PDN have to be sized differently. Else, deviations in drive strengths of '0' and '1' are going to be present, with V_{cm} skewing towards the most pronounced rail. The former is especially true for high DR designs, where size of TX devices is large, implying a higher current deviation between PUN and PDN, and thus higher V_{cm} shift. Ideally, the common mode of SLVS should be centered at $\frac{1}{2}V_{dd}$ to ensure symmetric output characteristics. Luckily, the effective current ratio between NFET and PFET devices stays roughly constant w.r.t majority of changes in external and internal environment, thereafter, simplifying centering of V_{cm} across different conditions.

Lastly, the approximate chip area of the SLVS has to be estimated. As differential topology has been used, meaning that number of active components per line has almost doubled, the first order area can be estimated to be double CTT area. This is confirmed by driver average current being the same as CTT, but the number of active transistors per line being doubled. Nevertheless, as SLVS is

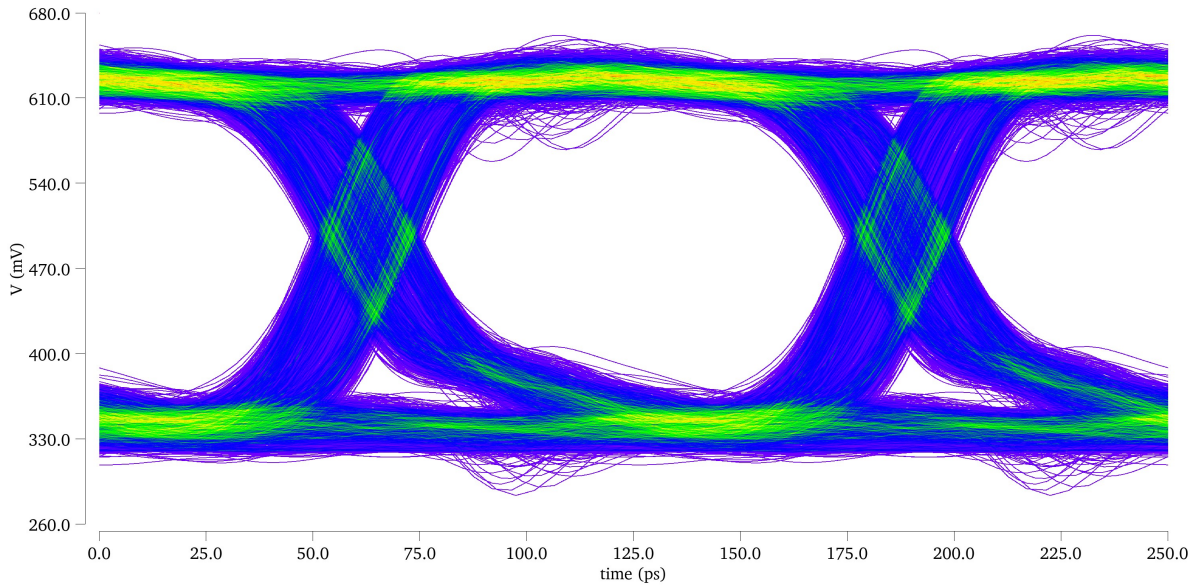


Figure 3.9: SLVS RX input eye for quasi-matched termination of 100 Ω. Simulation conditions used: Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

more immune to noise and EMI caused degeneration, at high DR the ratio between differential and single ended topologies is expected to be lower than 2. With SLVS area and power estimated, the other commonly used high speed differential topology CML can be investigated. SLVS area can also be seen to be twice that of CTT by comparing effective r_o. Single transistors output resistance can be estimated to be $\frac{375\ mV}{2.5\ mA} = 150\ [\Omega]$ for a centered swing, implying that 1 of 2 simultaneously active transistors is equivalent to half inverter size. Thereafter, half of the H-bridge amounts to exactly the inverter area of CTT in low DR conditions.

Summary: power dissipation is 2.5 mW, r_o is 150 Ω taken twice and average last stage current is 2.5 mA.

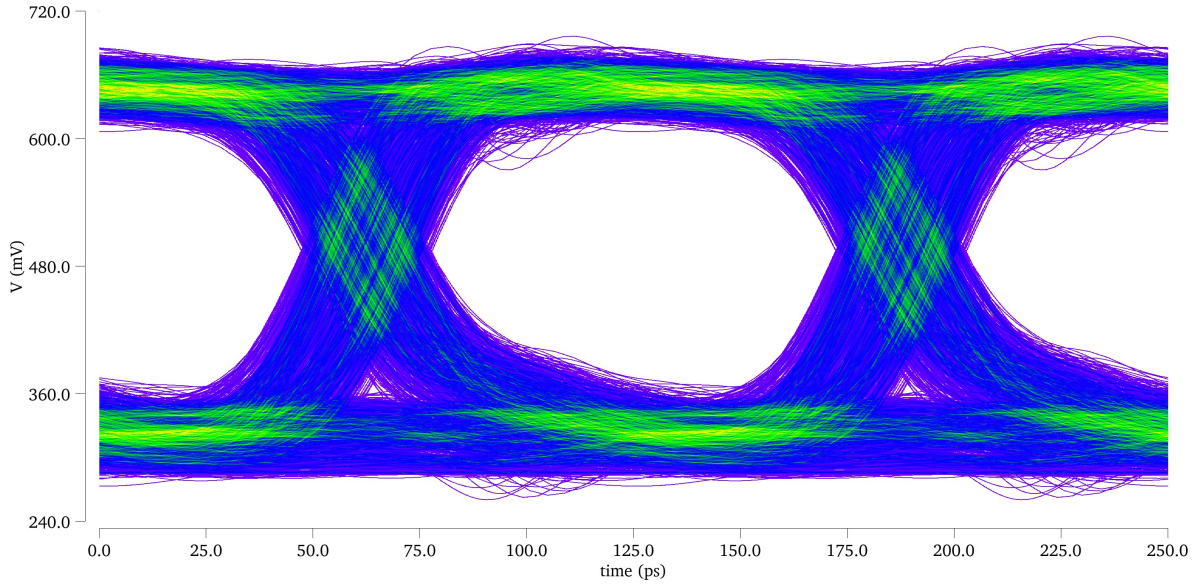


Figure 3.10: SLVS RX input eye for non-matched termination of 135 Ω. Simulation conditions used: Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

3.3.6. Sizing and Power Analysis of CML

CML is effectively a differential edition of SFD, thus both its power and area can be estimated to be double that of single ended version. With 250 mV swing and TERM of 50 Ω, CML average current can be seen to be approximately 5 mA. As one of the two branches is going to be on at any point in time, the duty factor is equal to 100%, causing continuous power consumption of 5 mW for 1 V power supply. Bear in mind, to improve signal quality, a parallel source termination can be added. However, the power consumption would almost double similarly as for SFD.

The common mode voltage level is equivalent to $V_{dd} - \frac{V_{swing}}{2}$, leading to increased RX currents and thus higher quality output, permitting to reduce differential swing. Nevertheless, with benchmark set to 250 mV, no alteration to differential operations of CML is made - analysing the optimal swing is left as a recommendation for future research.

CML input transistor covered area is ought to be similar to SLVS spanned area. Here, the average current is double that of SLVS, but the amount of devices is halved, making CML differential pair approximately equal to H-bridge total dimensions. Both input transistors operate in saturation region with the swing requirement defined in Subsection 3.2.1, meaning CML is able to sink more current per unit area than SLVS transistors can source/sink. Hence, the devices are going to be faster, but not necessarily smaller. Similarly as for SFD, r_o is approximately equal to 150 Ω when saturated and 75 Ω when swing has been reduced to center V_{cm} at half V_{dd}

Similarly as for SFD, the tail transistor has to be large, as it barely operates in saturation region. Moreover, the tail has to be able to sink currents of up to two times input transistor I_D, thus current mirror area is at least as large as cumulative differential pair size. Thereafter, overall CML area can exceed SLVS chip area even up to two times for moderate DR designs.

Summary: power dissipation is 5 mW, r_o is 150 Ω when saturated, 75 Ω when V_{cm} is set to $\frac{1}{2}V_{dd}$ and average last stage current is 5 mA.

3.3.7. Choice of a Differential Signalling Topology

From the analysis above, one can clearly see that SLVS provides both lower area and power consumption that CML for the same V_{out} swing. Thus, SLVS is selected for further investigation of the I/O topologies. Bear in mind, CML can reach speeds beyond those of SLVS due to its superior current sinking abilities. Thereafter, for extremely high speed applications, CML topology is preferred [22]. With both SES and DS selected, final RX topology compatible to both topologies can be chosen.

3.4. Choice of Receiver Circuitry

As briefly stated in Chapter 2, processing unit and memory are assumed to be separate entities, inferring that for a point-to-point interface, TX will be located on one chip, while RX will be embedded on the other IC. Typically memory and PU are manufactured by two different companies, meaning that, to-and-fro transmission is not strictly symmetric as long as signalling topology and termination conventions are met. Thereafter, rather than focusing on mimicking outsourced RX circuitry in full detail, one has to make sure that RX input signal quality standards are satisfied instead as discussed in Subsection 3.2.1.

With former in mind, not much attention is given to designing a PU receiver - the loading conditions of a real system are merely mimicked by using a placeholder circuitry. As noted in Subsection 3.3.4 the intrinsic input capacitance of RX can be effectively neglected in loading analysis, further solidifying the lowered focus on RX design. Using a topology suitable for both single ended and differential signalling is preferred, such that analysis and simulation implementation can be simplified to a single design. As readout of the system would be performed from a single ended line, it is decided to use a differential-to-single ended topology as the first stage of the amplifier. Note, the 2nd stage is of no interest in this thesis, since it does not affect the RX input node characteristics in a notable way.

To perform differential-to-single ended conversion, OTA is deemed sufficient (refer to Section 2.3). To increase RX gain, OTA would be followed by chain of inverters or a common drain/gate (depends on output readout type) topology, if either '1' or '0' has to be strengthened. Moreover, using inverters as final stages of the amplifier partially digitizes the signal by increasing slew rate and providing rail-to-rail swing. Thus, if no error correction or signal delay adjustment is required, the output signal can theoretically be sampled right after inverter chain.

Note, to improve signal quality a differential pre-amplifier can be used for DS. Increase in RX input capacitance by using pre-amplifier rather than OTA is negligible due their comparable input transistor

sizes - pre-amplifier can be omitted from PU's RX without any decline in accuracy of analysis.

To obtain a sufficient quality signal at RX output, sizing of OTA has to be performed such that low output impedance and high enough gain is achieved at target frequency to satisfy slew rate and speed conditions. To meet the gain requirement, OTA input transistor size is set to be large to provide an increase in amplifier's transconductance. g_m has a direct proportionality relation to OTA gain as seen in approximate gain equation of size-symmetric OTA (Equation 3.6). Nevertheless, as increase in size also reduces equivalent output resistance, amplifier gain is limited by technology's transistor intrinsic gain - both r_o and g_m are drain current dependant. Thereafter, to achieve the highest gain possible PUN and PDN sizes have to be balanced to yield optimum equivalent output resistance as purely changing size of a single device does not affect the gain value.

$$A = g_m r_o \tag{3.6}$$

To make the design more realistically loaded, it is decided to add a parasitic capacitive contribution right after the OTA (100 fF) partially mimicking dense wide-trace multi-layer routing and ESD protection circuitry. With the parasitics installed, the signal gets significantly degraded due to lack of current, hence limiting OTA attainable speed. As output capacitance of the OTA is negligible compared to 100 fF value, the response quality can be improved by simply increasing the size of output adjacent transistors (Reduces $\tau \propto$ RC). Note, RX area has to increase significantly to provide sufficient peak current to the load so that speed and slew requirements are met.

By knowing the total current provided to the load after imposition of artificial parasitic load, RX power consumption can be determined as average tail current multiplied by power supply voltage, assuming that current through tail is almost constant. As RX circuitry used is only mimicking a loading condition, the power consumption is determined from simulations without any prior theoretical estimations. The power consumption of RX on average is determined to be 1-1.5 mW.

To reduce power consumption and increase current stability of RX, cascoded CMs can be used in the tail of the OTA [100]. By performing a simulation for nominal design of CTT, use of cascoded current mirror results in RX power reduction being halved. However, the overall area of RX increases by approximately 40% (barely saturated transistor) following close to usual area vs. power trade-off. The TX and RX input did not undergo any noticeable changes. Moreover, the dynamic swing of RX is negatively affected with each added current mirror cascode - CMs saturation requirements are fixed to a certain voltage, e.g. V_{ds} is always 50 mV larger than $V_{gs} - V_{th}$.

RX input signal quality requirements on memory side are more strict (e.g. higher slew rate, wider opening) implying that memory RX can be weaker compared to PU RX [13]. Thus OTA followed by inverter chain is deemed to be sufficient as memory RX for this project. Note depending on how ***write*** action in memory is executed, level shifting might be required to provide high voltage necessary to perform current push-through in the thick oxide memory devices [24]. The technology type of logic vs. memory devices is different irrespective of whether both logic and memory are manufactured on different wafers and then bonded/connected or on the same wafer in CMOS under array structure [8].

RX output quality requirements are undefined because post RX circuitry is absent from the analysis. Thus, it is impossible to determine exact properties RX output signal must exhibit for the next stage to perform operations properly. Here only the scenario in which RX would be sized is provided.

3.5. Termination, Transmission Line and Packaging Sizing and Design

With both SES and DS transmission topologies chosen and RX circuitry settled upon, design of passive system components as TL, TERM and PACK has to be defined and finalized. In the following subsections limitations of each component and their final design used in this thesis are elaborated on.

3.5.1. Transmission Line Limitations, Design and Implementation in Simulations

In order to prevent major design limitations from being imposed on PCB circuit designers, NAND I/O interface has to be designed assuming close to worst case external conditions of the system. Such approach guarantees a complete freedom in the choice of routing trace style (microstrip, co-planar waveguide, stripline) and layout.

Nevertheless, certain limitations have to be set to assure design feasibility is not compromised. For

instance, it was decided (see Section 3.1) that memory chip and PU are located a finite distance from one another, resulting in TL length of 40 mm (straight line I/O pin to side distance being ≈6 mm at current standards [13]), to make certain that chip's logic area is not increased beyond a necessary margin.

To determine what type of TL imposes the most devastating effect on TX circuitry, dominant loss mechanisms have to be identified and analysed. With increase in TX area, it should be possible to overcome challenges laid by TL to a certain degree, such that sufficient quality signal at RX input is provided. Thereafter, a trade-off between various possible TL structures has to be established, where signal sensitivity to trace type selection is the key parameter.

The two main loss mechanisms present across the length of TL are insertion loss and crosstalk. The insertion loss is attenuation of signal along the line caused by factors as non-ideal components, radiation and reflection, to name a few [101]. The non-ideal components cause power dissipation due to finite values of realistic materials. For instance, conductors with limited conductivity lead to series resistance, thus energy lost in form of heat, while dielectric impurities and imperfections lead to parallel conductance responsible for current leakage from signal to return path.

Radiation loss is generated by having a weak coupling between signal and return (GND) paths, leading to electromagnetic energy being transferred to other conductive material in the surrounding environment [102]. Radiation loss is the main insertion loss mechanism if thick dielectric separation between conductors is used and air is employed as top dielectric - low dielectric constant leads to larger fringing fields.

Lastly, the reflections, in majority of the cases being most dominant insertion loss mechanism, are caused by discontinuities on the TL. For example, having TERM, which impedance does not match that of the TL leads to signal reflections being generated at the component interface. The back-propagating signal interferes with the incoming signal, causing an amplitude variation and thus reducing signal quality [44].

Total insertion loss factor varies largely on TL dimension and type - for microstrip line, the dominant loss effect is usually radiation or conductor, while for a stripline radiation losses are negligible. In case a very small cross-section stripline is used, its insertion losses at high frequencies are going to be generally higher than those of microstrip line due to immense conductor losses caused by both skin effect (t remains the same) and high DC resistance. [103]

The crosstalk has been separated from the insertion loss, as it is caused by coupled neighbouring lines rather than the line itself. Any two adjacent signal lines are going to be electromagnetically coupled. However, if high frequency signals are transferred along two neighboring lines, signal fluctuations can cause a rise or fall of instantaneous current on the line. As voltage response is delayed with respect to current change, variation in characteristic impedance can be observed [64]. Thereafter, crosstalk not only causes under- and overshoot of voltage response, but it also temporarily increases reflections on the line due to additional miss-match in the system.

Alternatively, change in impedance can be seen to be caused by line effective self-inductance and self-capacitance being altered as a result of mutual inductance and capacitance between the lines. The mutual parasitics are characterised by TL common electromagnetic field, leading to overall Z_0 reduction for odd mode coupling and increase for even mode coupling. The effects can be imagined using the following logic - in odd mode coupling the lines are more attracted to each other increasing the overall capacitance between them, but reducing loop inductance as one of the lines is seen as a partial return path (see Equation 2.50). The opposite is true for even mode coupling as lines get repelled from each other. The net effect on the characteristic impedance can be visualized using Equation 2.29.

Crosstalk is similar to radiation loss in its nature, thereafter, it is more detrimental for microstrip lines than striplines. With crosstalk becoming the most dominant signal degradation mechanism in high speed design [23], microstrip line is predicted to provide the most sever signal degradation. Thereafter, using microstrip line for I/O characterisation purposes allows complete freedom for PCB designers in choosing any TL type they desire. This is especially true if vias right under BGA pads are used going directly into intermediate PCB layers, completely bypassing use of edge planes [14]. Pin spread for conventional packages [13] allows for ground pin insertion in empty gaps for higher return path coupling, inferring a reduction of loop inductance.

The assumption above can be seen to hold when equal impedance simulations for a microstrip line in Figure 3.11 and stripline Figure 3.12 are compared. The signal eye at the output is significantly more

[14] URL `https://www.mclpcb.com/blog/vias-bga-pads/` [cited on 9th of August 2023]

clean and open in stripline case for iso-area conditions. Note, microstrip line also results in a larger PCB area as for $Z_0 = 50\ \Omega$ trace width has to be approximately double the dielectric separation between signal and return paths, while for stripline more than opposite ratio is true [50]. The only benefit of microstrip line over stripline is faster signal propagation speed across TL, since the effective dielectric constant is lower (see Equation 2.29). However, by adjusting sampling clock to have exactly the same delay as data, propagation delay miss-match is corrected, leading to proper sampling and/or dynamic amplification.

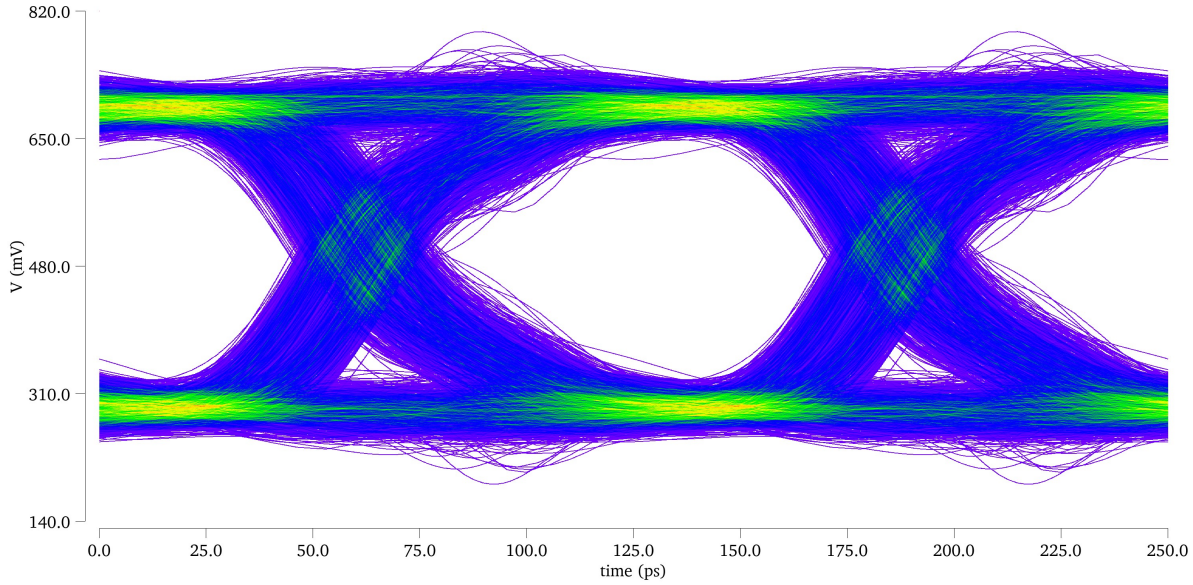


Figure 3.11: Signal eye of RX input for CTT topology when TL is of microstrip configuration. Simulation conditions used: Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

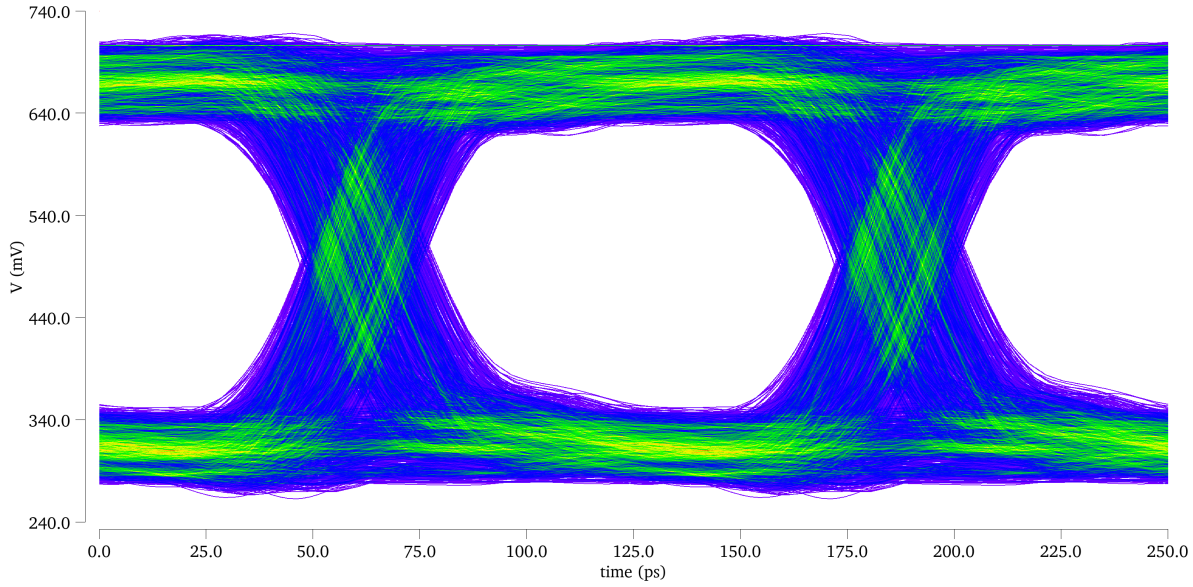


Figure 3.12: Signal eye of RX input for CTT topology when TL is of stripline configuration. Simulation conditions used: Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%,

To compute the characteristic impedance of TL, commercially available calculators implementing microstip formulas given in Equation 2.12 [15] and embedded microstrip formulas in Equation 2.17 [16] are

used. The former calculator also returns the value of ϵ_{eff} due the presence of 2 different dielectrics. Observe, embedded variation of microstrip line is used in the design as in-package traces are completely immersed into dielectric material.

The dimensions of TL are taken such that Z_0 = 50 Ω and manufacturing costs are minimized (bulk manufacturing - poolable option)[17]. Thereafter, to ensure minimum trace width and best signal-to-ground coupling, slightly larger than minimum dielectric thickness of commercial PCBs [18] is used for modelling TL in this thesis - dielectric thickness is set to 100 μm. From here, it can be concluded, that trace width has to be approximately 200 μm. The actual width has to be re-iterated when dielectric material has been selected.

Conventionally, spacing between two adjacent lines has to be at least the same size as trace width [19]. However, in this thesis, it is assumed that gaps between tracers are at lowest $\frac{3}{4}$ of width assuming large manufacturing tolerances. The latter also leads to most severe crosstalk permissible. Finally, the thickness of the TL is set to 35 μm, which correspond to conventional PCB trace height. Note, choosing smaller thickness than 35 μm causes manufacturing tolerances to grow rapidly - designing for 18 μm thick trace can lead to large deviations, even up to full trace thickness, resulting in 35 μm design [20].

When selecting a dielectric material, poolable option is preferred. Additionally, lowest loss tangent material should be selected to ensure that dielectric losses can be neglected. For this purpose, NP155F [21] is used, which have an $\epsilon_r \approx 4$ and $\tan\delta \approx 0.012$. With the aforementioned numbers, G term of TL is deemed negligible. Bear in mind, using NP155F requires copper trace width reduction of 20 μm to ensure $Z_0 = 50\ \Omega$ when *ncline* approximation is used, resulting in 180 μm actual width used.

Lastly, average copper roughness factor of 1.5 μm is applied, which is equivalent to copper trace acquired using liquid etching [54]. By plugging in all the aforementioned parameters into equations given in Section 2.4, following TL parameters are obtained: $R = 183\ \Omega$/m, $L = 315$ nH/m and $C = 107$ pF/m, resulting in $Z_0 = 54\ \Omega$. Note, 10% deviation from target impedance can be observed, which is caused by use of statistical relations as mentioned in Section 2.4.

When using Cadence® Virtuoso® 2D field solver utilizing finite element method with crosstalk incorporated, Z_0 of 50 Ω is acquired. Interestingly, majority of theoretical assumptions provided in theoretical computations overlap with Cadence® Virtuoso® implementation of TL. Thereafter, RLCG parameters undergo only a slight variation, resulting in $R = 145\ \Omega$/m, $L = 280$ nH/m and $C = 115$ pF/m, totaling into $Z_0 = 49\ \Omega$. Observe, the values above have to be converted to numbers per mm to ease expression of total TL RLC parameters. The mutual components between directly adjacent aggressor and victim lines are on average $R = 20\ \Omega$/m, $L = 50$ nH/m and $C = 10$ pF/m, and once removed lines - $R = 5\ \Omega$/m, $L = 15$ nH/m and $C = 1$ pF/m. From the above, one can conclude that in case guard lines are used, only the adjacent aggressor line influence has to be taken into account for Z_0 computations.

Neglecting any aggressor line past directly neighbouring signal line, characteristic impedance of victim line for worst case odd mode coupling is 37 Ω, while worst even mode coupling $Z_0 = 63\ \Omega$. Total resistance of TL can be noticed to be small (6 Ω) and thus, insertion loss is deemed negligible in comparison to crosstalk caused amplitude changes. With TL parasitics established, PACK contributions to signal degradation have to be computed.

3.5.2. Packaging Limitations and Design Considerations

This thesis project is performed exclusively in simulation environment - no tape-out and testing is performed. Thereafter, FC packaging is a straightforward choice for final design analysis.

Notice, only 2 components of packaging are unknown - BGA parasitics are defined in Subsection 2.5.1. For trace parameters, the same analysis as provided in Subsection 3.5.1 is used, with only

[15] URL https://chemandy.com/calculators/microstrip-transmission-line-calculator-hartley27.htm [cited on 9th of December 2022]

[16] URL https://www.allaboutcircuits.com/tools/embedded-microstrip-impedance-calculator/ [cited on 9th of December 2022]

[17] URL https://www.eurocircuits.com/blog/pcb-services-based-on-material-choice/ [cited on 5th of December 2022]

[18] URL https://www.protoexpress.com/blog/ipc-class-2-vs-class-3-different-design-rules/ [cited on 9th of August 2023]

[19] URL https://www.eurocircuits.com/pcb-design-guidelines/classification/ [cited on 5th of December 2022]

[20] URL https://www.eurocircuits.com/blog/tolerances-on-copper-thickness-on-a-pcb/ [cited on 5th of December 2022]

[21] URL http://www.eurocircuits.com/wp-content/uploads/ec2015/ecImage/document/NP155F-D-5-2012.jpg [cited on 9th of August 2023]

change being formula set adaptation to EMS line. Package's dielectric filling is assumed to have the same properties as NP155F, conductor is assumed to have the same R_a of 1.5 μm and thickness of 35 μm. The only difference is trace width and dielectric height of the substrate. The width is chosen such, that 2 traces can fit between 2 adjacent external BGAs and all the gaps would have equal dimensions. Hence, internal package trace dimensions are chosen to be 90 μm, which corresponds to $\frac{1}{2}$ of TL width, making the equidistant gaps approximately 30-40 μm wide. For simplicity, dielectric height of package trace was assumed to be $\frac{1}{2}$ of PCB dielectric height, resulting in 50 μm and total height of 135 μm. The former approach provides Z_0 of 47 Ω ($R = 270$ Ω/m, $L = 300$ nH/m and $C = 140$ pF/m), which is assumed to be close enough to 50 Ω mark, thereafter, no further adaptations are performed.

The trace lengths are assumed to be 5 mm and 2 mm to ensure that internal connection can span a distance from package edge till 2nd center-most pin position [13]. Crosstalk is neglected for on-package traces to simplify the design. Thereby only distributed RLC model is used to approximate package trace, where 2D field-solver returned values given above are used as a reference. Equation 2.17 formula returns questionable results when immersion distance (total height) becomes large, thus, the empirical equation set was eliminated from the analysis.

Lastly, via parasitics have to be determined. Note, only 4 dimension parameters are required: h_{via}, r_{via}, r_p and r_{ap}. The dimensions could be directly extracted from PCB classification, assuming that the same via-trace ratios apply to both in-package traces and TL [18]. Thereafter, assuming vias with h_{via} of 135 μm, r_{via} of 15 μm, r_p of 50 μm and r_{ap} of 65 μm, R_{via} is 0.05 Ω, L_{via} is 38 pH and π-bridge capacitance C_{via} is 18 fF. Notice, parasitics of via and BGA are relatively low compared to internal traces or TL, thus they have almost negligible effect on overall system performance. Nevertheless, for full coverage of the system, both BGA and via parasitics are implemented.

3.5.3. Termination Limitations and Design Considerations

The most important consideration on TERM is whether to put it on chip or outside of it. As shown in Section 2.6, external TERM leads to severe reflections, which suffocate the signal and cause data reliability issues. Hence, first and foremost, only on-chip termination is assumed for all designs analysed. For DS the effect of external TERM is highly severe as can be seen when comparing simulations employing external (see Figure 3.13) vs. on-chip (see Figure 3.14) termination.

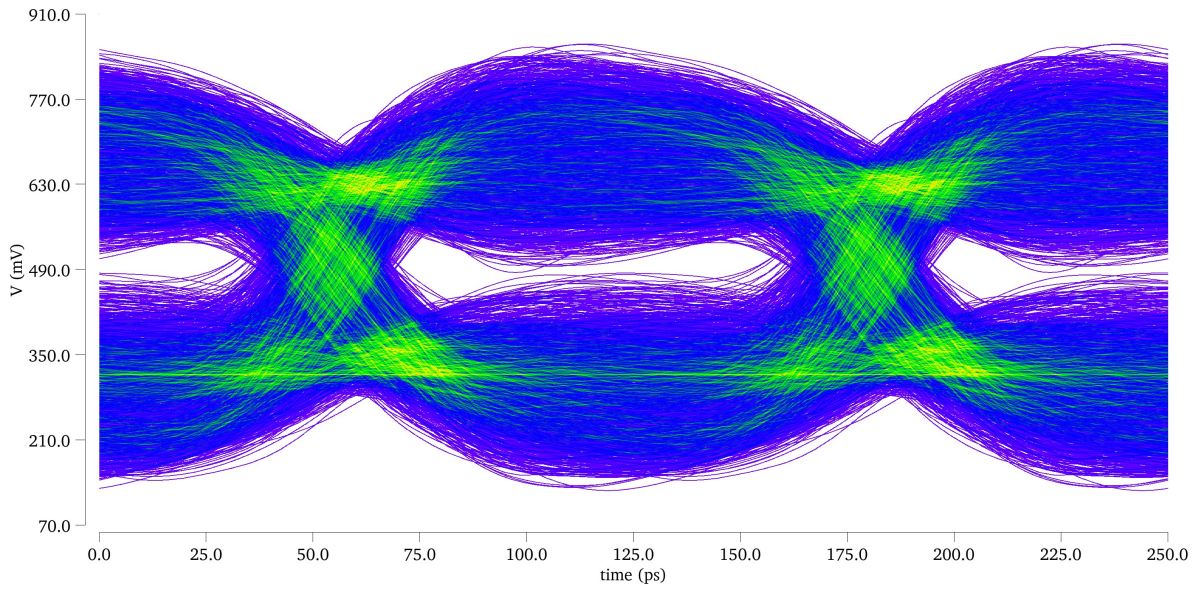


Figure 3.13: Example of transient signal eye when termination is located outside of chip. Simulation conditions (refer to Section 4.1): SLVS topology, Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

Next, trade-off between active and passive termination has to made. Observe, in full system implementation, active termination has to mimic operations of a passive termination with two main goals in mind - achieving high voltage-current linearity and placing TERM as close to RX amplifier as possible.

Passive resistances not always can be directly placed next to logic circuitry due to required clearances, implying that their actual impedance might deviate from the desired value due to parasitics of routing. However, as tape-out of a chip is not performed, exact design rule checks are not clearly known, hence passive termination can be assumed to be located infinitesimally close to RX active circuitry.

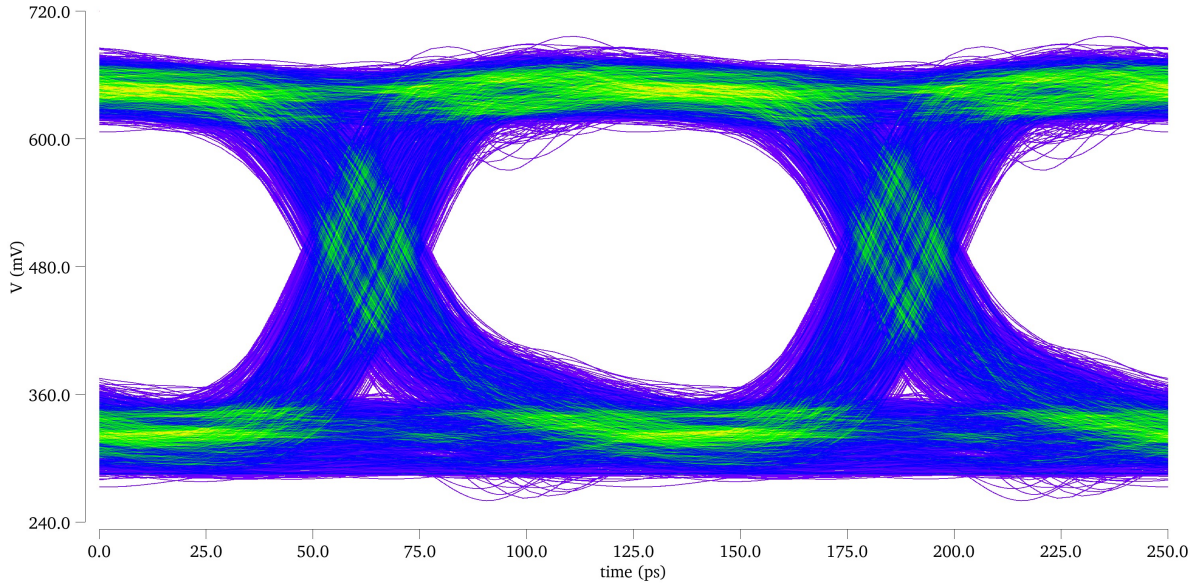


Figure 3.14: Example of transient signal eye when termination is located on-chip. Simulation conditions (refer to Section 4.1): SLVS topology, Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

Moreover, using a single symmetric active termination is likely not going to provide the desired linearity of TERM. Linearity of symmetric termination varies w.r.t V_{th} of the device - the lower the threshold voltage, the less difference is present between saturation and triode currents at $V_s = \frac{V_{dd}}{2}$. For a device with low V_{th}, boundary between linear region and saturation current has merged making one inverse of the other, generating highly linear voltage-current relation as is shown in Figure 3.15. Increasing V_{th} leads to lower linearity of symmetric TERM as can be seen in Figure 3.16.

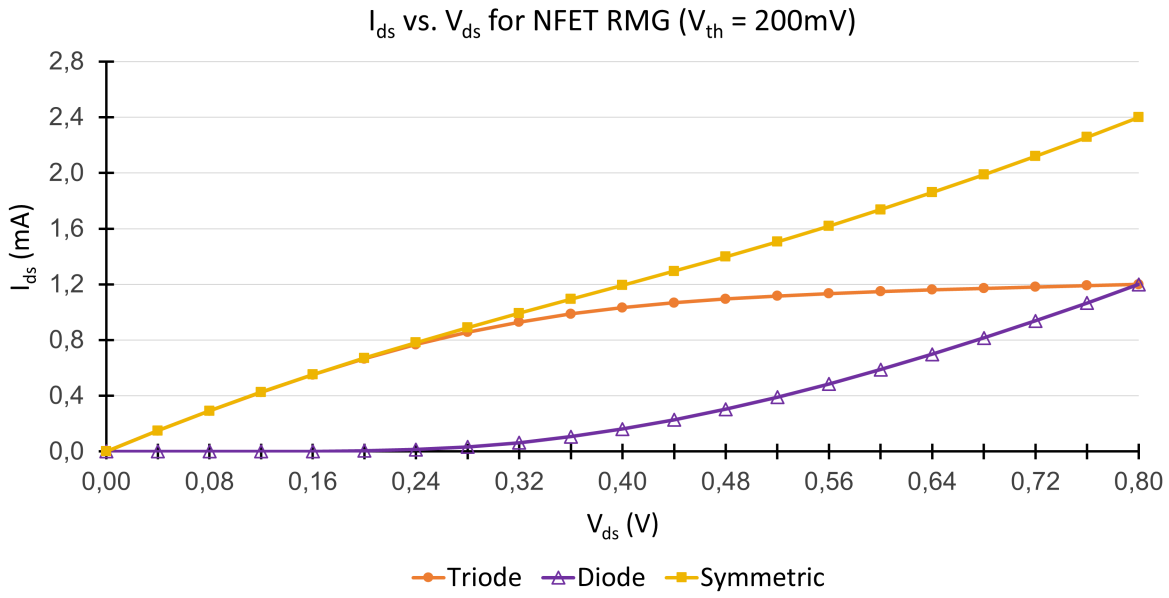


Figure 3.15: Total current of symmetric TERM for Imec in-house 14nm FinFET technology, V_{th} = 200 mV

From the figures, one can derive that low V_{th} devices are ideal for making linear active TERM. Nevertheless, power consumption of such configuration is going to be extremely high as both 'on' and 'off' currents of the device have been increased. For large number of data lines the power consumption would add up, leading to immense power budget, which could be infeasible to satisfy. On the other hand, having only adequate linearity requires controllable parallel impedances which generate the required impedance value continuously. Thereafter, implementation of active TERM can be concluded to be more time and resource consuming. With the thesis goal being not only design implementation but also sensitivity analysis, it is decided to use only passive termination for design exploration. Full scale investigation of suitable active TERM in FinFET technology is left as a future research topic.

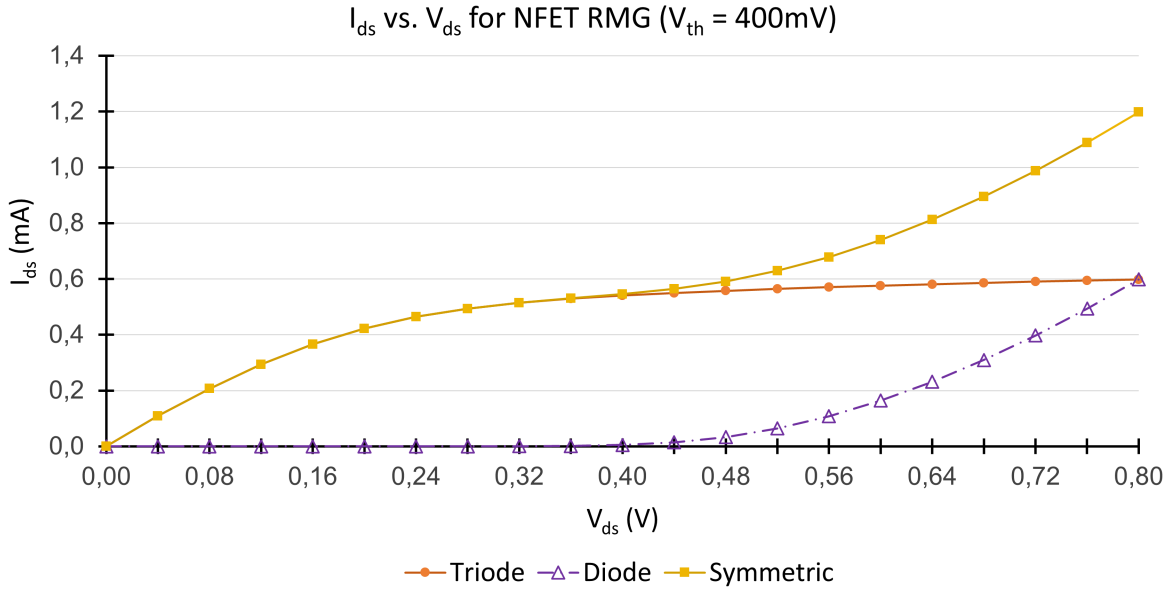


Figure 3.16: Total current of symmetric TERM for Imec in-house 14nm FinFET technology, V_{th} = 400 mV

Lastly, only the TERM showed in the topology schematics in Section 2.2 are implemented. Source termination is not needed per se, if driver has been designed with care, since TX last stage can behave as a matched resistance. With the former considerations in mind, the full system is ready for simulations and sizing for design case. However, before obtaining initial results, available device technology is discussed.

3.6. Discussion on Device Type Selection and Analysis

It was noted in Chapter 2 that in this thesis both RMG and GF devices are going to be explored for applications in NAND I/O circuitry. Here more elaboration on both devices is provided to understand current limitations of using FinFET devices in 3D NAND CuA structure. Notice, further provided material is based on both planar and FinFET devices and thus shall be used with caution. The net effects of various manufacturing process are largely the same whether planar or multi-gate device is used, as HK/MG GF and RMG device characteristics are mainly defined by thermal budget applied - time and temperature of annealing.

The main differences between RMG and GF devices is their attainable V_{th} value and process complexity. In case of HK/MG GF, V_{th} is elevated due to the high thermal budget experienced by the gate stack - immense temperature applied for a long time can cause dielectric layer degradation [104]. Extreme thermal budget is impacting PFET V_{th} more severely than NFET, especially if thin effective oxide thickness (EOT) device is used, where gate oxide is susceptible to regrowth due to Silicon oxidation during thermal anneal [105]. Increased V_{th} of PFET can be prevented if Silicon-Germanium compound is used for channel formation, as it allows optimization of device bandgap by shifting valance band upwards [106]. Nevertheless, using Silicon-Germanium increases manufacturing process complexity, which is the main advantage of GF vs. RMG. Alternatively, hybrid GF and RMG device can be gener-

ated, where GF NFET and RMG PFET is used to avoid full RMG process complexity [107].

RMG manufacturing process is commonly more involved than GF, as additional chemical polishing steps are used due to inclusion of sacrificial dummy gate and dual metal deposition [108]. However, the complexity is outweighed by increased carrier mobility (thus, higher I_D) and improved device reliability - both achieved by having temperature unaltered gate stack [109]. The former implies, that RMG device integration in 3D structures as CuA would be challenging as logic cells would have to withstand memory's thermal budget. That is also the main reason there are currently no existing 3D NAND I/O circuitry based on RMG exhibiting thermal stability [16]. Thermally capable FinFET RMG device able to tolerate DRAM thermal budget has been presented recently [110], indicating that major breakthroughs in RMG thermal stability are only yet to come.

When considering device performance per unit area, RMG devices display better drive strength, whether those are planar devices [16] or FinFET devices [104] and their variations [111]. Thereafter, for intensive performance and power applications RMG device is a better choice, while low power use cases is more GF niche [19]. Thus, RMG would be perfectly suited for high-speed I/O driver located on memory side, where area is strictly limited and RX circuitry where significant amplification is required. On the other hand, GF would be ideal for TX circuitry on PU side where area restrictions are less severe, thus, allowing to cut on process complexity costs. With this in mind, target case results can be acquired and discussed as is done in Section 3.7.

3.7. Design Case Simulation Results

Combining all the defined design parameters and characteristics in the preceding sections, simulations for CTT and SLVS designs can be performed. For each design, 3 different FinFET device types - FinFET GF, RMG, GF extended - are going to be used, to determine the lowest area-power performance just meeting required signal quality. GF extended refers to device where conduction channel has been elongated further than the gate boundaries. As next, CTT and SLVS designs are also simulated with planar RMG technology, allowing to determine FinFET provided active area savings.

Observe, simulation conditions for executed schematic runs are provided in figure captions for eased overview of design variable changes (if any). In this section, the following simulation conditions have been used for all simulations: Imec low V_{th} FinFET 14 nm tech., V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temperature = 25°C, DR = 8Gb/s, input jitter = 16%. With this, active area footprint, dynamic power and total power are found as presented in Table 3.1.

Table 3.1: Results of I/O topologies for various devices reaching 8 Gb/s transmission speed

Topology	Device Type	TX Active Area Footprint (μm^2)	Dynamic TX Power (mW)	Total Design Power (mW)
CTT	RMG	0.31	2.3	7.9
	GF	0.44	2.5	8.0
	GF_ext	0.50	2.5	8.0
	RMG (Planar)	2.52	2.3	7.9
SLVS	RMG	0.39	3.1	4.3
	GF	0.51	3.2	4.4
	GF_ext	0.58	3.1	4.2
	RMG (Planar)	3.29	3.1	4.3

It can be seen that TX active area of CTT topology for RMG devices is by approximately 42%

and 61% lower than systems using GF and GF extended technology. The large difference in area can be attributed to reduced drive strength of GF and GF extended devices in comparison to RMG, when exactly the same dynamic current has been applied. To provide iso-performance conditions, the same signal swing has to be generated over similarly scaled TERM resistances, implying that current consumption of the designs has to be set almost exactly the same. Thereafter, total power consumption between designs of different devices is found to be approximately equal (see Table 3.1). The slight deviation is caused by the change in design sensitivity to crosstalk, which is attenuated/enhanced by intrinsic capacitance increase/decrease in case device type is changed and thus TX area altered. Self loading is almost negligible as intrinsic capacitance variation is minor relatively to load capacitance which consists of both PACK and TL capacitive contributions.

When looking at SLVS design, RMG design is not as superior over GF and GF extended designs, providing active area savings of 30%. The main cause for drop in RMG performance is caused by difference in effective current of PFET and NFET devices. Compared to GF, RMG sees a larger gap in current production between NFET and PFET, implying that PDN has to be increased in size to counteract higher PFET current generation. When input buffer is used, which inverts data signal, PUN of H-bridge has to be increased rather than PDN to balance and elevate V_{cm} at RX input. Centering of V_{cm} for CTT was performed using TERM, thereafter, only SLVS sees diminishing RMG superiority over GF designs.

Lastly, when comparing active areas of FinFET RMG vs. planar RMG designs, an immense difference - factor 8 - can be noticed. Note, only the active area footprint of TX is provided here, thereafter, the numbers are representative of the approximate ratio. In realistic design the difference in area would decline sharply with introduction of ESD, sampling and multiplexing circuits.

The overall gap between FinFET and planar devices is generated by multitude of factors. It has to be noted that $\frac{W}{L}$ ratio of TX circuitries between different technologies is similar. Thereafter, if L value is reduced by a certain factor, so is the width. Hence, a factor of 1.7 squared is caused by difference in channel length values between planar 45 nm and FinFET 14 nm technologies. Another factor of 1.5 is achieved by having almost equivalent PFET and NFET current characteristics for FinFETs - PFET vs. NFET currents differ by a factor of ≈2 in planar devices. Lastly, footprint area of FinFET benefits by a factor almost 2 over planar technology due to going 3D, where effective area is mainly extruded upwards. As fin pitch (distance between fins) is commonly ≈2 times smaller than fin height, active area is going to be approximately twice larger than footprint area. The exact enhancement of going 3D can be determined as $\frac{8.1}{1.7^2 \cdot 1.5} = 1.88$ for CTT topology and $\frac{8.4}{1.7^2 \cdot 1.5} = 1.95$ for SLVS topology. Even though the area downsizing seems infeasible, the actual footprint reduction is rational, when factors are analysed separately and superimposed on one another.

The main reason for the possibility of direct comparison is the shared V_{dd} value between planar and FinFET devices in this study. In reality, FinFET power would be 20% lower than that of 45 nm planar devices, while area would be closer to 4-5 times larger. Here, the area reduction would be proportional to squared reduction in current as per Equation 2.4.

When comparing acquired dynamic power values with theoretically estimated ones in Section 3.3 a deviation of 24% for SLVS and staggering 144% for CTT can be noticed. Also difference in total power can be observed, however, to determine exact variation, RX power has to be expressed first. OTA amplifier power amounts for approximately 1.1-1.2 mW, which can be directly determined from SLVS topology as $P_{tot} - P_{dyn}$. Thereafter, variation in total power of CTT topology can be determined to be only 7%, indicating that total power estimation strategy provided in Subsection 2.2.1 is accurate. The former implies that dynamic power estimations have to be revised since too ideal assumptions were taken. Such scenario was already predicted in Subsection 2.2.1, where dynamic power was said to hold if and only if TX r_o guesstimation is correct. Additionally, both SLVS and CTT P_{dyn} are increasing due to use of high frequency data signals - the current is directly forwarded towards TERM less frequently due to alternative current path from power supply to ground with both PUN and PDN being operational simultaneously. For DR closer to DC operating point, measured power value converges to theoretical estimations as showed in Section 4.2. For high DR designs RX input eye rails have a higher spread (see Section 4.2), implying that assumed average swing values are not entirely accurate.

Additionally to RX input results, signal quality just at the output of the TX was briefly investigated. Such exploration is required to understand what and how large system adaptations to current design are required to obtain a design, which can be made multidrop configuration compatible. Signal eye at the driver output for CTT topology (see Figure 3.17) can be seen to be almost sufficient to enable

multidrop operations as eye opening reaches 300 mV in almost the whole eye width. On the other hand SLVS topology (see Figure 3.18) sees higher one-sided source reflection, thus its TX input eye amplitude can be seen to lack, not even reaching 200 mV margin.

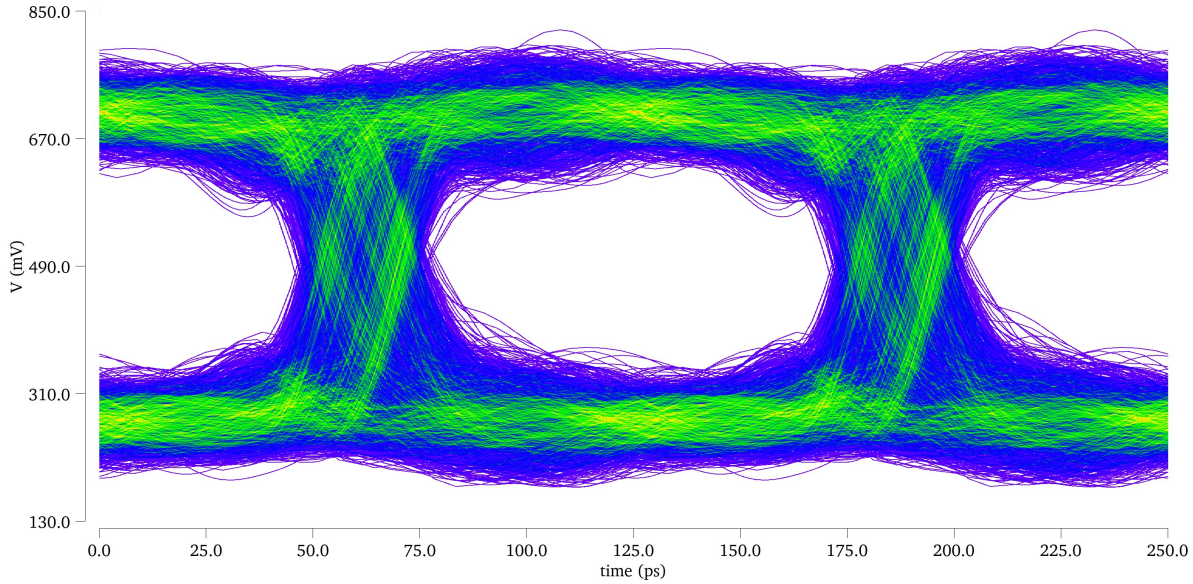


Figure 3.17: CTT example of signal eye at driver output. Simulation conditions (refer to Section 4.1): Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

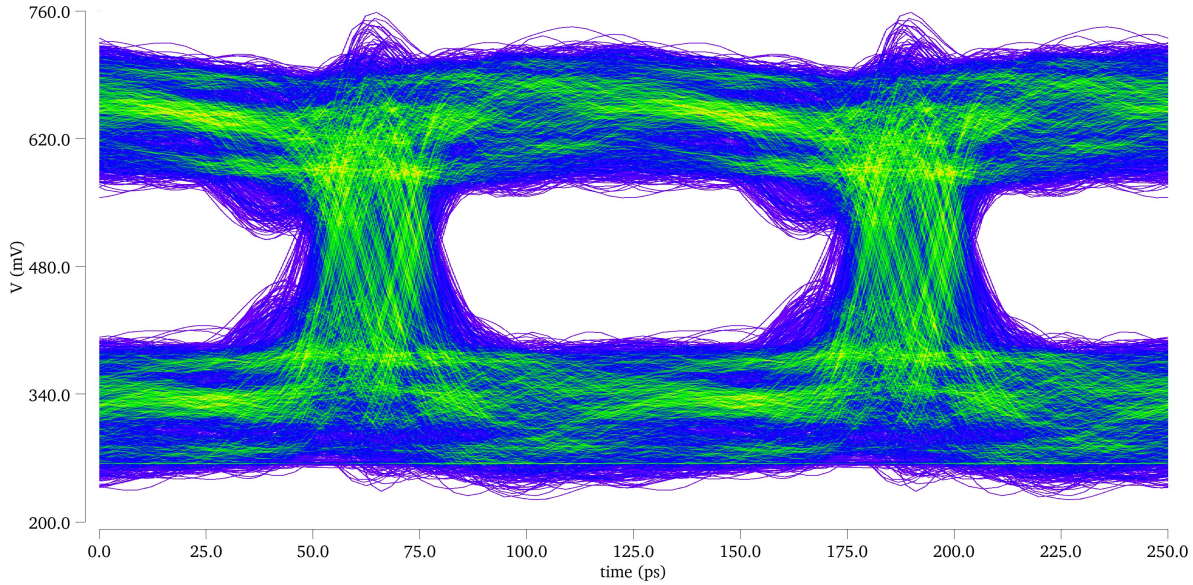


Figure 3.18: SLVS example of signal eye at driver output. Simulation conditions (refer to Section 4.1): Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} =1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

Lastly, one has to note that active area footprint is determined excluding dummy transistor area. As it is not know how compact the layout of the TX circuitry would be, exact influence of dummy transistors on the total area cannot be predicted. However, it is known that such transistors would be required to reduce channel damages caused by etching process of the gate [112]. Uniform etching is required to ensure that the right amount of operational fins/fingers are present, especially if highly sensitive signal

node is designed (barely saturated). Thus, the actual silicon area of TX would be increased by a certain margin. Also, it has to be understood that actual ratio between total areas of planar vs. FinFET would not see a difference of 8 as other circuit elements as ESD, sampling, TERM and control would take up more combined area than TX driver itself. Use of smaller technology would provide a benefit for some of the other parts of the circuit as well, but the exact ratios are unknown and hence left as part of future research. Moreover, metal gate stack's covered area is also omitted from the analysis.

With the results covered, one can decide for what applications each of the topology - CTT and SLVS - is going to be better suited. Seeing that at 8 Gb/s the active area footprint of CTT is smaller by 25% while power is almost doubled (for FinFET RMG) in comparison to SLVS, one can conclude that overall SLVS topology is substantially more superior. Even though DS is better in terms of active TX area-power product than SES, it has to be noted that total area-power product gap will increase towards 2 with further integration of other system modules. For instance, DS has to have 2 complementary TL present, implying that its covered PCB area is going to be almost twice as large as SES one. As a matter of fact, for high amount of data lines, PCB area ratio converges to a number $\in$[1, 2] due to presence of single guard line in between signal lines. The latter can be visualized using a formula $\frac{4n+2m-3}{2n+2m-3}$, where n denotes number of signal and guard lines and m denotes number of gaps between signal lines and guard lines. Note, it was assumed that gap between complementary signal line pair in DS is approximately equal to a signal line width, hence simplifying calculations for DS area. In case m is significantly larger than n, one can notice that PCB area of SES and DS is going to be the same, while in case n » m, overall area ratio can be seen to be 2. For the particular design case $n \approx m$, thereby, PCB area ratio between SES and DS is 1.5, making DS a better choice for optimal system performance due to both, internal and external area-power optimization. For very tight area budget where power consumption is not an issue CTT could still be used.

With target case results obtained, sensitivity of the design to various parameters has to be investigated. More elaboration on this can be found in the next chapter.

4. Sensitivity Analysis

Using the design definition provided in Chapter 3 and initial results shown in Section 3.7 further analysis of the system can be established. The next logical step in line is sensitivity analysis where impact on design performance due to variations of system parameters is investigated. With design sensitivity predictions provided in Section 2.7 results of SA and their accompanying discussion can be found in the following sections ordered: Section 4.2 for DR variations; Section 4.3 for TL length and guard spacing variations; Section 4.4 for design sensitivity to V_{th}; Section 4.5 for sensitivity w.r.t jitter; Section 4.6 for voltage variations; Section 4.7 for investigation on process changes; Section 4.8 for design sensitivity to temperature. However, before that SA evaluation procedure and nominal conditions have to be clearly detailed as is done in Section 4.1.

4.1. Sensitivity Analysis Strategy: Reference Case and Basis of Evaluation

Before even beginning SA, one can immediately raise the question - if target case of 8 Gb/s was developed to meet minimal compatibility conditions, how can the system tolerate any negatively impacting variation of design parameters? The answer is simple - the SA is performed in an unconventional way. Rather than simply varying parameter values or environmental conditions and seeing their effect on signal quality, it is decided to keep quality requirements constant and determine how much area is required to compensate supposed performance drop. Or on the contrary, how much active area can be relaxed to take full advantage of system performance enhancement. With the above strategy a factor of safety can be determined, which indicates the necessary area scaling to make the nominal configuration applicable for vast amount of use cases with more strict/relaxed requirements. Thereby, area of the active circuity can be set such that particular environmental, manufacturing and technology conditions are met.

From the aforementioned, one can also conclude that design target case provided in Chapter 3 is the most viable option to be used as a reference for SA as it has been already defined and analysed. Thereafter, nominal conditions are set to be the same as before: Imec low V_{th} FinFET 14 nm tech. equivalent, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temperature = 25°C, DR = 8Gb/s, input jitter = 16%. To evaluate whether expected trends are observed, theoretical predictions mentioned in Section 2.7 are going to be used. Bear in mind, deviations from expected trends are predicted - only first order relations were assumed ignoring majority of possible second order effects. With all the aforementioned in mind, design sensitivity to DR is investigated first.

4.2. Design Sensitivity to Data Rate Variation

To begin with design DR sweep, a range for which evaluation of design is attractive has to be defined. As there are currently no of-the-shelf I/O designs close to target DR of 8 Gb/s, at the given moment there is no significant benefit of exploring design conditions for higher than target DR. Exploring DR below 8 Gb/s provides an additional insight on alternative system configurations which can be benchmarked against existing designs. Thereafter, DR sweep is performed in a DR $\in$ [2, 8] Gb/s with a step of 1 Gb/s. Higher DR values are left for future research.

When tweaking only the DR of the design for SLVS and CTT topologies, TX active area footprint variations for various FinFET device types depicted in Figure 4.1 were obtained. Here one can indeed notice that area of CTT grows more rapidly than that of SLVS as predicted in Subsection 2.7.1. Note, Figure 4.1 also depicts industry's nominal power supply condition curve where V_{dd} is taken to be 0.8 V. As expected, common V_{dd} conditions require more active area to balance drive strength reduction caused by lower power supply. The power, on the other hand, is approximately 20% lower which has to be the case when going from 1 V power supply to 0.8 V V_{dd} as depicted in Figure 4.2.

Figure 4.2 clearly depicts that current for the same topology using different devices has to be the same to ensure proper system operations. The power levels are perfectly coinciding for SLVS topology when equivalent simulation conditions are present. On the other hand, CTT RMG power curve sees a deviation up to 10% in comparison to GF and GF extended which could be caused due to multitude of reasons as non-optimal sizing of GF and GF extended device cases and higher miss-match of TERM and TX driver impedance among others. Nonetheless, as the discrepancy is rather small, determination of the exact effect is deemed unnecessary. Moreover, the dynamic power is converging to $\approx$1 mW for CTT and 2.5 mW for SLVS in low DR designs as expected per DC power analysis provided in

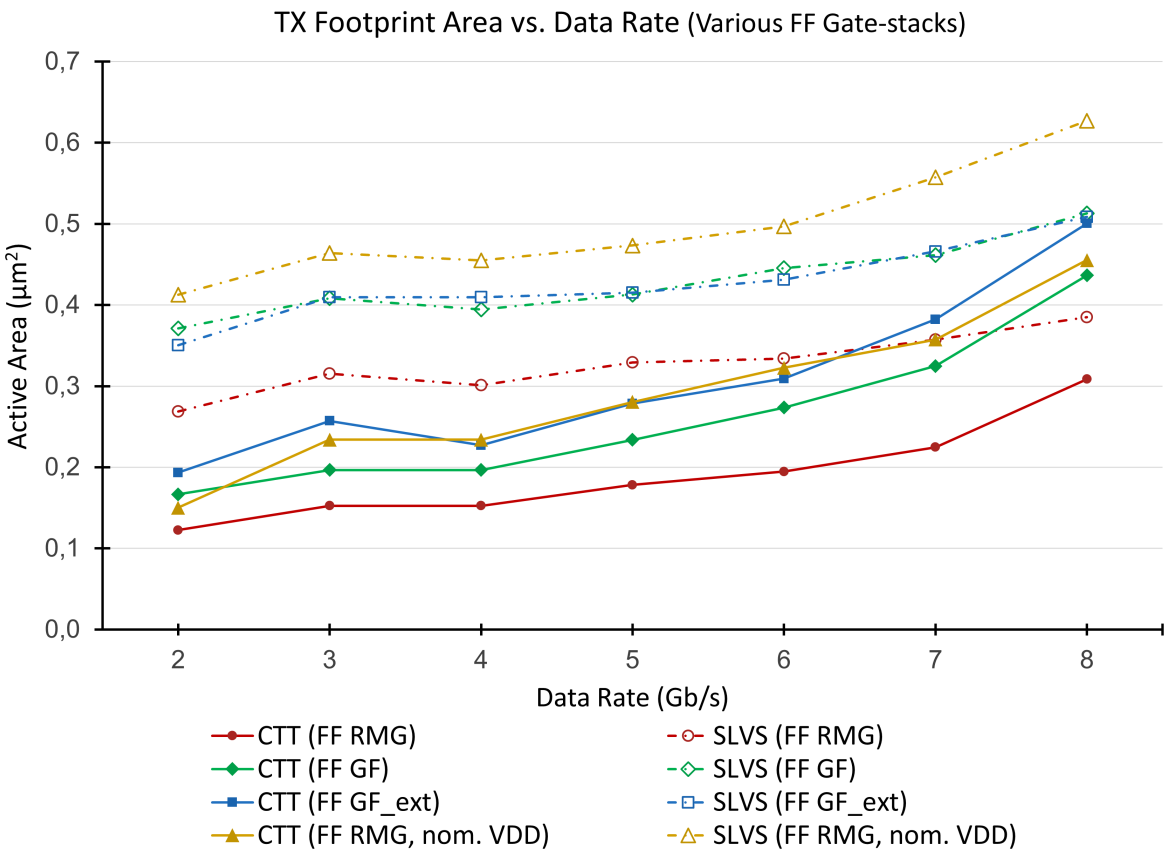


Figure 4.1: CTT and SLVS topology TX active area footprint w.r.t DR $\in$ [2, 8] Gb/s for various FinFET devices

Section 2.2.

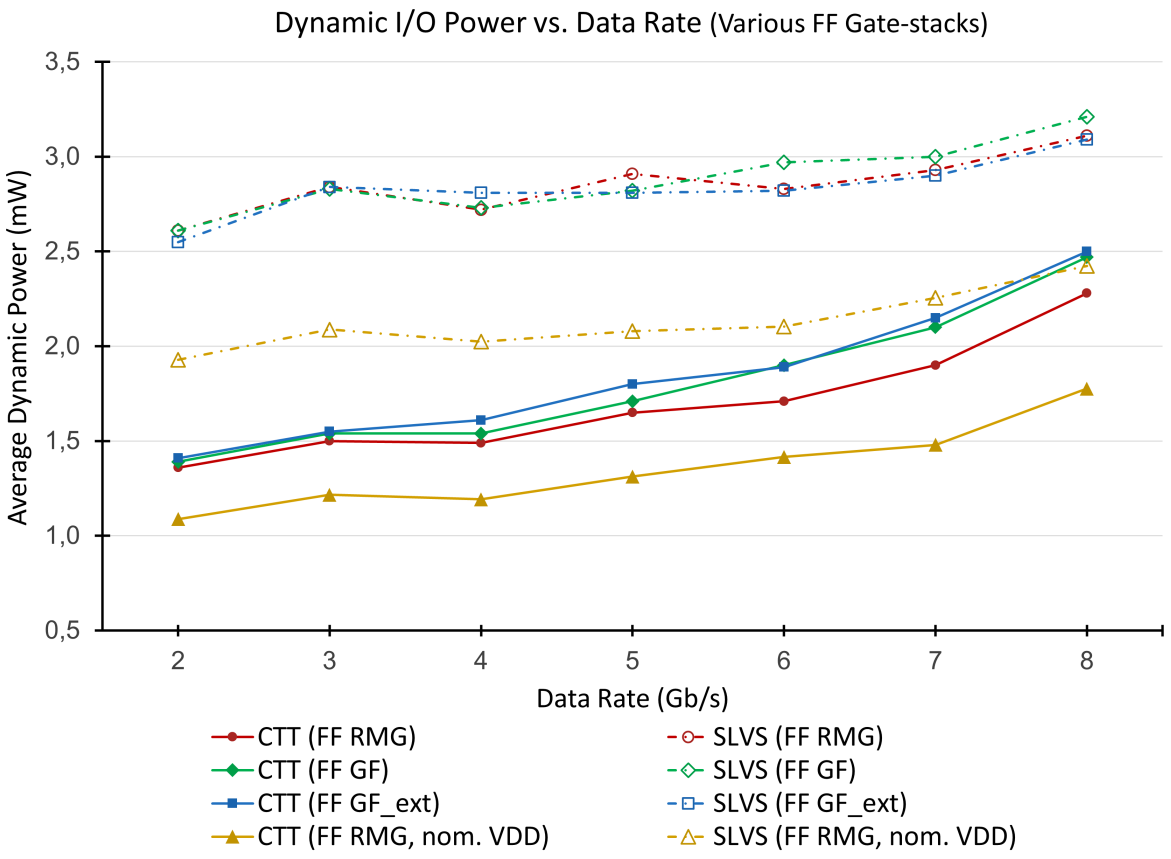
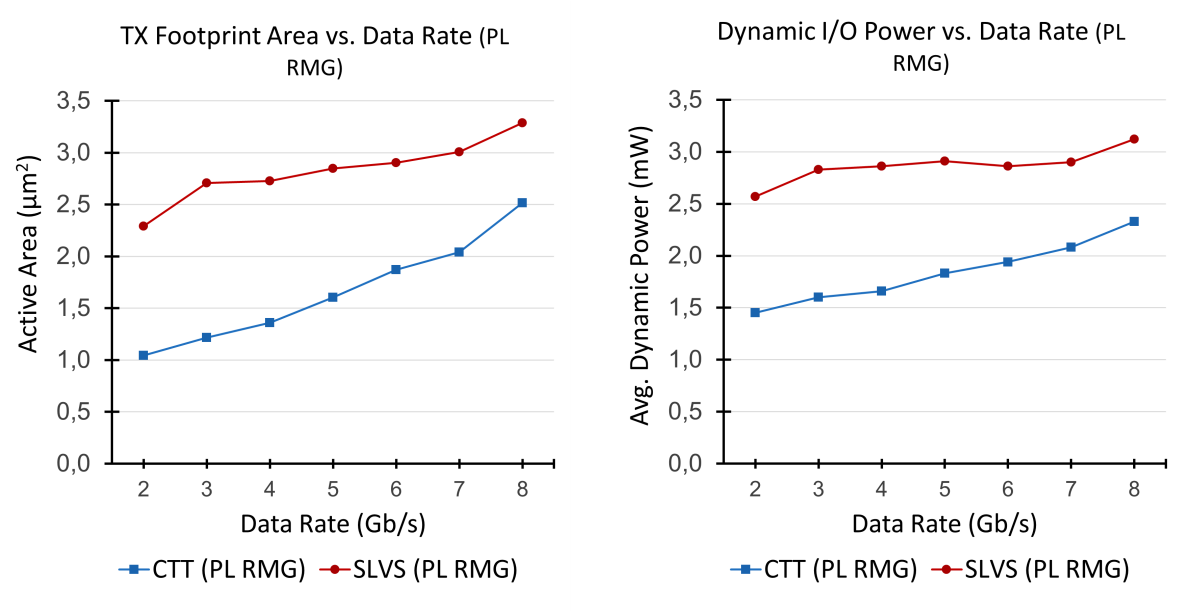


Figure 4.2: CTT and SLVS topology TX dynamic power w.r.t DR $\in$ [2, 8] Gb/s for various FinFET devices

Alike trends for both area and power can be noticed for planar devices as depicted in Figure 4.3a and Figure 4.3b. Note, both planar and FinFET devices exhibit a sudden jump of area and power at 3 Gb/s, which seems to be caused by additional reflections or resonance on the TL upon investigation of signal eye. Observe, the dynamic power graphs provided in this chapter might give counter intuitive representation of actual power ratios between topology schemes. Power variation trend in total power plots cannot be visualized as clearly (see Figure 4.4a) since static and RX power remains approximately constant throughout parameter variations, as neither RX or TERM undergoes almost any changes in their respective sizing. Hence, only dynamic power contributions are provided from this point onward.



(a) CTT and SLVS topology TX active area footprint w.r.t DR ∈ [2, 8] Gb/s for planar RMG

(b) CTT and SLVS topology TX dynamic power w.r.t DR ∈ [2, 8] Gb/s for planar RMG

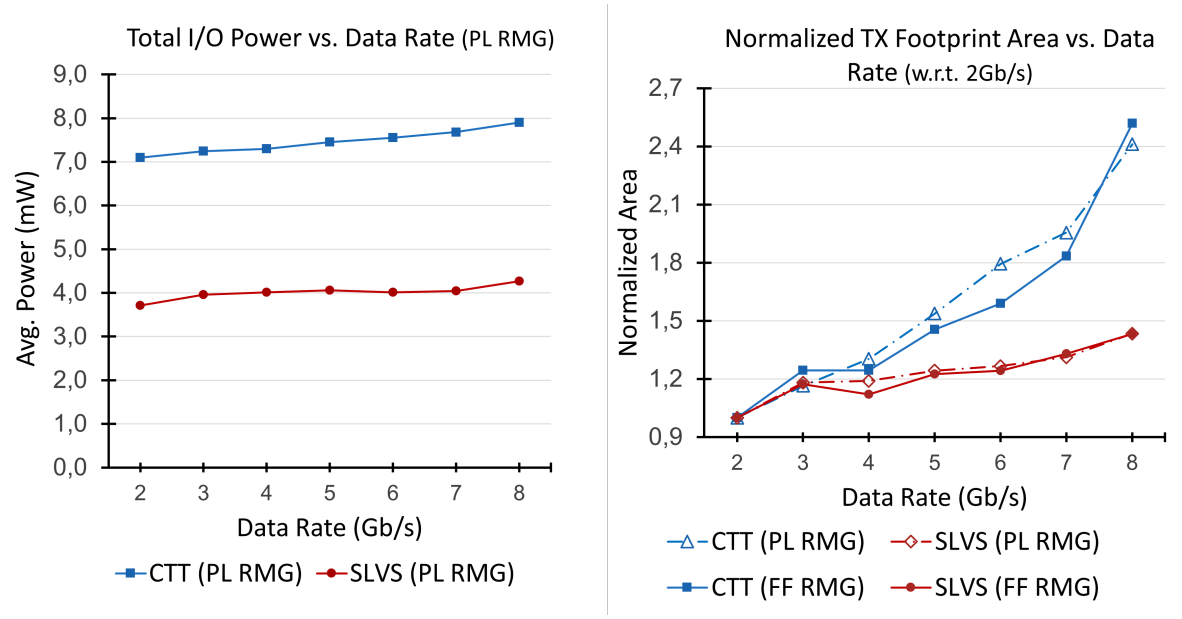
Figure 4.3: Graph indicating SLVS and CTT planar RMG configuration's a) TX active area footprint b) dynamic power for DR ∈ [2, 8] Gb/s

To determine area growth trend accuracy FinFET RMG and planar RMG devices have to be compared. For this purpose relative area change w.r.t 2 Gb/s case for both devices is plotted in a single plot, which can be seen in Figure 4.4b. It can be noticed that relative trends almost perfectly overlap between planar and FinFET devices which is to be expected - both technologies are still FET devices following the same characteristic behaviour. Note, the closely matching relative increase trend between FinFET and planar devices validates the sizing strategy used in this project. Hence, performance comparison between planar and FinFET can be performed in terms of area-power product if applied V_{dd} is different or directly if the same power supply voltage values are used. The slight deviations in the relative trends can be neglected, since the exact origin of the spread cannot be precisely indicated - it might be due to incomplete optimization of values.

Observe, SLVS exhibits significantly stronger EMI rejection and thereby its increase is almost linear w.r.t 2Gb/s as anticipated already in Subsection 2.7.1. From Figure 4.4b, one can conclude that at a DR shortly beyond 8 Gb/s, SLVS area would match that of the CTT due to the high difference in relative area growth trend - at this point SLVS becomes completely dominant over CTT. For low DR a choice between SES and DS can still be performed, as proportional trade-off between area and power is still present, while for very high speed interconnects (DR > 8 Gb/s) DS will always be a better option.

An interesting thing to note is self-loading of FET intrinsic capacitors. FinFET internal capacitance per unit area is larger than the one seen in planar MOSFET [97] due the increased amount of interacting surfaces. Thereafter, potentially the area of the FinFET should have a sharper incline with respect to the data rate. Nevertheless, as the total area of planar devices is substantially larger, the exact role of self-loading cannot be derived. Moreover, intrinsic capacitors of the devices are significantly smaller than TL generated capacitance, effectively diminishing the influence of device parasitics to zero. Minor deviations in under- and overshoot presence as well as slew rate steepness can be observed when comparing planar and FinFET devices - FinFET devices have sharper slew rate response, but more crosstalk induced power rail variations. The former indicates that total FinFET intrinsic capacitances are lower than those of planar alternatives in same DR designs.

Lastly, required output signal swing values to ensure proper operations differ per DR used as can be seen when comparing Figure 4.5 and Figure 4.6. The main cause of the deviation is frequency accompanying effects as crosstalk and lower time interval of voltage application at transistor gates causing higher variation in slew rate.



(a) CTT and SLVS topology total power w.r.t DR $\in$ [2, 8] Gb/s for planar RMG

(b) SLVS and CTT relative increase in area w.r.t 2 Gb/s for planar and FinFET RMG technology

Figure 4.4: Graph indicating SLVS and CTT a) planar RMG configuration's total power b) relative area w.r.t 2 Gb/s for RMG devices

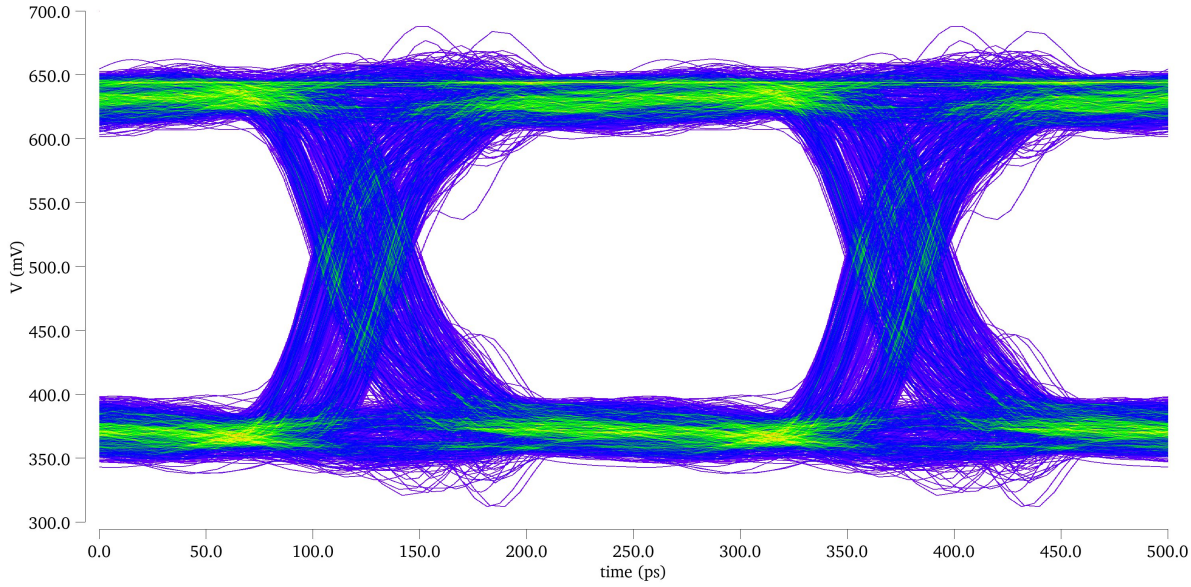


Figure 4.5: CTT topology signal eye at RX input for 4 Gb/s, FinFET RMG. Simulation conditions (refer to Section 4.1): Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 4Gb/s, input jitter = 16%

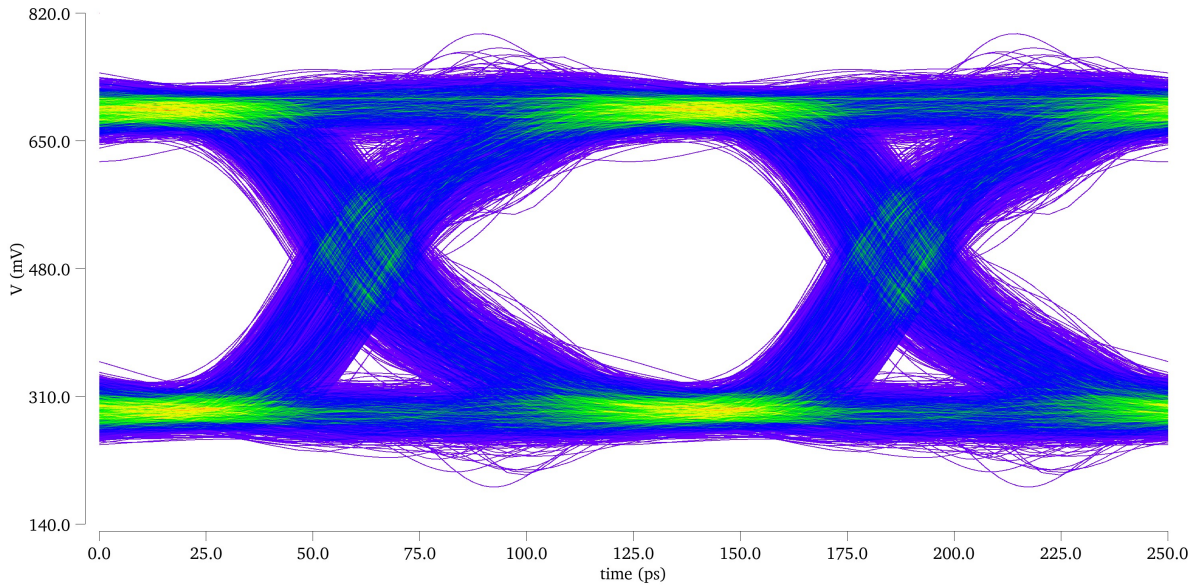


Figure 4.6: CTT topology signal eye at RX input for 8 Gb/s, FinFET RMG. Simulation conditions (refer to Section 4.1): Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

4.3. Design Sensitivity to Transmission Line Parameter Variation

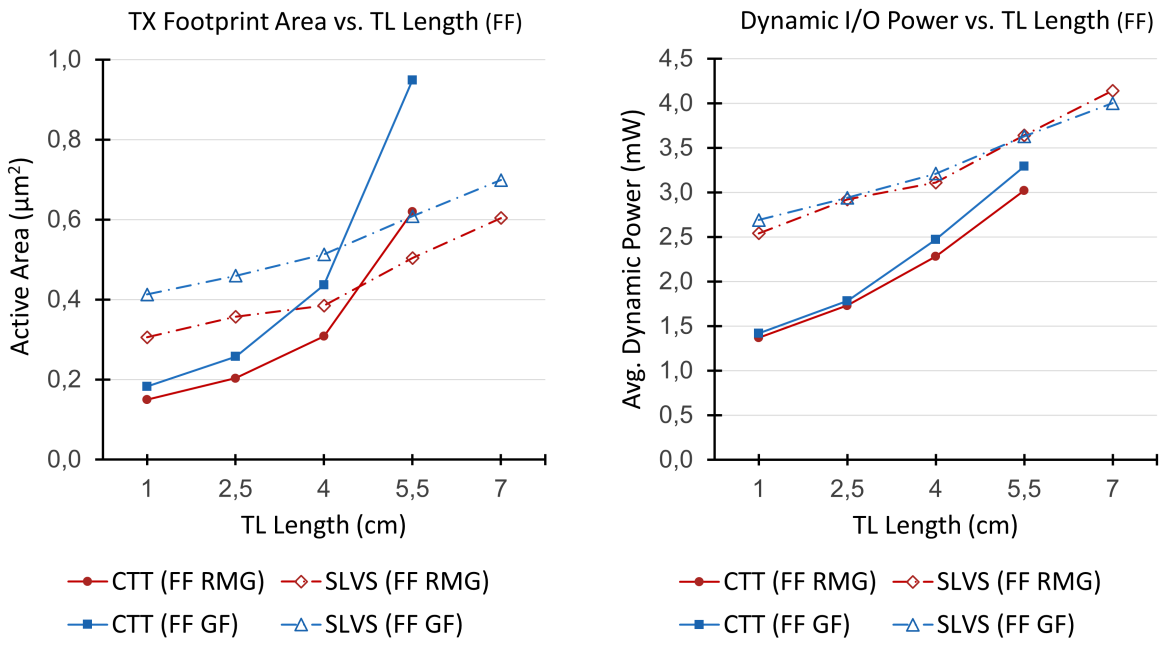
Transmission line structure provides a lot of parameters which can be sized and swept individually - dielectric height, dielectric material, trace width, thickness, etc. However, as the time-sapn of the project is limited, only the most interesting parameters - TL length and adjacent trace spacing - are analysed. The former two parameters affect potential routing and PCB layout configuration the most, hence, putting them higher in priority list of parameters for investigation. First, the TL length sweep is discussed and then the guard spacing is covered.

4.3.1. Transmission Line Length Variation

First things first - use of TERM allowed to successfully eliminate any standing waves on TL for very short traces. Thereafter, investigation of small traces is not appealing anymore as they are always guaranteed to provide a better performance than long traces. Thereafter, minimum TL length value is set to be 1.5 cm, which is approximately side-to-side distance of a commercial NAND memory chip [13]. The upper range limit is determined by finding at what TL length CTT topology fails to provide a decent output at TX driver size twice that of SLVS. With iterative approach, the CTT failing range was determined to be somewhere in the vicinity of 6-7 cm. With TL sweep range $\in$ [1.5, 7] cm and defining step size to be 1.5 cm, area and power variation w.r.t TL length can be seen in Figure 4.7a and Figure 4.7b respectively. Bear in mind, CTT values for 7 cm TL are absent, since design area and power tends towards infinity and thus is omitted from the plot.

In Figure 4.7a and Figure 4.7b, one can notice that power and area increase w.r.t TL length for SLVS topology is indeed close to linear as predicted in Subsection 2.7.2. Also, relative area difference between RMG and GF devices remains constant indicating that proper scaling was performed in both sizing cases. On the other hand, CTT topology exhibits some degree of linearity only for TL length $\in$[1.5-4] cm - above this range active area grows rapidly. Already at 5.5 cm point, SLVS topology can be seen to require less area than CTT. The inferior EMI rejection of CTT makes it an unattractive selection for signalling topology if relatively long interconnects are required for data transmission. Note, high linearity for DS was observed even beyond 10 cm TL length mark (omitted here).

Planar devices see a similar trend, where CTT becomes a completely inferior topology to SLVS at 5.5 cm due to the immense area growth as shown in Figure 4.8a and Figure 4.8b. When comparing FinFET and Planar device dissipated dynamic power, they can be seen to be largely the same for both CTT and SLVS. Thereafter, area factor of $\approx$[8, 9] holds in the entire range of TL length sweep. Notice, 5.5 cm case for planar devices sees a slightly different power and thus also design area curve increase.

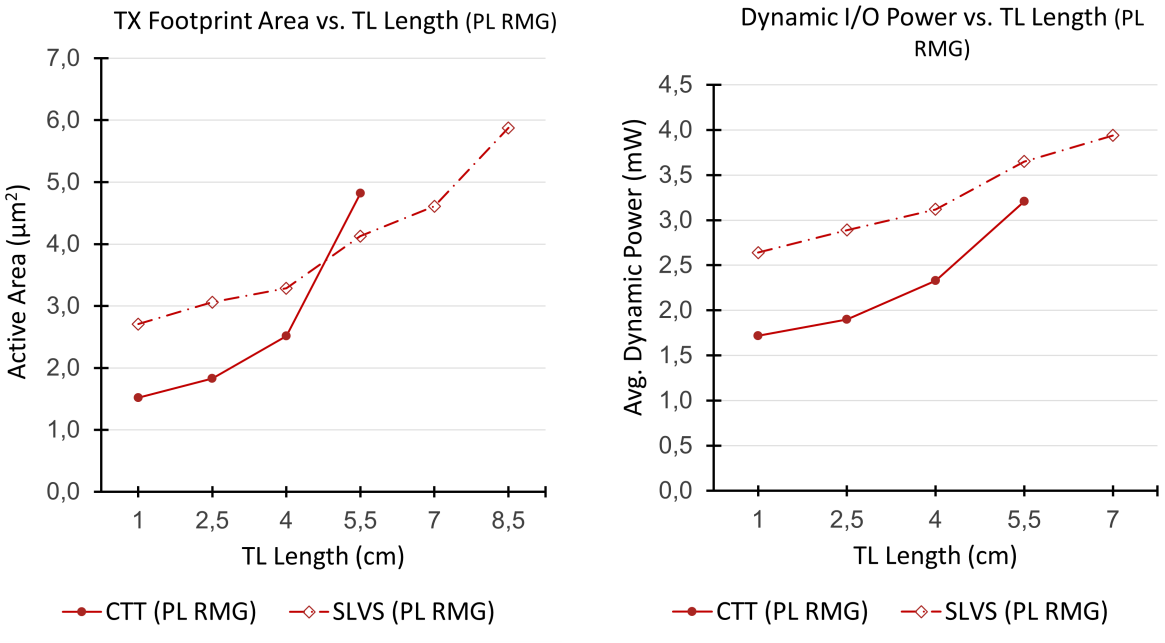


(a) CTT and SLVS topology TX active area footprint w.r.t TL length ∈ [1.5, 7] cm for FinFET devices

(b) CTT and SLVS topology dynamic power w.r.t TL length ∈ [1.5, 7] cm for FinFET devices

Figure 4.7: TL length sweep caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET devices

It was noted late that for this particular design case poor optimization has been performed, meaning that a re-run of the simulation is required.



(a) CTT and SLVS topology TX active area footprint w.r.t TL length ∈ [1.5, 7] cm for Planar RMG

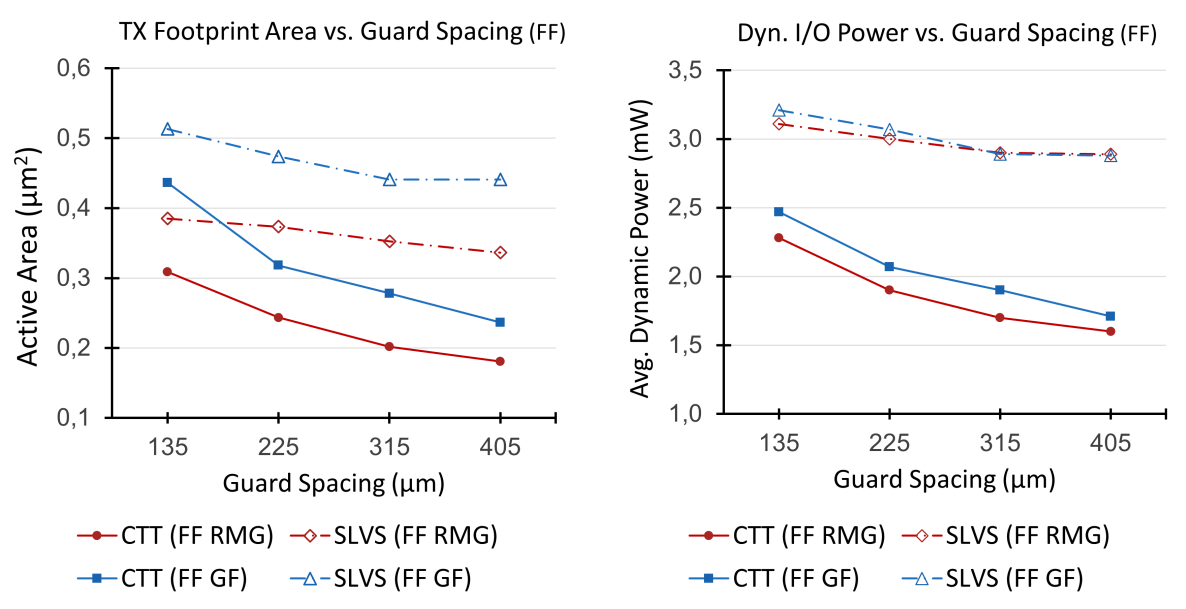
(b) CTT and SLVS topology dynamic power w.r.t TL length ∈ [1.5, 7] cm for Planar RMG

Figure 4.8: TL length sweep caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for planar devices

4.3.2. Guard Line Spacing Variation

Variations of guard line spacing provides similar magnitude but inverted effect to that observed in Subsection 4.3.1. However, contrary to TL length, guard spacing caused area variations converge to a finite value (e.g no crosstalk area) rather than infinity, implying that too large separation range is not needed for the investigation. Distance used in Section 3.7 is equal to 135 μm or 75% of trace width, which coincides with the minimum allowable guard spacing according to adapted manufacturing rules (see Subsection 3.5.1 [18]). Hence, the lower bound of guard spacing and simultaneously the step size is set to be 135 μm.

In order to keep TL length largely unchanged, maximum guard spacing value has to be set not more than 2-3 times larger than trace width. With swing chosen to be $\in$ [135, 405] μm, area and power relaxation response for FinFET devices can be seen in Figure 4.9a and Figure 4.9b. As expected, CTT sees a larger benefit from spreading signal lines more apart than SLVS does. Also, the trend is rather linear to the first degree as predicted in Subsection 2.7.3. It can be seen that SLVS design is close to saturation margin, which is usually reached at a spacing around 5 trace widths [86].



(a) CTT and SLVS topology TX active area footprint w.r.t guard line spacing $\in$ [135, 405] μm for FinFET devices

(b) CTT and SLVS topology dynamic power w.r.t guard line spacing $\in$ [135, 405] μm for FinFET devices

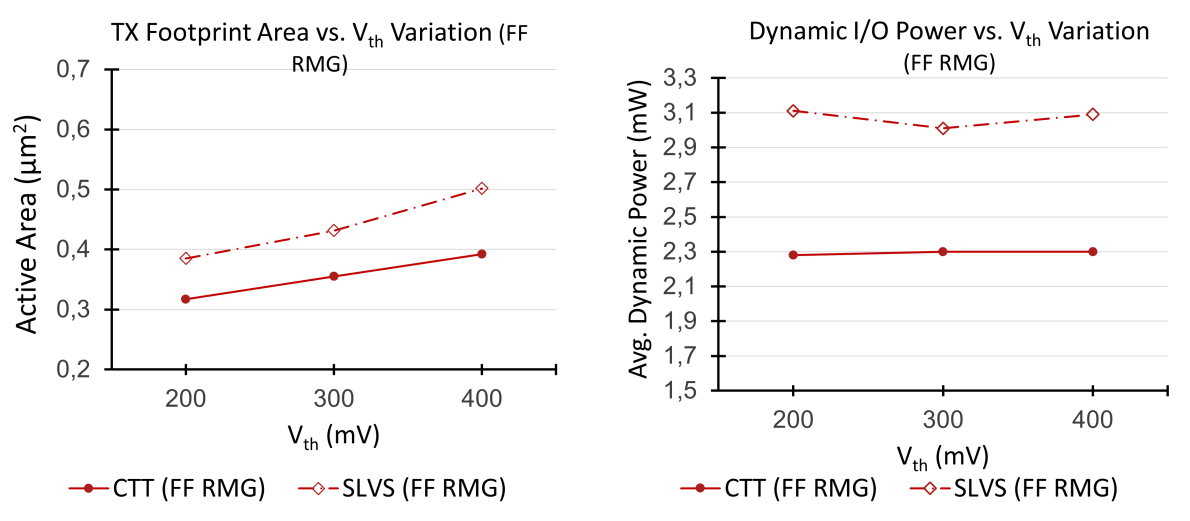
Figure 4.9: Guard line spacing change caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET devices

Linear relaxation trends for both topologies were also observed in planar 45 nm technology. However, as the results lack re-optimization to newest sizing standards for CTT (skewed static TERM), graphical representations are omitted from this thesis. The latter implies dynamic power incompatibility, which is due to higher output swing used. The SLVS on the other hand, showed exactly the same power characteristics, as correct sizing techniques were used. Thereafter, the FinFET to planar ratio was still equal to $\approx$8. From this and TL length analysis, one can conclude that maximum possible routing separation for a given PCB area and shortest possible trace has to be used in order to obtain the highest signal performance at the output.

4.4. Design Sensitivity to V_{th} Variation

Threshold voltage cannot be directly varied for all device types - it requires changing gate metal work function and thus the gate metal itself [27]. However, in case low V_{th} device is given, temperature annealing can be used to partially increase V_{th} to a desired value [104]. As Imec 14 nm FinFET device is characterised to have 200 mV V_{th}, it is assumed that V_{th} in range [200, 400] mV could be obtained with additional manufacturing steps. Thereafter, with the swing defined, required FinFET area increase with increasing V_{th} can be found in Figure 4.10a. The current levels are approximately the same for all

3 V_{th} cases (see Figure 4.10b), indicating that area compensated V_{th} variation in full effect. However, contrary to expectations, the area trend is more linear than quadratic.



(a) CTT and SLVS topology TX active area footprint w.r.t $V_{th} \in$ [200, 400] mV for FinFET RMG

(b) CTT and SLVS topology dynamic power w.r.t $V_{th} \in$ [200, 400] mV for FinFET RMG

Figure 4.10: V_{th} change caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET RMG

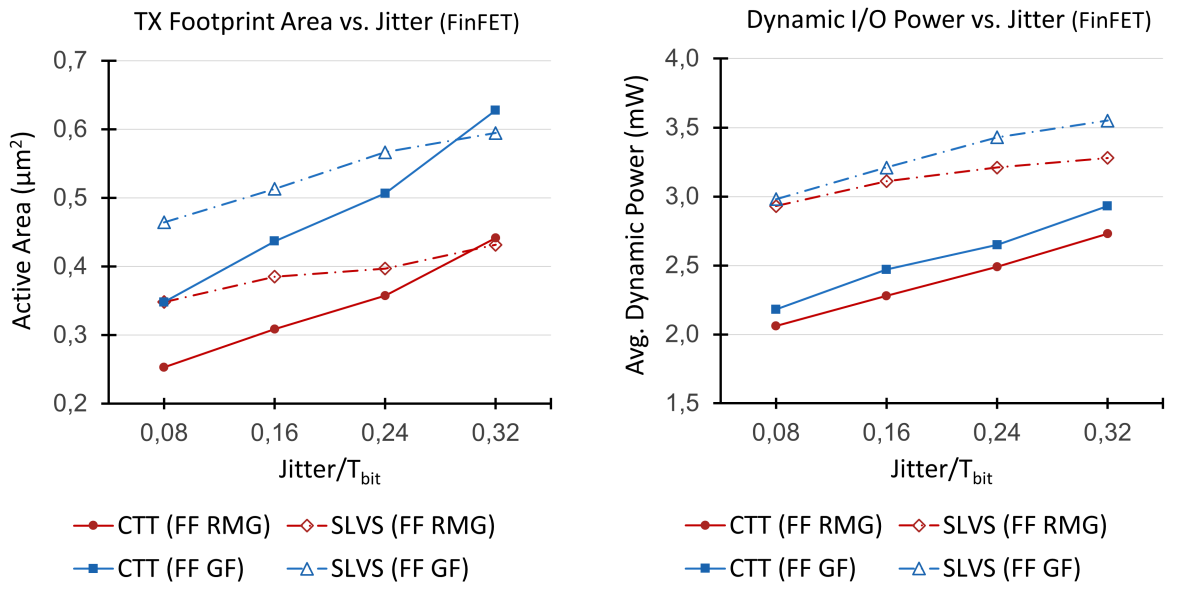
The most likely reason for full linearity is the operational regime of transistors. TX drivers of SLVS and CTT mainly operate in triode region, where current is linearly proportional to V_{th} rather than quadratically as is the case for saturation. Thereby, as slew rate is not the limiting compliance requirement, TX area should increase linearly w.r.t V_{th} opposite of what was predicted in Subsection 2.7.7, where saturation conditions were set as a reference.

Observe, threshold voltage variation was performed only for the RMG model as manufacturing implementation of either technology cannot be ensured. Thereafter, simulations of one compact model were deemed to be representative of the expected area-V_{th} relation. Further exploration in the matter is left as future research. However, at this point one can already indicate that due the linear performance/area dependence on V_{th} high threshold devices are preferred in large chip designs to limit leakage current, while low V_{th} devices are more suitable in high-performance area-restricted design cases.

4.5. Design Sensitivity to Jitter Variation

Design sensitivity to jitter is hard to predict - as it is not always clear whether signal hold time is violated and slew rate is the main driving mechanism of signal quality compatibility. Thereafter, applying jitter in the bounds of [8, 32] % of bit width value, area and power compensation for performance correction can be seen in Figure 4.11a and Figure 4.11b. It can be noticed that SLVS design is less susceptible to noise (including jitter) compared to CTT as for jitter equal to 32% of bit width TX active area of DS topology becomes smaller than that of SES system.

Note, both CTT and SLVS react almost linearly to jitter variations only with differing magnitude. The latter implies that hold time was violated and increase in area was required to increase both slew rate and swing values. As mentioned in Subsection 2.7.8 high gain stages increase the slew rate and thus permit to correct hold time violations from taking place. The exact reason for the obtained trend is rather impossible to explain, it can be only assumed that compatibility corners were breached and slight change in slew rate was sufficient to perform necessary corrections. As exact jitter in the system is merely assumed, this investigation allows to determine the approximate factor of safety which has to be applied if more than design case jitter is anticipated.



(a) CTT and SLVS topology TX active area footprint w.r.t jitter ∈ [8, 32] % for FinFET devices

(b) CTT and SLVS topology dynamic power w.r.t jitter ∈ [8, 32] % for FinFET devices

Figure 4.11: Jitter change caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET devices

4.6. Design Sensitivity to Voltage Variation

Power supply voltage level variation is one of the most common parameter fluctuation present in any electronic system. Both over- and under-voltage lead to impaired chip reliability, with former degrading system life span and latter being responsible for data signal errors. Thereafter, it is important to investigate all potential signal outcomes of different voltage variation modes for a significantly large swing.

The three different voltage variation modes were already mentioned in Subsection 2.7.4, where case 1 (see Subsection 4.6.1) corresponds to only V_{in} variations, case 2 (refer to Subsection 4.6.2) considers only V_{dd} changes and case 3 (look at Subsection 4.6.3) looks at both input and power supply voltage simultaneous fluctuations. For simplicity, the range of voltage variations is set to be the same for all cases - approximately 20% of under- and over-voltage from industry nominal V_{dd} of 0.8 V. Thereafter, voltage sweep range is chosen to be [0.6, 1] V.

Note, as device characterisation voltage (1 V) is used in design case explored in Section 3.7, which serves as reference conditions, entirety of the following investigation can be said to look at under-voltage scenarios. Voltage variations with 0.8 V being the reference should be executed, however, such SA configuration is left for future research. Additionally, voltage variations are performed only for FinFET devices as nominal power supply voltage of planar devices is higher, implying that direct comparison between voltage-area trends cannot be performed.

4.6.1. Skewing of Pre-input Voltage

When only input voltage undergoes changes, output signal waveform is expected to be skewed in comparison to nominal conditions as stated in Subsection 2.7.4. When input voltage has been skewed by 300 mV, the RX input eye diagram for SLVS topology is degrades as shown in Figure 4.12, while CTT signal skewing is shown in Figure 4.13. Observe, CTT experiences lower V_{cm} shift compared to SLVS which can be attributed to two factors: CTT TERM sets a stable common mode voltage and CTT consists of more TX driver stages than SLVS. Lack of stages and floating TERM causes SLVS topology to be more susceptible to independent (1 out of 2) voltage variations, reducing its superiority over CTT topology.

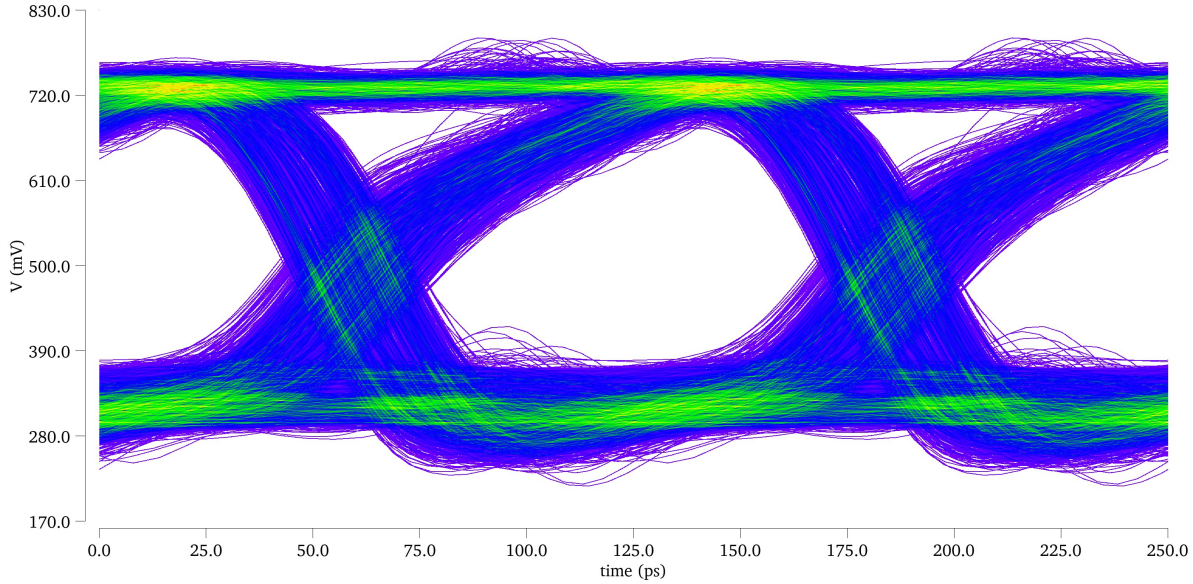


Figure 4.12: SLVS example of signal eye skewness for reduced V_{in}. Simulation conditions (refer to Section 4.1): Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = 0.7 V, V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

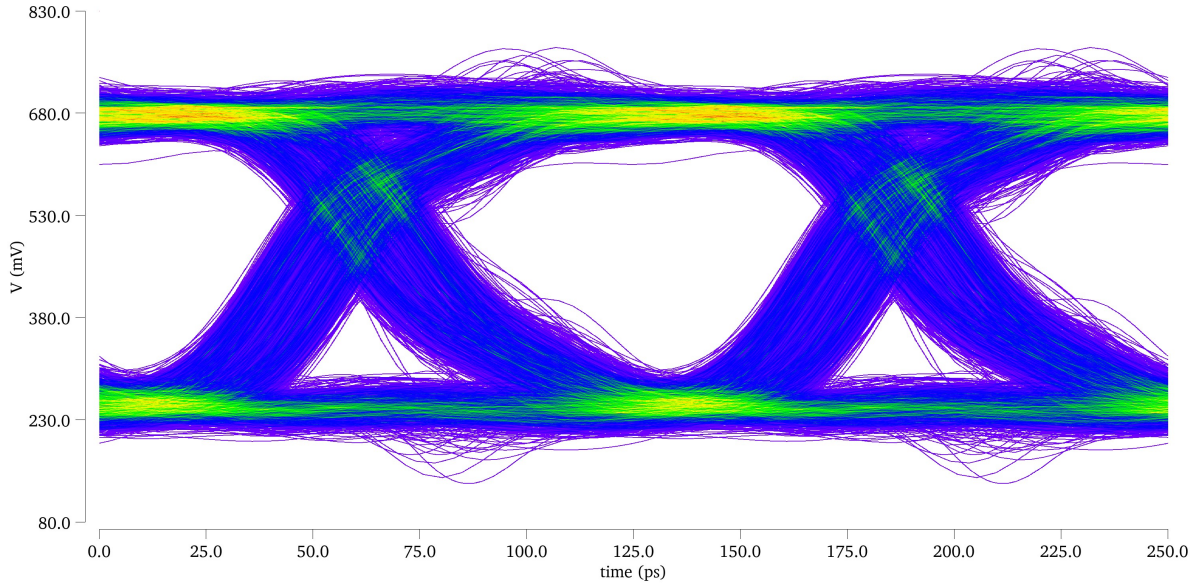
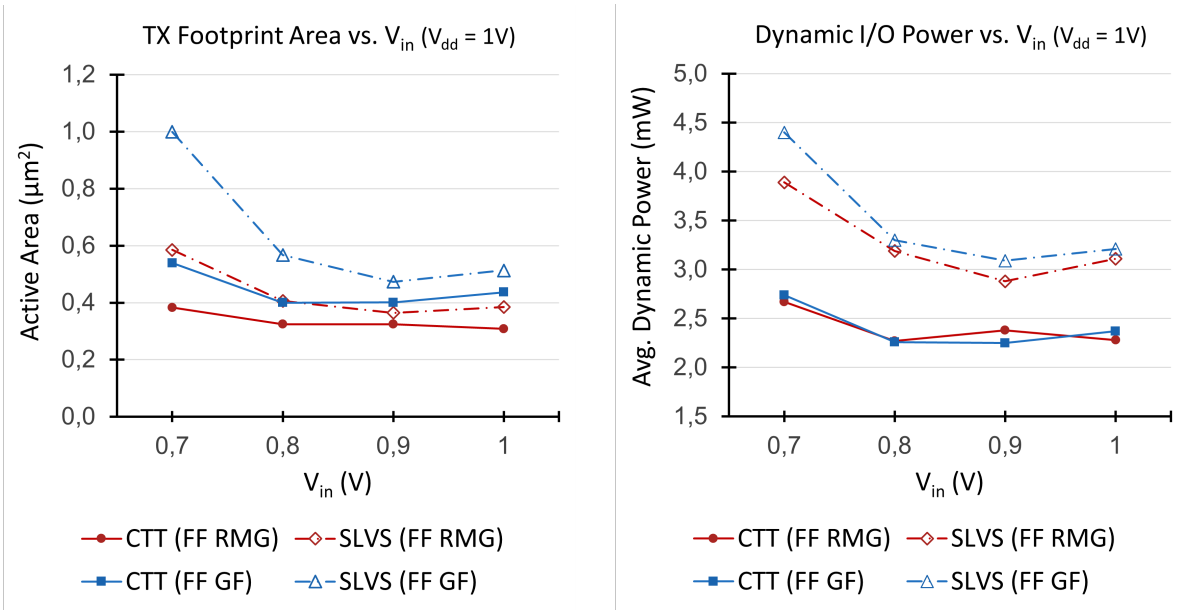


Figure 4.13: CTT example of signal eye skewness for reduced V_{in}. Simulation conditions (refer to Section 4.1): Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = 0.7 V, V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

Required area and power compensation for FinFET devices upon V_{in} variations is depicted in Figure 4.14a and Figure 4.14b respectively. It can be immediately noticed that power for V_{in} range of [0.8, 1] V is rather linear for CTT as expected in Subsection 2.7.4. In 0.7 V case 20% increase in power can be observed which can be associated with the severe skewedness observed in Figure 4.13. Power trend for SLVS represents a bell curved shape, which is completely amiss the predictions stated in Subsection 2.7.4. The trend cannot be explained with high accuracy, however, it is attributed to presence of floating TERM and only 2 stages at the input.

Even more, the H-bridge stage itself is already skewed with higher PUN than PDN sizes due to inverting buffer stage as mentioned in Section 3.7. It can be noticed that V_{in} variations benefit from

this effect under the condition if voltage variations are small - inverting buffer causes down-skewed eye which has to be further corrected by PUN. Observe, area variations follow exactly the same trend as power depicted in Figure 4.14b, since V_{in} is not directly connected to last stage of TX driver. Thereafter, area increase of driver causes proportional increase in power according to both linear (Equation 2.6) and saturation (Equation 2.4) region currents.



(a) CTT and SLVS topology TX active area footprint w.r.t $V_{in} \in$ [0.7, 1] V for FinFET devices

(b) CTT and SLVS topology dynamic power w.r.t $V_{in} \in$ [0.7, 1] V for FinFET devices

Figure 4.14: V_{in} fluctuation caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET devices

Notice, 600 mV case is neglected in figures above as for such case SLVS signal was skewed beyond repair - no TX active area size was able to provide sufficient drive strength to correct the severely unbalanced current generation. In conclusion, CTT topology is slightly less susceptible to performance degradation than SLVS with regards to input voltage fluctuations. In case high variation of data signals is expected, CTT topology is more suited for I/O applications than SLVS due to its more superior V_{cm} biasing.

4.6.2. Skewing of Driver Power Supply

As mentioned in Subsection 2.7.4, varying only TX V_{dd} causes skewing of drive strength similarly as in V_{in} case. Skewedness of RX input eye diagram for SLVS can be seen in Figure 4.15, while that for CTT is shown in Figure 4.16. Observe, signal eyes look almost like a flipped versions of plots given in Figure 4.12 and Figure 4.13 as predicted in Subsection 2.7.4. Here, also SLVS topology sees a higher V_{cm} variation than CTT, similarly as observed in Subsection 4.6.1.

Area and power correction to ensure proper operations can be observed in Figure 4.17a and Figure 4.17b. It can be noticed that power varies almost linearly with slight exponential tendencies - the same prediction was already stated in Subsection 2.7.4. The non-linear behaviour is caused by current values not being constant throughout variation of power supply voltage. The former is true due to the skewing of the eye - to be compatible with compliance mask, signal swing and slew rate has to be increased by increasing the current. Else, both the minimum slew rate steepness and hold time are violated due to lagging rising/falling edge of RX input.

Active area has to balance the quadratic V_{dd} relation to I_D generation, and thus it can be noticed to follow quasi-quadratic trend. Bear in mind, both linear and saturation currents are almost quadratically dependent on V_{dd} - changes on both V_{gs} and V_{ds} are proportional to V_{dd} fluctuations. The slight divergence from the expected trend is caused by the higher current requirement stated above.

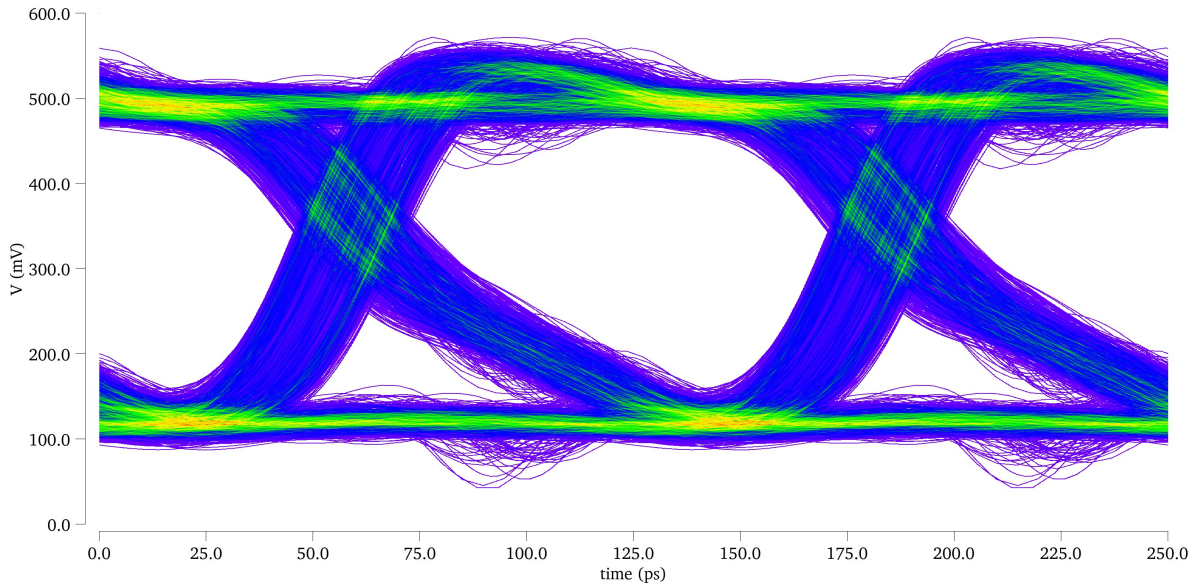


Figure 4.15: SLVS example of signal eye skewness for reduced V_{dd}. Simulation conditions (refer to Section 4.1): Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = 1 V, V_{dd} = 0.7 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

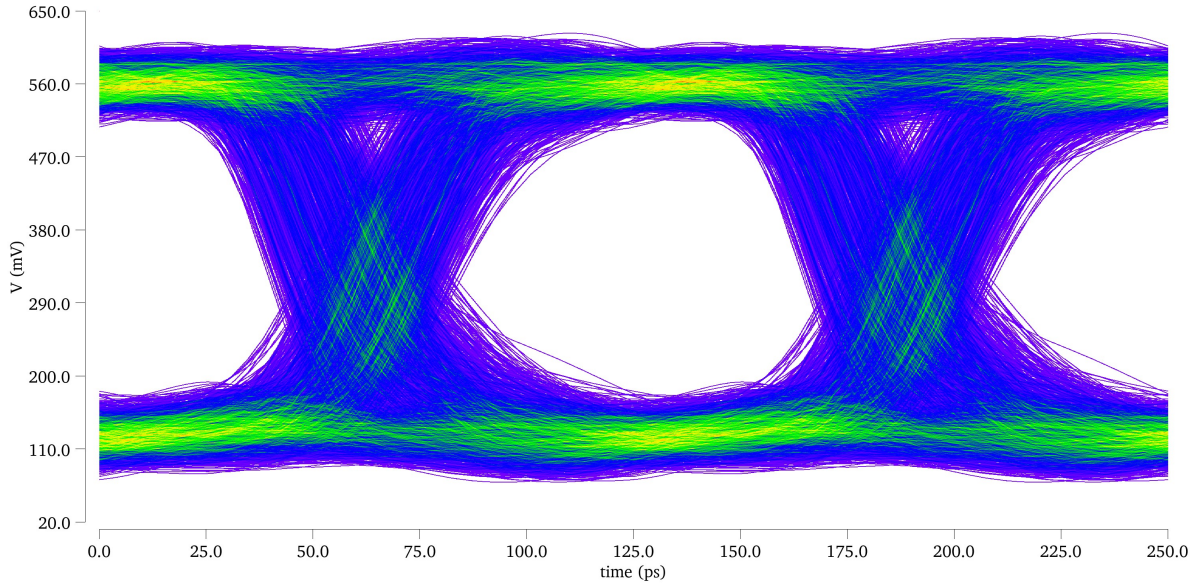
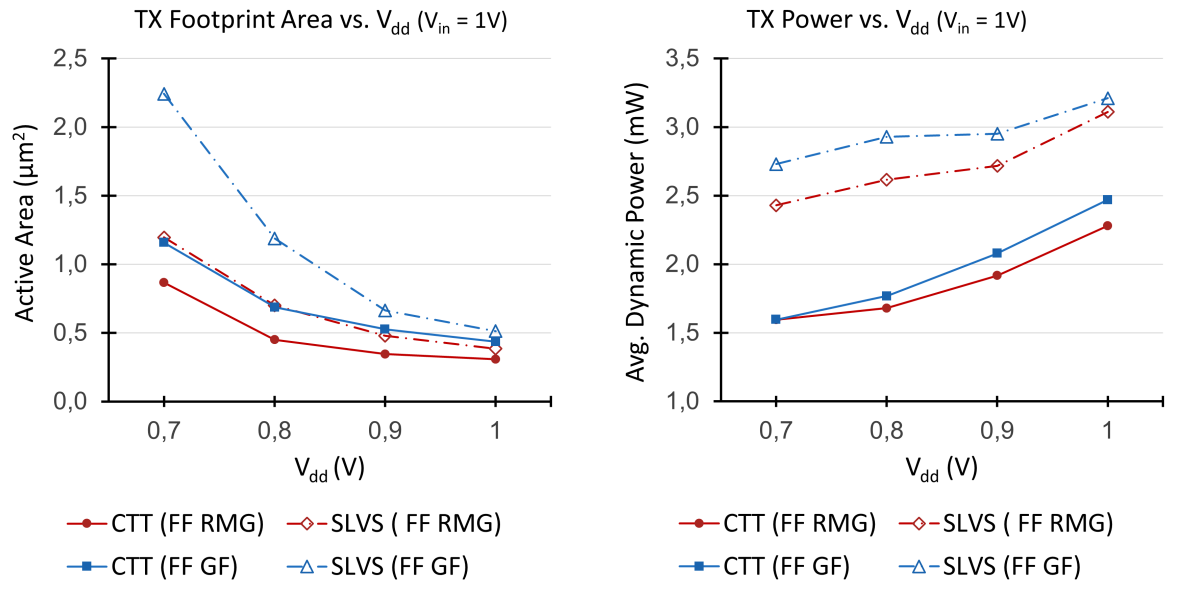


Figure 4.16: CTT example of signal eye skewness for reduced V_{dd}. Simulation conditions (refer to Section 4.1): Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = 1 V, V_{dd} = 0.7 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

Here, once again signal eye of SLVS at 600 mV was skewed too drastically to obtain any TX area able to provide sufficient quality output. From the graphs above, one can conclude that CTT is a better choice if an independent voltage variation is present in the system, simply because SLVS experiences higher signal drive strength skewing due to reasons already mentioned in Subsection 4.6.1.



(a) CTT and SLVS topology TX active area footprint w.r.t $V_{dd} \in$ [0.7, 1] V for FinFET devices

(b) CTT and SLVS topology dynamic power w.r.t $V_{dd} \in$ [0.7, 1] V for FinFET devices

Figure 4.17: V_{dd} fluctuation caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET devices

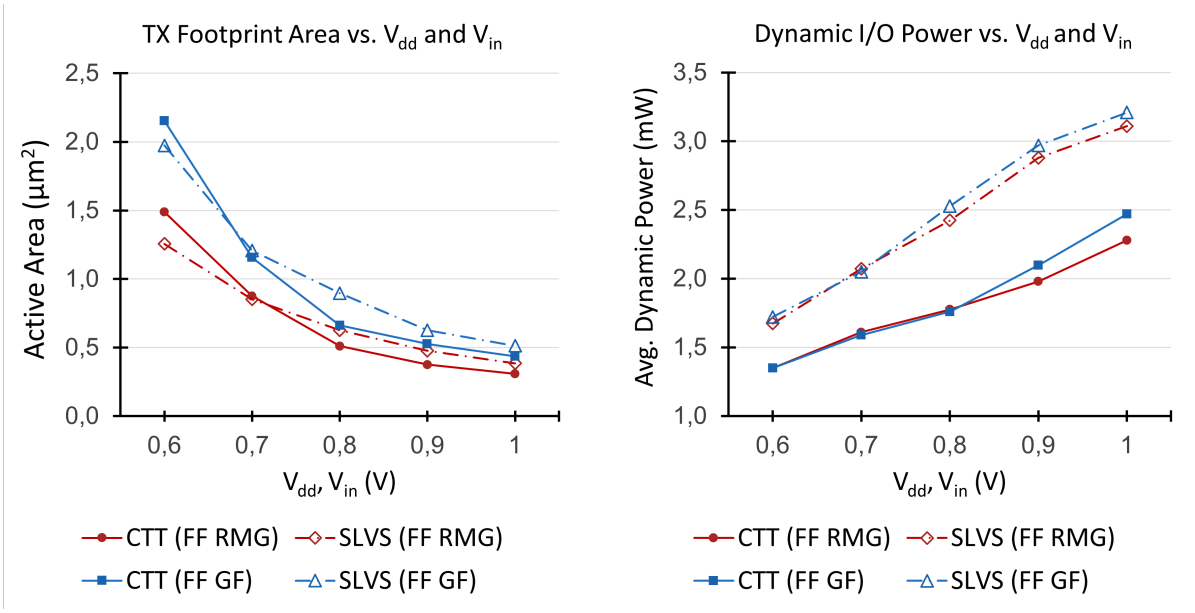
4.6.3. Equivalent Technology Scaling

Lastly, voltage variation of both V_{dd} and V_{in} simultaneously has to be explored. In this case no skewing of the output is present as equivalent drive strength is seen by both PUN and PDN. Thereafter, power is expected to vary with almost full proportionality to V_{dd} which is the case as shown in Figure 4.18b. Notice, current value is almost constant for CTT designs for the entire range of exploration. However, in case of SLVS the current actually does not remain constant - it reduces alongside the power supply voltage. The latter is also the main reason why SLVS area becomes smaller than that of CTT at $V_{dd} = V_{in}$ of 0.6 V as depicted in Figure 4.18a.

Attainable SLVS current reduction could be associated with higher matching of the circuit at lower TX driver area and all TX transistors operating in saturation conditions continuously. As source TERM is not used, driver output impedance is directly connected to the entire impedance network. Thereafter, assuming that TX impedance of reference case is too high providing miss-matched conditions, increase in TX area succeeded in lowering discontinuities and hence reducing severity of reflections. With this in mind, the area could be slightly relaxed to meet the minimum quality requirements.

The second point made above can be explained by looking at an example. For instance, shifting $V_{dd} = V_{in}$ down by 200 mV causes a reduction in V_{cm} of 100 mV as $V_{cm} \approx \frac{V_{dd}}{2}$. Hence, $|V_{ds} - (V_{gs} - V_{th})|$ reduces by 100 mV as $V_{ds} \propto \Delta V_{cm}$ while $V_{gs} \propto V_{dd}$, implying that transistors generate more current per unit area in iso-performance conditions. The same does not apply to CTT as it employs parallel TERM which tightly controls both the equivalent RX input impedance and data signal slew rate conditions. Bear in mind, reduced V_{dd} decreases signal-to-noise ratio present in the system, thereafter one has to look at SLVS increased performance for lower V_{dd} with caution.

With all the above in mind, SLVS can be said to be more superior in case ideal voltage fluctuations are present where both V_{dd} and V_{in} vary simultaneously. As this is likely not the case in real life system due to propagation delays and separate nodes seeing different noise contributions, CTT can be seen to be mildly better choice if meeting tight voltage variation budget is number one priority. In case low power supply fluctuations are expected, SLVS topology stays superior over CTT in terms of area-power product.



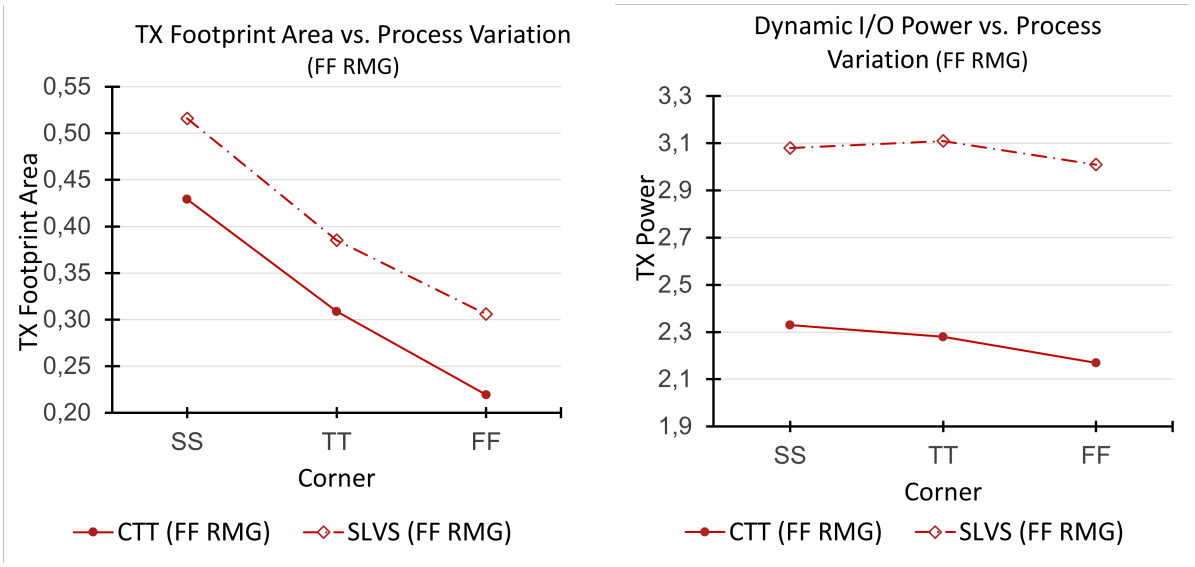
(a) CTT and SLVS topology TX active area footprint w.r.t V_{dd} and V_{in} ∈ [0.6, 1] V for FinFET devices

(b) CTT and SLVS topology dynamic power w.r.t V_{dd} and V_{in} ∈ [0.6, 1] V for FinFET devices

Figure 4.18: V_{dd} fluctuation caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET devices

4.7. Design Sensitivity to Process Variation

Usually process, temperature and voltage variations are investigated simultaneously as all three parameters have inter-dependency of one another. However, in this thesis it is assumed that all PVT parameters can be split and analysed separately. Thereafter, FF and SS process corners are first explored independently of temperature and voltage.



(a) CTT and SLVS topology TX active area footprint w.r.t FF and SS process for FinFET devices

(b) CTT and SLVS topology dynamic power w.r.t FF and SS process for FinFET devices

Figure 4.19: Process variation caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET devices

To determine process variation effect, technology file used in simulations has to be explored. Given FinFET model process corner variation is set up in a relatively crude way - when corner is changed, only the dimensions and doping of transistor are varied by a certain degree. More accurate compact

models have been developed, however, they require higher level access which is not provided during this internship. Thereby the results provided further have to be used only as a first hand estimation of the expected variations.

Area and power change with FF and SS in comparison to TT can be seen in Figure 4.19a and Figure 4.19b respectively. One can note the almost linear dependence of area with respect to process corner is present which has to do with the implementation of compact model. No further details on model implementation can be disclosed.

Power for FF corner can be seen to reduce slightly for both CTT and SLVS topologies, however, as the decrease in current is small, the exact cause for observable discrepancy cannot be pinpointed. Thereafter, neither of the topologies reacts differently to the corners, implying that area-power product ratio between SLVS and CTT remains unchanged.

4.8. Design Sensitivity to Temperature Variation

As mentioned in Subsection 2.7.5, location of ZTC point is the first step in determination of temperature related circuit performance variations. The graph of Imec 14 nm FinFET model I_D vs. applied gate voltage for several temperatures can be seen in Figure 4.20. Immediately the absence of ZTC can be observed, implying that no matter circuit biasing conditions, system performance is always increasing with reduced temperature. It could be that instead of being completely absent, ZTC is shifted to the far left side (cut-off region), where it cannot be seen. As a result of missing ZTC, I_D varies linearly upon equivalent increments of temperature, inclining one to believe that FinFET behaviour upon temperature variation is not implemented in compact model to full effect.

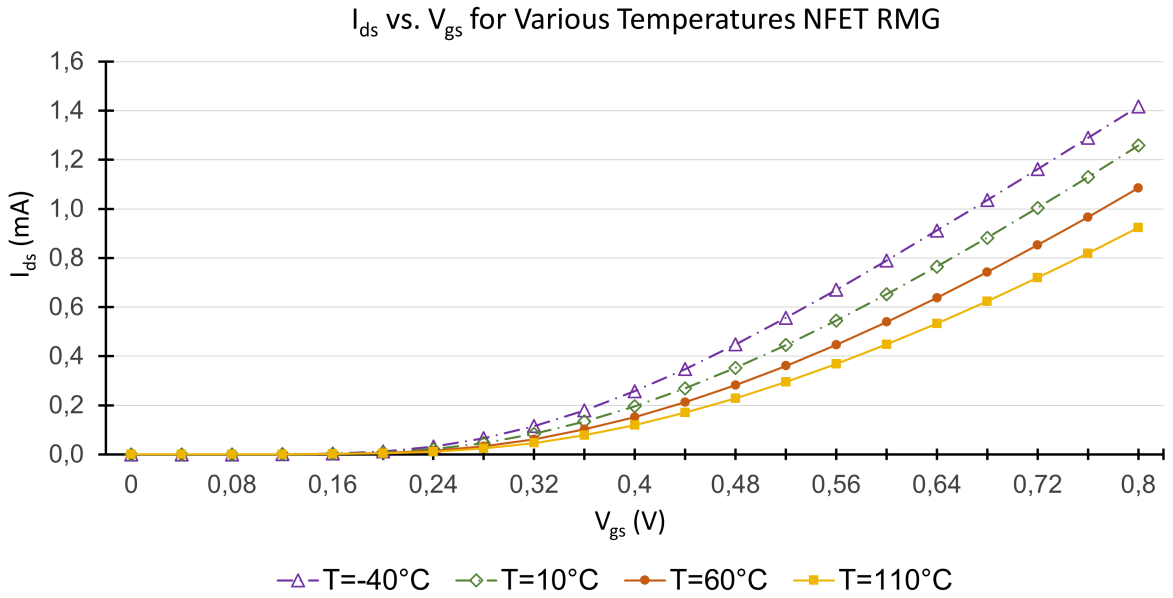
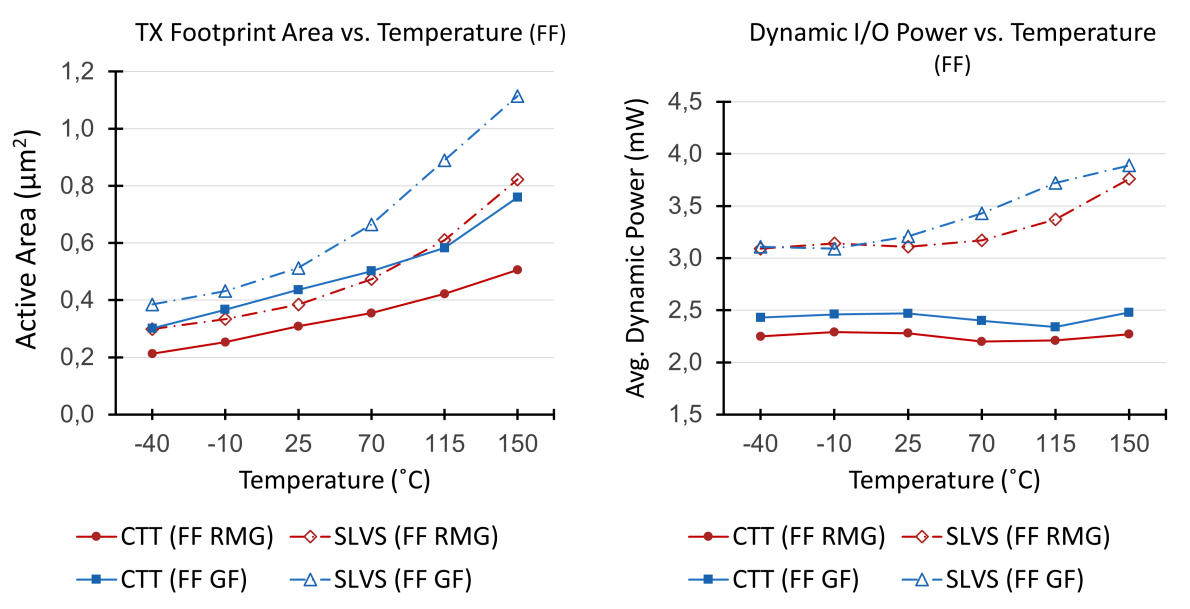


Figure 4.20: Imec 14 nm NFET drain current vs. applied gate voltage for various system temperatures

With the aforementioned in mind, circuit area and power compensation w.r.t temperature can be seen in Figure 4.21a and Figure 4.21b. The CTT power can be seen to remain largely constant, which leads to linear increase in area w.r.t temperature as expected. Nonetheless, SLVS topology shows non-linearity above 115°C, which can be explained by skewing the design experiences (similarly as for V_{dd} or V_{in} changes). An example of RX input eye at 150 °C can be seen in Figure 4.22. Note, the skewing is less severe than in voltage source variations, however, it still leads to increase in required current to obtain proper output. With this, one can conclude, that CTT topology is slightly less susceptible to temperature variations than SLVS for the given compact model.



(a) CTT and SLVS topology TX active area footprint w.r.t temperature ∈ [-40, 150] °C for FinFET devices

(b) CTT and SLVS topology dynamic power w.r.t temperature ∈ [-40, 150] °C for FinFET devices

Figure 4.21: Temperature variation caused CTT and SLVS topology variation of a) TX active area footprint b) dynamic power for FinFET devices

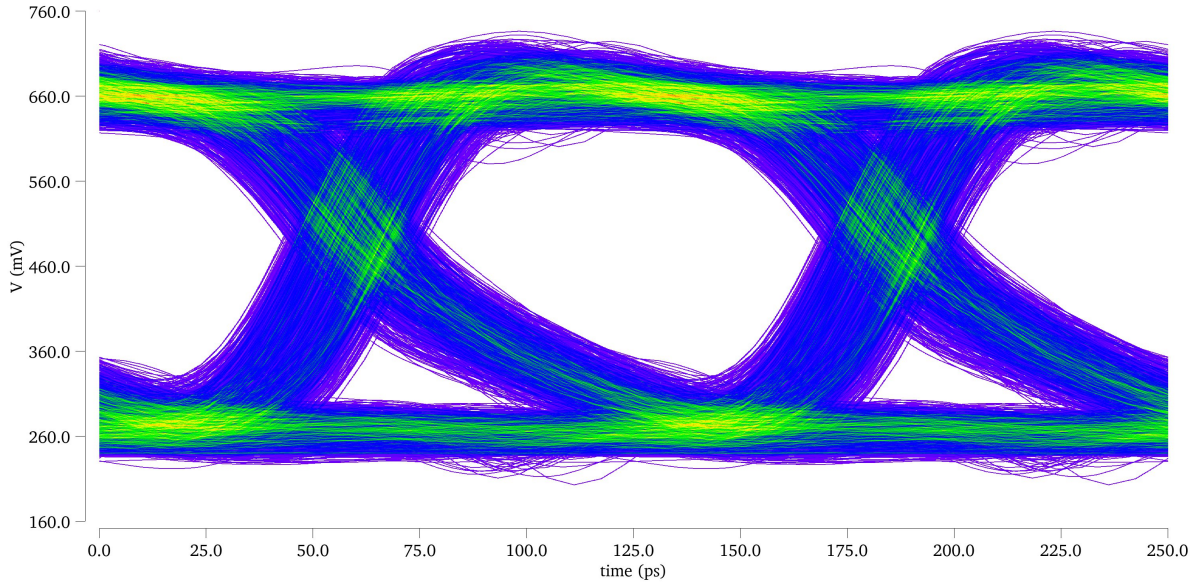


Figure 4.22: Signal eye of RX input for SLVS for 150°C temperature. Simulation conditions used: Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 150°C, DR = 8Gb/s, input jitter = 16%

4.9. Critical Corner Determination and Analysis

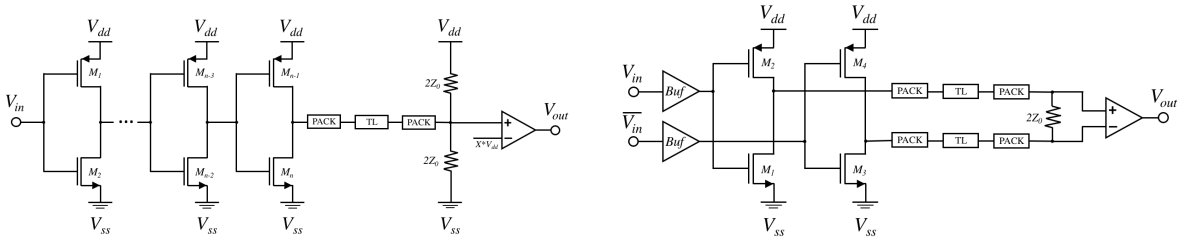
Since both process and temperature dependence is not fully implemented into the compact model, there is no valid point to perform complete PVT analysis of the system. Thereafter, determination and investigation of critical design corners is left for future research when access to more accurate process and temperature influence on circuitry is granted.

With both target case defined and sensitivity analysis performed, conclusions of the study can be drawn. Summary of main take-aways and recommendations can be found in Chapter 5.

5. Conclusions and Future Directions

The goal of this project was to develop and analyse a 3D NAND compatible I/O structure able to reach 8 Gb/s transmission speed using thermally stable Imec in-house developed 14 nm FinFET (Fin Field Effect Transistor) technology equivalent devices. Three different manufacturing flow device types were considered: replacement metal gate (RMG), gate-first (GF) and GF extended. Analysis part consists of design sensitivity exploration of several parameters such as data rate (DR), transmission line (TL) length variation, voltage, temperature, process to name a few. Lastly, the results were benchmarked against thermally stable planar 45 nm technology to determine the benefit of using a lower FET technology node in NAND I/O development.

To ensure an unbiased evaluation of the devices, two I/O topologies were used. Comparing performance of devices across multiple designs allows to certify that resulting enhancement in system behaviour by change of device type is accurate. For this purpose one single ended signalling (SES) topology named center tapped termination (CTT, shown in Figure 5.1a) and one differential signalling (DS) topology called scalable low-voltage signalling (SLVS, shown in Figure 5.1b) were used. In addition to confirming performance comparison validity, use of SES and DS allowed to investigate system susceptibility to frequency dependent parameters such as noise and crosstalk.



(a) Schematic of CTT topology

(b) Schematic of SLVS topology

Figure 5.1: Depiction of a) CTT b) SLVS signalling topology used in simulations

To determine whether required DR has been reached, compliance eye mask (CEM) was developed, which indicates minimum requirements of signal swing, hold time and slew rate. Before that however signal eye diagram at a particular circuit node is obtained by slicing full transient signal into period long intervals (unit interval (UI)), overlapping and centering them. As a result, the waveform looks like Figure 5.2, where internal hexagon represents CEM. The minimum requirements set for the particular investigation are: eye width has to be larger than $\frac{1}{2}$UI and eye height has to be at least 200 mV centered in a continuous time interval of 30% UI. Signal has to be evaluated at the receiver input node as accustomed in commercial standards [13].

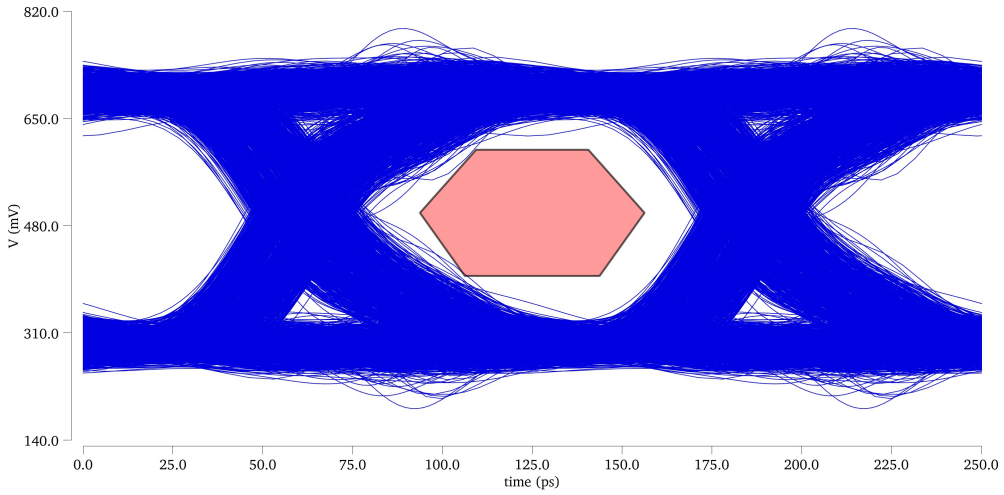


Figure 5.2: An eye diagram including compatibility eye mask for CTT topology with following simulation conditions: Imec low V_{th} FinFET 14 nm tech. RMG, V_{in} = V_{dd} = 1 V, TL_{len} = 4 cm, line spacing = 135 um, corner:TT, temp = 25°C, DR = 8Gb/s, input jitter = 16%

With the aforementioned in mind, test setup including both amplitude and time noise modules as well as electromagnetic coupling implementation between adjacent signal traces was made. Power supply voltage was set to 1 V, which corresponds to FinFET characterisation voltage value, TL length chosen to be 40 mm and temperature, process are set to nominal industry conditions (25°C, TT). Simulations were performed for 3200 bits which is significantly too low for verification of standard bit error rate - 10^{-9}. Nonetheless, the chosen number of bits was determined sufficient for simulations in this thesis.

Using above defined simulation model and signal quality requirements, necessary DR of 8 Gb/s was observed to be reachable by both planar 45 nm devices and FinFET 14 nm devices. The main difference between designs of different device technologies was determined to be transmitter (TX) active area footprint, while dynamic power remained largely unchanged as can be seen in Table 5.1. Similar power values across different topologies indicate that sizing was performed properly since current to generate a pre-defined swing on shared signalling topology has to be approximately the same when equivalent power supply voltage is applied. When comparing active area between FinFET RMG and planar RMG devices of DS topology, ratio of 8.4 can be found, while CTT comparison bears difference of 8.1 times. Thereafter, it can be said that even though the ratio seems unlikely high, it is accurate. The exact generation of the ratio can be determined by comparing device type properties: FinFET vs. planar channel length reduction causes active area decrease of 1.7^2, going 3D enhances ratio by another factor of $\approx$2 and removing PFET mobility inferiority provides a factor of 1.5 boost. With this in mind, one can conclude that FinFET devices would relax I/O area requirements allowing to achieve higher DR systems for strictly limited area cases or increase memory density for iso-DR design. In case industry nominal power supply values (0.8 V for FinFET, 1 V for planar) are used, systems using FinFET benefit by both lower power and area than equivalent planar designs - power consumption is reduced by 20%, while area sees a 4-5 time reduction.

When comparing FinFET devices in between themselves, RMG device can be seen to provide approximately 30%/50% better area-power product performance in SLVS and 40%/60% in CTT vs. GF/GF extended devices. The discrepancy between SLVS and CTT is caused by SLVS TX driver area skewing, which is required due to slightly unbalanced NFET and PFET drive strengths in FinFETs. CTT does not require any correction since its Thevenin (double parallel) termination sets the common mode voltage - SLVS termination is floating, preventing it from controlling output common mode. As process complexity of RMG devices is higher than that of GF due to additional chemical polishing steps of dummy gates, the overall manufacturing costs of full RMG vs. GF system might be higher. Thereafter, area-cost trade-off has to be performed, which is left as future research. From all the above one can derive that RMG devices are more suitable for low area designs requiring high performance, while GF and GF extended are ideal for low-budget and decent performance applications.

Table 5.1: Results of I/O topologies for various devices reaching 8 Gb/s transmission speed

Topology	Device Type	TX Active Area Footprint (μm)	Dynamic TX Power (mW)	Total Design Power (mW)
CTT	RMG	0.31	2.3	7.9
	GF	0.44	2.5	8.0
	GF_ext	0.50	2.5	8.0
	RMG (Planar)	2.52	2.3	7.9
SLVS	RMG	0.39	3.1	4.3
	GF	0.51	3.2	4.4
	GF_ext	0.58	3.1	4.2
	RMG (Planar)	3.29	3.1	4.3

When comparing performance exhibited by DS vs. SES, it can be noticed that DS is overall more superior in terms of area-power product for any device as total power ratio differs by a factor of 1.8 while the active area footprints are approximately of the same magnitude. However, total area of DS system will see a larger difference w.r.t SES than the currently observable ratio due to requirement of additional signal paths on the external interconnect, realized via printed circuit board (PCB). If guard lines are used, the ratio between CTT and SLVS PCB areas ranges in [1, 2] depending on the exact sizing and separation of signal paths. With DS being more superior than SES in terms of crosstalk and noise rejection, SLVS is the choice for high speed interconnects with DR in the vicinity or exceeding 8 Gb/s.

The former can be stated with high certainty as gap between CTT and SLVS shrinks with increasing DR as can be found in Figure 5.3, where FinFET TX active area footprint is provided vs. DR. To better visualize it, comparison of relative area increase w.r.t DR is depicted in Figure 5.4. Observe, SLVS relative area trend is linear and smaller than that of CTT, implying that SLVS area is predicted to catch up with CTT for a DR slightly above 8 Gb/s.

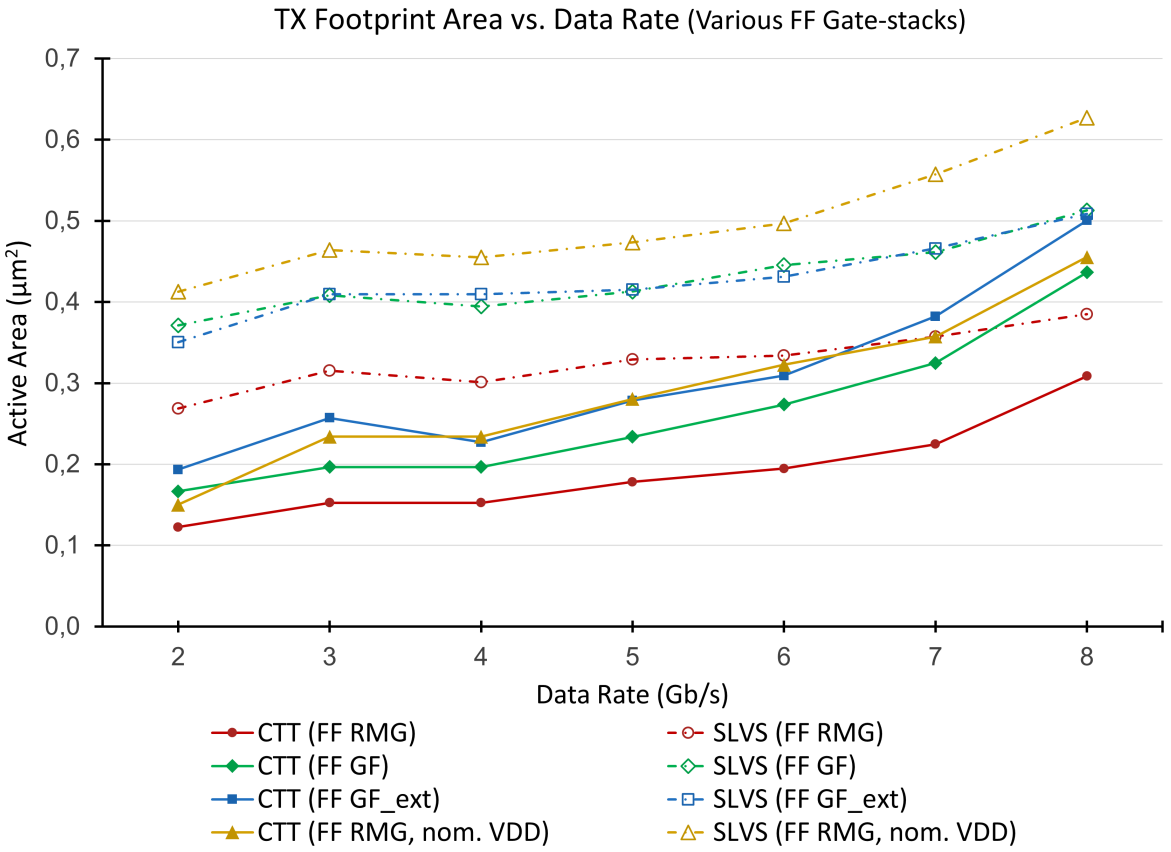


Figure 5.3: CTT and SLVS topology TX active area footprint w.r.t DR $\in$ [2, 8] Gb/s for various FinFET devices

Similarly, SLVS has been shown to be a better option in case large TL length is required as CTT topology fails to provide comprehensible output when TL reaches approximately 6-7 cm. DS shows inherently high immunity towards electro-magnetically induced noise, thus showing almost linear area growth with TL length expansion. To reduce SLVS superiority, spacing between adjacent signal paths can be increased. Since CTT is more susceptible to crosstalk, increased gap between two neighboring signal lines leads to quicker reduction in mutual capacitive and inductive components, allowing for better coupling with the return path. Bear in mind, if separation between lines is increased too much, CTT and SLVS PCB area becomes similar and TL length has to be increased, making CTT less appealing option.

As compact model device dependency on process and temperature variation is implemented with first order approximations, sensitivity analysis of these parameters provide inconclusive results. Voltage

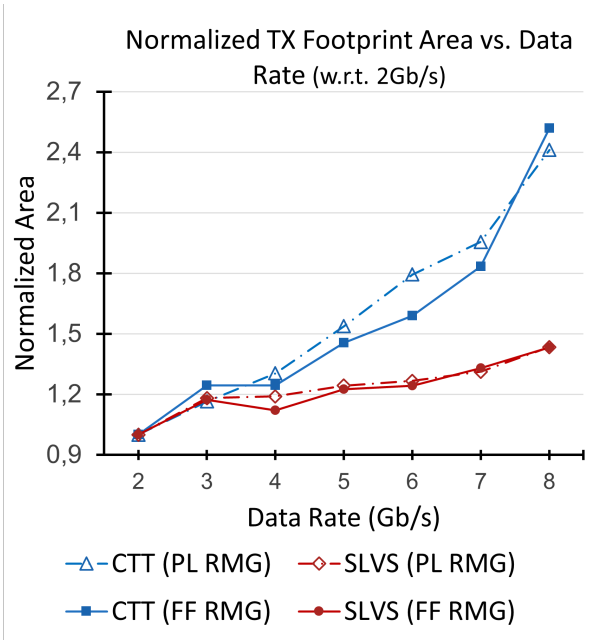


Figure 5.4: SLVS and CTT relative increase in area w.r.t 2 Gb/s for planar and FinFET RMG technology

variations also have to be inspected with caution - while simultaneous input and power supply voltage variation provides better scaling factor for SLVS than CTT, independent scaling degrades SLVS output signal severely. SLVS is found to be more prone to signal skewing due to floating common mode voltage, implying that substantially different input and power supply voltages can damage the output signal beyond repair.

Lastly, it has also been shown that SLVS is less sensitive to time dependent bit width variations also known as jitter. This is to be expected with DS having a stronger amplitude and time noise immunity than SES. Additionally, upon investigation of system reaction to device threshold voltage, both SVLS and CTT based designs experienced area variations with equivalent slope.

To improve the design and the current analysis further, several actions can be taken. First, active on-die termination (ODT) can be explored, which was omitted due to its implementation complexity. As active ODT can be varied throughout circuit operations, it can provide better impedance matching, which reduces reflection caused degradation. Moreover, more toplogies than the given ones should be investigated, as broader dataset would boost the validity of results obtained in this master thesis and allow to better determine what configuration is better suited for particular application (e.g. low area, low power, low cost). Investigation of design sensitivity to other parameters as TL impedance would provide additional insights on system response to external conditions. With more parameters investigated, the likelihood of obtaining operational circuit upon manufacturing increases as potential performance degrading mechanisms have been taken into account accordingly. Lastly, generating a layout of the system covered in this thesis would allow to explore the effect of internal interconnect parastitics on highly voltage level sensitive nodes and confirm whether the developed designs would exhibit the same performance in a manufactured chip.

References

[1] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, “Big data: The management revolution,” *Harvard business review*, vol. 90, no. 10, pp. 60–68, 2012.

[2] C. P. Chen and C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on big data,” *Information sciences*, vol. 275, pp. 314–347, 2014.

[3] D. Efnusheva, A. Cholakoska, and A. Tentov, “A survey of different approaches for overcoming the processor-memory bottleneck,” *International Journal of Computer Science and Information Technology*, vol. 9, no. 2, pp. 151–163, 2017.

[4] S. Inaba, “3d flash memory for data-intensive applications,” in *2018 IEEE International Memory Workshop (IMW)*, IEEE, 2018, pp. 1–4.

[5] L. M. Grupp, J. D. Davis, and S. Swanson, “The bleak future of nand flash memory.,” in *FAST*, vol. 7, 2012, pp. 10–2.

[6] S. S. Rizvi and T.-S. Chung, “Flash ssd vs hdd: High performance oriented modern embedded and multimedia storage systems,” in *2010 2nd International Conference on Computer Engineering and Technology*, IEEE, vol. 7, 2010, pp. V7–297.

[7] Y. Li and K. N. Quader, “Nand flash memory: Challenges and opportunities,” *Computer*, vol. 46, no. 8, pp. 23–29, 2013.

[8] K. Parat and C. Dennison, “A floating gate based 3d nand technology with cmos under array,” in *2015 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2015, pp. 3–3.

[9] H. Huh, C. Jeon, C. Yang, *et al.*, “A 64gb nand flash memory with 800mb/s synchronous ddr interface,” in *2012 4th IEEE International Memory Workshop*, IEEE, 2012, pp. 1–4.

[10] J. Cho, D. C. Kang, J. Park, *et al.*, “30.3 a 512gb 3b/cell 7 th-generation 3d-nand flash memory with 184mb/s write throughput and 2.0 gb/s interface,” in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, vol. 64, 2021, pp. 426–428.

[11] K. Kim, “The smallest engine transforming humanity: The past, present, and future,” in *2021 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2021, pp. 1–1.

[12] J.-W. Park, D. Kim, S. Ok, *et al.*, “30.1 a 176-stacked 512gb 3b/cell 3d-nand flash with 10.8 gb/mm 2 density with a peripheral circuit under cell array architecture,” in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, vol. 64, 2021, pp. 422–423.

[13] ONFI, *Open NAND Flash Interface Specifications*. 2022.

[14] V. G. Oklobdzija and R. K. Krishnamurthy, *High-performance energy-efficient microprocessor design*. Springer Science & Business Media, 2007.

[15] S.-H. Lee, “Technology scaling challenges and opportunities of memory devices,” in *2016 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2016, pp. 1–1.

[16] A. Spessot, S. M. Salahuddin, R. Escobar, *et al.*, “Thermally stable, packaged aware lv hkmg platforms benchmark to enable low power i/o for next 3d nand generations,” in *2022 IEEE International Memory Workshop (IMW)*, IEEE, 2022, pp. 1–4.

[17] T. Higuchi, T. Kodama, K. Kato, *et al.*, “30.4 a 1tb 3b/cell 3d-flash memory in a 170+ word-line-layer technology,” in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, vol. 64, 2021, pp. 428–430.

[18] K. Mistry, C. Allen, C. Auth, *et al.*, “A 45nm logic technology with high-k+ metal gate transistors, strained silicon, 9 cu interconnect layers, 193nm dry patterning, and 100% pb-free packaging,” in *2007 IEEE International Electron Devices Meeting*, IEEE, 2007, pp. 247–250.

[19] T. Y. Hoffman *et al.*, “Integrating high-k/metal gates: Gate-first or gate-last?” *Solid State Technology*, vol. 53, no. 3, pp. 20–21, 2010.

[20] D. Triyoso, R. Carter, J. Kluth, *et al.*, “Factors impacting threshold voltage in advanced cmos integration: Gate last (finfet) vs. gate first (fdsoi),” *ECS Transactions*, vol. 69, no. 5, p. 103, 2015.

[21] W. J. Dally, W. J. Dally, and J. W. Poulton, *Digital systems engineering*. Cambridge university press, 1998.

[22] F. Yuan, *CMOS current-mode circuits for data communications*. Springer Science & Business Media, 2007.

[23] K. C. Yong, W. C. Song, B. E. Cheah, and M. F. Ain, "Signaling analysis of inter-chip i/o package routing for multi-chip package," in *2012 4th Asia Symposium on Quality Electronic Design (ASQED)*, IEEE, 2012, pp. 243–248.

[24] B. Razavi, *Fundamentals of microelectronics*. John Wiley & Sons, 2021.

[25] M. N. Reddy and D. K. Panda, "A comprehensive review on finfet in terms of its device structure and performance matrices," *Silicon*, vol. 14, no. 18, pp. 12 015–12 030, 2022.

[26] B. Gunning, L. Yuan, T. Nguyen, and T. Wong, "A cmos low-voltage-swing transmission-line transceiver," in *1992 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, IEEE, 1992, pp. 58–59.

[27] J. M. Rabaey, *Digital integrated circuits a design perspective*. 1999.

[28] T. Granberg *et al.*, *Handbook of digital techniques for high-speed design*. Wharton School Pub, 2005.

[29] Y.-C. Cho, Y.-C. Bae, B.-M. Moon, *et al.*, "A sub-1.0 v 20nm 5gb/s/pin post-lpddr3 i/o interface with low voltage-swing terminated logic and adaptive calibration scheme for mobile application," in *2013 Symposium on VLSI Circuits*, IEEE, 2013, pp. C240–C241.

[30] I. S. Association *et al.*, "Ieee standard for low-voltage differential signals (lvds) for scalable coherent interface (sci)," *IEEE Std 1596.3-1996*, 1996.

[31] A. Boni, A. Pierazzi, and D. Vecchi, "Lvds i/o interface for gb/s-per-pin operation in 0.35-/spl mu/m cmos," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 4, pp. 706–711, 2001.

[32] W.-Y. Tsai, H. Liu, X. Li, and V. Narayanan, "Low-power high-speed current mode logic using tunnel-fets," in *2014 22nd International Conference on Very Large Scale Integration (VLSI-SoC)*, IEEE, 2014, pp. 1–6.

[33] H. H. Muller, W. K. Owens, and P. Verhofstadt, "Fully-compensated emitter-coupled logic: Eliminating the drawbacks of conventional ecl," *IEEE Journal of Solid-State Circuits*, vol. 8, no. 5, pp. 362–367, 1973.

[34] B. Razavi, Y. Ota, and R. G. Swartz, "Design techniques for low-voltage high-speed digital bipolar circuits," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 3, pp. 332–339, 1994.

[35] B. Razavi, *Design of Analog CMOS Integrated Circuits*. McGraw-Hill Education, 2000.

[36] J. Nebhen, M. Masmoudi, W. Rahajandraibe, and K. Aguir, "High dc-gain two-stage ota using positive feedback and split-length transistor techniques," in *Advances in Data Science, Cyber Security and IT Applications: First International Conference on Computing, ICC 2019, Riyadh, Saudi Arabia, December 10–12, 2019, Proceedings, Part II 1*, Springer, 2019, pp. 286–302.

[37] H.-Y. Joo, S.-J. Bae, Y.-S. Sohn, *et al.*, "18.1 a 20nm 9gb/s/pin 8gb gddr5 dram with an nbti monitor, jitter reduction techniques and improved power distribution," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, 2016, pp. 314–315.

[38] I. Sinclair, *Electronics simplified*. Newnes, 2011, Chapter 4.

[39] J. Montanaro, R. T. Witek, K. Anne, *et al.*, "A 160-mhz, 32-b, 0.5-w cmos risc microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1703–1714, 1996.

[40] B. Nikolic, V. G. Oklobdzija, V. Stojanovic, W. Jia, J. K.-S. Chiu, and M. M.-T. Leung, "Improved sense-amplifier-based flip-flop: Design and measurements," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 6, pp. 876–884, 2000.

[41] C.-L. Hsu and M.-H. Ho, "High-speed sense amplifier for sram applications," in *The 2004 IEEE Asia-Pacific Conference on Circuits and Systems, 2004. Proceedings.*, IEEE, vol. 1, 2004, pp. 577–580.

[42] M. S. Akter, K. A. Makinwa, and K. Bult, "A capacitively degenerated 100-db linear 20–150 ms/s dynamic amplifier," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 1115–1126, 2017.

[43] H. W. Johnson, M. Graham, *et al.*, *High-speed digital design: a handbook of black magic*. Prentice Hall Englewood Cliffs, NJ, 1993, vol. 155.

[44] H. Johnson, H. W. Johnson, and M. Graham, *High-speed signal propagation: advanced black magic*. Prentice Hall Professional, 2003.

[45] H. Zhang, F. Che, T. Lin, and W. Zhao, *Modeling, Analysis, Design, and Tests for Electronics Packaging beyond Moore*. Woodhead Publishing, 2019.

[46] M. Steer, *Microwave and RF design*. NC State University, 2019.

[47] C. Robinson, T. Piwtorak, and B. Souid, "Synthesis and analysis of microstrip and stripline transmission line structures,"

[48] I. J. Bahl, "A designer's guide to microstrip line.," 1977.

[49] IPC-2141A, "Design guide for high-speed controlled impedance circuit boards," in *Ipc*, 2004, pp. 16–19.

[50] B. C. Wadell, *Transmission line design handbook*. Artech House Microwave Library, 1991.

[51] Z. Piatek, B. Baron, T. Szczegielniak, D. Kusiak, and A. Pasierbek, "Self inductance of long conductor of rectangular cross section," *Przeglad Elektrotechniczny*, vol. 88, no. 8, pp. 323–326, 2012.

[52] M. N. Sadiku, S. M. Musa, and S. R. Nelatury, "Comparison of approximate formulas for the capacitance of microstrip line," in *Proceedings 2007 IEEE SoutheastCon*, IEEE, 2007, pp. 427–432.

[53] E. Hammerstad and O. Jensen, "Accurate models for microstrip computer-aided design," in *1980 IEEE MTT-S International Microwave Symposium Digest*, IEEE, 1980, pp. 407–409.

[54] L. M. Al-Hadhrami and A. Quddus, "Role of solution hydrodynamics on the deposition of caso4 scale on copper substrate," *Desalination and Water Treatment*, vol. 21, no. 1-3, pp. 238–246, 2010.

[55] J. W. Dally, *Packaging of Electronic Systems: A Mechanical Engineering Approach*. McGraw-Hill, 1990.

[56] M. M. Pajovic, "A closed-form equation for estimating capacitance of signal vias in arbitrarily multilayered pcbs," *IEEE Transactions on electromagnetic Compatibility*, vol. 50, no. 4, pp. 966–973, 2008.

[57] G. Heinrich and S. Dickmann, "Lumped models for vias in multilayered pcbs," in *2009 IEEE International Symposium on Electromagnetic Compatibility*, IEEE, 2009, pp. 33–38.

[58] J. R. Miller, I. Novak, and T.-Y. Chou, "Calculating partial inductance of vias for printed circuit board modeling," in *2002 IEEE 11th Topical Meeting on Electrical Performance of Electronic Packaging*, IEEE, 2002, pp. 123–126.

[59] M. E. Goldfarb and R. A. Pucel, "Modeling via hole grounds in microstrip," *IEEE microwave and guided wave letters*, vol. 1, no. 6, pp. 135–137, 1991.

[60] J. W. Dally, P. Lall, and J. C. Suhling, *Mechanical design of electronic systems*. College House Enterprises, 2008.

[61] G. R. Blackwell, *The electronic packaging handbook*. CRC Press, 2017.

[62] C. R. Paul, *Inductance: loop and partial*. John Wiley & Sons, 2011.

[63] E. B. Rosa, *The self and mutual inductances of linear conductors*. US Department of Commerce and Labor, Bureau of Standards, 1908.

[64] S. H. Hall, G. W. Hall, J. A. McCall, *et al.*, *High-speed digital system design: a handbook of interconnect theory and design practices*. Wiley New York, 2000.

[65] H. Zhang, S. Krooswyk, and J. Ou, *High speed digital design: design of high speed interconnects and signaling*. Elsevier, 2015.

[66] J. G. Maneatis and M. A. Horowitz, "Precise delay generation using coupled oscillators," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 12, pp. 1273–1282, 1993.

[67] Z. Chen and G. Katopis, "A comparison of performance potentials of single ended vs. differential signaling," in *Electrical Performance of Electronic Packaging-2004*, IEEE, 2004, pp. 185–188.

[68] H. B. Bakoglu, *Circuits, interconnections, and packaging for VLSI*. Addison-Wesley, 1990.

[69] K. T. Tang and E. G. Friedman, “Simultaneous switching noise in on-chip cmos power distribution networks,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 10, no. 4, pp. 487–493, 2002.

[70] V. H. Vega-Gonzalez, R. Torres-Torres, and A. S. Sanchez, “Analysis of the electrical performance of multi-coupled high-speed interconnects for sop,” in *2009 52nd IEEE International Midwest Symposium on Circuits and Systems*, IEEE, 2009, pp. 1030–1033.

[71] G. F. Taylor, T. Arabi, H. Greub, R. Muyshondt, A. Manthe, and P. Aminzadeh, “Reliability and performance tradeoffs in the design of on-chip power delivery and interconnects,” in *IEEE 8th Topical Meeting on Electrical Performance of Electronic Packaging (Cat. No. 99TH8412)*, IEEE, 1999, pp. 49–52.

[72] M. Saint-Laurent and M. Swaminathan, “Impact of power-supply noise on timing in high-frequency microprocessors,” *IEEE Transactions on Advanced Packaging*, vol. 27, no. 1, pp. 135–144, 2004.

[73] J. Kim, B. Choi, H. Kim, *et al.*, “Separated role of on-chip and on-pcb decoupling capacitors for reduction of radiated emission on printed circuit board,” in *2001 IEEE EMC International Symposium. Symposium Record. International Symposium on Electromagnetic Compatibility (Cat. No. 01CH37161)*, IEEE, vol. 1, 2001, pp. 531–536.

[74] M. Popovich, A. Mezhiba, and E. G. Friedman, *Power distribution networks with on-chip decoupling capacitors*. Springer Science & Business Media, 2007.

[75] R. R. Tummala, E. J. Rymaszewski, and A. Klopfenstein, “Microelectronics packaging handbook,” *Van*, 1990.

[76] L. D. Smith, R. E. Anderson, D. W. Forehand, T. J. Pelc, and T. Roy, “Power distribution system design methodology and capacitor selection for modern cmos technology,” *IEEE Transactions on Advanced Packaging*, vol. 22, no. 3, pp. 284–291, 1999.

[77] F. Heiman and H. Miiller, “Temperature dependence of n-type mos transistors,” *IEEE Transactions on Electron Devices*, vol. 12, no. 3, pp. 142–148, 1965.

[78] I. Filanovsky and A. Allam, “Mutual compensation of mobility and threshold voltage temperature effects with applications in cmos circuits,” *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 7, pp. 876–884, 2001.

[79] Z. Prijić, S. Dimitrijev, and N. Stojadinović, “The determination of zero temperature coefficient point in cmos transistors,” *Microelectronics Reliability*, vol. 32, no. 6, pp. 769–773, 1992.

[80] F. Salehuddin, I. Ahmad, F. Hamid, A. Zaharim, U. Hashim, P. Apte, *et al.*, “Optimization of input process parameters variation on threshold voltage in 45 nm nmos device,” *International Journal of the Physical Sciences*, vol. 6, no. 30, pp. 7026–7034, 2011.

[81] A. A. Mutlu and M. Rahman, “Statistical methods for the estimation of process variation effects on circuit operation,” *IEEE Transactions on Electronics Packaging Manufacturing*, vol. 28, no. 4, pp. 364–375, 2005.

[82] X. Bai, P. Patel, and X. Zhang, “A new statistical setup and hold time definition,” in *2012 IEEE International Conference on IC Design & Technology*, IEEE, 2012, pp. 1–4.

[83] R. Gonzalez, B. M. Gordon, and M. A. Horowitz, “Supply and threshold voltage scaling for low power cmos,” *IEEE Journal of Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, 1997.

[84] M. P. Li, *Jitter, noise, and signal integrity at high-speed*. Pearson Education, 2007.

[85] K. S. D. Oh and X. C. C. Yuan, *High-Speed Signaling: Jitter Modeling, Analysis, and Budgeting*. Prentice Hall, 2011.

[86] A. Athavale and C. Christensen, “High-speed serial i/o made simple,” *Xilinx inc*, vol. 4, 2005.

[87] N. H. Weste and D. Harris, *CMOS VLSI design: a circuits and systems perspective*. Pearson Education India, 2015.

[88] S. De, R. Das, R. K. Varshney, and T. Schneider, “Design and simulation of thermo-optic phase shifters with low thermal crosstalk for dense photonic integration,” *IEEE Access*, vol. 8, pp. 141 632–141 640, 2020.

[89] K. Joardar, “A simple approach to modeling cross-talk in integrated circuits,” *IEEE Journal of Solid-State Circuits*, vol. 29, no. 10, pp. 1212–1219, 1994.

[90] A. Deutsch, H. H. Smith, C. W. Surovic, *et al.*, “Frequency-dependent crosstalk simulation for on-chip interconnections,” *IEEE transactions on advanced packaging*, vol. 22, no. 3, pp. 292–308, 1999.

[91] A. R. Chada, *Modeling and estimation of crosstalk across a channel with multiple, non-parallel coupling and crossings of multiple aggressors in practical PCBS*. Missouri University of Science and Technology, 2014.

[92] D. Norte, “Near-end crosstalk considerations for coupled microstriplines,” 2011.

[93] T. Cabara, W. C. Fischer, J. Harrington, and W. W. Troutman, “Forming damped lrc parasitic circuits in simultaneously switched cmos output buffers,” *IEEE Journal of Solid-State Circuits*, vol. 32, no. 3, pp. 407–418, 1997.

[94] R. Downing, P. Gebler, and G. Katopis, “Decoupling capacitor effects on switching noise,” *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, vol. 16, no. 5, pp. 484–489, 1993.

[95] K. Bock, G. Groeseneken, and H. Maes, “Esd protection methodology for deep-sub-micron cmos,” *Microelectronics Reliability*, vol. 38, no. 6-8, pp. 997–1007, 1998.

[96] M.-D. Ker and C.-Y. Chang, “Esd protection design for cmos rf integrated circuits using polysilicon diodes,” *Microelectronics Reliability*, vol. 42, no. 6, pp. 863–872, 2002.

[97] S. Salas, J. C. Tinoco, A. G. Martinez-Lopez, J. Alvarado, and J.-P. Raskin, “Fringing gate capacitance model for triple-gate finfet,” in *2013 IEEE 13th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems*, IEEE, 2013, pp. 90–92.

[98] S. Jimenez, A. Lemmon, B. Nelson, and B. Deboi, “Comprehensive characterization of mosfet intrinsic capacitances,” in *2021 IEEE Applied Power Electronics Conference and Exposition (APEC)*, IEEE, 2021, pp. 1524–1530.

[99] Y. Yan and T. H. Szymanski, “Low power high speed i/o interfaces in 0.18/spl mu/m cmos,” in *10th IEEE International Conference on Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003*, IEEE, vol. 2, 2003, pp. 826–829.

[100] M. Hebali, D. Berbara, M. A. Abid, *et al.*, “An ultra low power and high performance of cmos (6h-sic) current mirrors in bsim3v3 130nm technology,”

[101] R. W. Beatty, “Insertion loss concepts,” *Proceedings of the IEEE*, vol. 52, no. 6, pp. 663–671, 1964.

[102] J. Coonrod, “The effect of radiation losses on high frequency pcb performance,” *IPC Apex Expo*, 2014.

[103] J. Coonrod, “Insertion loss comparisons of common high frequency pcb constructions,” *IPC APEX EXPO*, 2013.

[104] G. Boccardi, R. Ritzenthaler, M. Togo, *et al.*, “Rmg technology integration in finfet devices,” 2012.

[105] L.-Å. Ragnarsson, Z. Li, J. Tseng, *et al.*, “Ultra low-eot (5 Å) gate-first and gate-last high performance cmos achieved by gate-electrode optimization,” in *2009 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2009, pp. 1–4.

[106] H. R. Harris, P. Kalra, P. Majhi, *et al.*, “Band-engineered low pmos v t with high-k/metal gates featured in a dual channel cmos integration scheme,” in *2007 IEEE Symposium on VLSI Technology*, IEEE, 2007, pp. 154–155.

[107] C. Lai, C. Lin, L. Cheng, *et al.*, “A novel “hybrid” high-k/metal gate process for 28nm high performance cmosfets,” in *2009 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2009, pp. 1–4.

[108] C. Auth, A. Cappellani, J.-S. Chun, *et al.*, “45nm high-k+ metal gate strain-enhanced transistors,” in *2008 Symposium on VLSI Technology*, IEEE, 2008, pp. 128–129.

[109] K. Choi, H. Jagannathan, C. Choi, *et al.*, “Extremely scaled gate-first high-k/metal gate stack with eot of 0.55 nm using novel interfacial layer scavenging techniques for 22nm technology node and beyond,” in *2009 Symposium on VLSI Technology*, IEEE, 2009, pp. 138–139.

[110] E. Capogreco, H. Arimura, R. Ritzenthaler, *et al.*, “Finfets with thermally stable rmg gate stack for future dram peripheral circuits,” in *2022 International Electron Devices Meeting (IEDM)*, IEEE, 2022, pp. 26–2.

[111] N. Waldron, C. Merckling, W. Guo, *et al.*, “An ingaas/inp quantum well finfet using the replacement fin process integrated in an rmg flow on 300mm si substrates,” in *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, IEEE, 2014, pp. 1–2.

[112] J. H. Lee, D.-G. Kim, H.-J. Lee, and C. S. Hwang, “Fabrication of a nano-scaled tri-gate field effect transistor using the step-down patterning and dummy gate processes,” *Microelectronic Engineering*, vol. 173, pp. 33–41, 2017.

A. Optimum Power Design of an 8 Gb/s NAND I/O Interconnect in 14 nm Fin-FET Technology

This appendix contains the paper that is submitted to IEEE Journal of Solid-State Circuits (JSSC), 26th October 2023.

Optimum Power Design of an 8 Gb/s NAND I/O Interconnect in 14 nm FinFET Technology

Arvis Jomerts[1,2,*], Nicolas Pantano[1], Ivick Guerra Gomez[1], Kavitha Soundra Pandiyan[1], Romain Ritzenthaler[1], Naoto Horiguchi[1], Said Hamdioui[2], Georgi Gaydadjiev[2], Alessio Spessot[1]
[1]*Interuniversitair Micro-Electronica Centrum vzw (imec)*, Belgium, [2]*Technische Universiteit Delft (TU Delft)*, Netherlands
*Corresponding Author (email: A.Jomerts@student.tudelft.nl / arvis.jomerts1999@gmail.com)

***Abstract*—Periphery area of CMOS under array manufactured 3D NAND memory must remain low to obtain a market competitive product, and it must be thermally stable to survive the entire thermal annealing process. Replacing standard planar logic devices with a smaller alternative such as FinFET is a promising solution to substantially improve Input/Output (I/O) performance of NAND memory chips without exceeding area limits. Thus, this work presents an 8 Gb/s I/O interconnect developed in imec's thermally stable 14 nm FinFET technology equivalent.**

A methodology to benchmark transmitter (TX) active area footprint and system power for different transistors across various topologies is proposed. Signal degradation mechanisms such as crosstalk and noise were considered; the largest system parasitic contributors - packaging and transmission line - were analysed. Throughput is validated using industry requirements. System parameters, to which the design is most sensitive, are identified and discussed. It is concluded that FinFET TX active area footprint is several times lower than equivalent thermally stable planar 45 nm design for iso-performance conditions. Research presented here establishes a baseline for technology scaling in NAND I/O and proposes guidelines for suitable topology selection depending on system constraints.

***Index Terms*—high-speed, I/O, FinFET, NAND, Flash**

I. Introduction

Rapid generation of industry valued data in vast quantities (Big Data) poses several challenges for the conventional computer architecture - not only the data storage capacity has to increase, but also the processing speed and transmission bandwidth (BW) between components must improve [1]. To compensate for the growing bottleneck in processor-memory BW, memory cell signal parallelization has to be intensified. Thereby, Input/Output (I/O) supported Data Rate (DR) becomes the limiting factor if no design adaptations are implemented [2]. Therefore, to increase the overall data processing ability of the system, I/O speed has to be enhanced by upgrading the design [3].

3D NAND flash memory is one of the most appealing candidates to alleviate Big Data posed challenges as it provides relatively high memory density and low latency when used as the main memory module [4]. Current 3D NAND I/O made using CMOS under array (CuA) manufacturing flow reaches DR up to 2.4 Gb/s [5] using planar technology. However, as CuA chip logic is located under memory cells, its area cannot be directly increased for a standard size chip without reducing memory density. Hence, a need to advance to a lower technology node emerges [3]. The periphery transistor also has to have mature manufacturing flow: as memory cells are located atop the periphery in CuA technique, logic is manufactured first. Thus, I/O circuitry has to endure the complete chip thermal annealing, inferring a requirement for thermal stability [6].

Considering the above, the aim of this work is to develop a 3D NAND compatible I/O able to reach 8 Gb/s DR, using thermally stable 14 nm technology FinFET equivalent [7]–[9]. The main contributions provided by this work are as follows:

- Evaluation of thermally stable 14 nm FinFET device performance in high-speed, temperature sensitive NAND I/O application.
- Determination of area benefits provided by FinFET devices in comparison to thermally stable 45 nm planar technology for iso-performance conditions.
- Assessment of FinFET based I/O design area sensitivities to system parameters and environmental conditions as DR, transmission line length and voltage.
- Development of optimum power I/O interconnect able to achieve 8 Gb/s BW using 2 carefully designed transmission topologies simulated in realistic environmental conditions.

Since there is currently no existing thermally stable FinFET suitable for 3D NAND applications, the current best alternative was used, assuming that future technology will have exactly the same characteristics [10].

This paper is organized in the following manner. Sec II briefly addresses the challenges of high-speed, FinFET based I/O design and discusses the chosen signaling topologies and its components. Here, also the imposed signal limitations and throughput validation methodology can be found. In Sec. III the 8 Gb/s designs are provided and discussed. System sensitiviete are elaborated and evaluated in Sec. IV. Lastly, the results found in this paper are concluded in Sec V.

II. Circuit Setup and Simulation Benchmark

An investigation into the challenges inherent to moving to a lower node and higher DR need to be addressed before a design analysis can be carried out. First, higher DR causes more cross-induced noise in adjacent signal lines - with hold-to-rise time ratio being constant and bit width reducing, higher slew rate transitions occur [12]. Hence, signal fluctuations are bound to increase as seen in Eq. (1), where L_{mut} is mutual inductance and $\frac{dI}{dt}$ is instantaneous current change [13].

$$\Delta V = L_{mut}\frac{dI}{dt} \tag{1}$$

Second, the use of smaller transistors leads to higher density routing, which further increases crosstalk effects [14]. Lastly, FinFET technology uses a lower power supply voltage (V_{dd}) and leads to overall lower intrinsic circuit capacitance, implying that the system is more susceptible to noise and crosstalk [15]. To counteract the aforementioned drawbacks, high signal swing must be attained compared to planar, low-speed designs, which comes at the cost of higher power consumption.

A. Toplogy Overview and Circuit Architecture

The memory-to-processor interconnect is established using conventional point-to-point signaling schemes, where the main novelty is the use of an advanced technology node. Exploration of multi-point operations is left for future investigation, as design adaptations would be required to cope with additional signal reflections causing higher signal quality degradation in the likes of inter-symbol interference [13].

To ensure unbiased benchmarking of different transistors, one Single Ended Signaling (SES) and one Differential Signaling (DS) topology is used in the analysis. As memory and processor are separate entities, the overall system consists of five main components: Transmitter (TX), Receiver (RX), Packaging (PACK), external Transmission Line (TL) and Termination (TERM) - see Fig. 1.

SES is implemented using Centre Tapped Termination (CTT) toplogy [16] seen in Fig. 1. Its TX is composed of a growing inverter chain, which ensures that the last TX stage can drive the high cumulative parasitic capacitance generated by TL and PACK. Double parallel TERM is used on the load side to strongly set the output common mode voltage (V_{cm}) to half V_{dd}. Set V_{cm} also imposes relatively high next stage drive strength and symmetric '1' and '0' signal characteristics. However, the drawback of CTT TERM is high static power dissipation due to a continuous current path between the power and the ground rail. Lower power consuming SES toplogies are available, nevertheless, their elevated susceptibility to noise, higher process, voltage and temperature dependence, unbalanced '0' and '1' signal power in combination with next stage drive strength inferiority to CTT are the main drawbacks when relatively high load capacitance has to be driven [5].

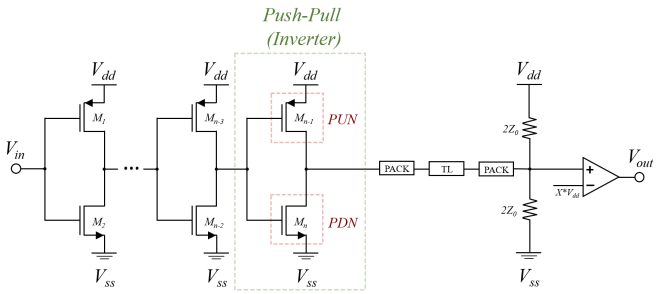


Fig. 1: Schematic of Proposed CTT Topology

To reduce total power consumption and improve noise and crosstalk immunity, DS topology such as Scalable Low-Voltage Signaling (SLVS) seen in Fig. 2 is proposed. It is worth noting that SLVS is a simplified version of the Low-Voltage Differential Signaling (LVDS) [17] standard, where the current imposing source at the tail is removed. Here, floating TERM is used to generate the next stage voltage swing, which reduces power consumption, but leads to drifting V_{cm}. Buffers were added to the design to convert amplitude noise into jitter such that ringing on TL is prevented. Otherwise, low frequency noise can partially pass through the TX (H-bridge) due to Miller effect. Nonetheless, use of a small buffer which size can hardly be adapted in the discrete sizing convention of FinFET leads to signal skewing post H-bridge if PFET and NFET effective currents differ. For simplicity, passive TERM is used for both CTT and SLVS designs. Investigation of active on-die TERM should be performed as one of the next steps in the analysis such that signal quality improvements due to adaptive impedance matching can be quantified.

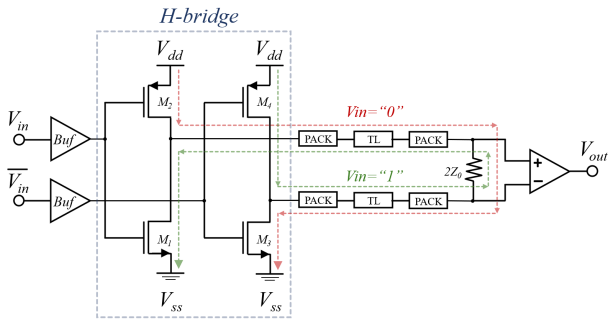


Fig. 2: Schematic of Proposed SLVS Topology

As throughput validation takes place at the input of RX [11], emulation of RX amplifier is not strictly required. Only load mimicking circuitry has to be added to ensure non-open output terminals. Moreover, RX input terminal capacitance is negligible in comparison to TX and PACK parasitics. Hence, simple 5 transistor Operational Transconductance Amplifier (OTA) [18] is used as RX due its applicability to both SES and DS. The used OTA schematic can be seen in Fig. 3

1) Transmission Line Model: TL implementation is performed using *Cadence® Virtuoso®* 2D field solver entities. Microstrip line is used as the underlying TL type of printed circuit board to be used. This imposes worst-case signal attenuation and degradation conditions. This guarantees that no limitations are set on the electrical engineers using the product under development. Moreover, the characteristic impedance (Z_0) is set to be 50 Ω to obtain the best attenuation and power efficiency trade-off, while simultaneously simplifying the system implementation with standard components.

Crosstalk is included by employing adjacent aggressor lines transmitting different bit sequences. To obtain worst-case conditions, artificial delay of 7% and 11% bit width is forced on the aggressor signals. Guard lines (grounds) are used in-between two signal paths to limit system analysis to 3 active

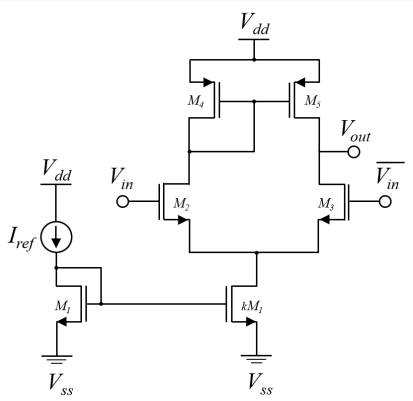


Fig. 3: 5 transistor OTA schematic

lines of interest [19]. The length for which traces run in parallel is set to be 4 cm (also TL total length) assuming the use of centre-most pins and full chip width spacing between the memory and the processor [11].

2) High-speed Applicable Packaging: To limit package parasitics, Flip Chip (FC) Ball Grid Array (BGA) package is used as shown in Fig. 4. In FC no bondwires are required, reducing path inductance which inhibits high-speed signal. Use of both internal and external BGA eliminates highly parasitic components (long pins or thin metallic pads) of surface mounted devices. FC packaging can be simplified to a three component model, which consists of BGA, trace and via, with mirror implementation around the via.

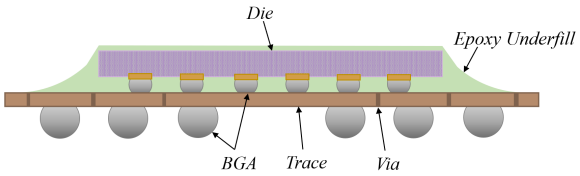


Fig. 4: FC packaging using both internal and external BGA

B. Signal Limitations

Both amplitude and time (jitter) noise have been introduced in the system by superimposing them on the input signal. Jitter was incorporated by first feeding an ideal signal into slew rate transformer made in VerilogA. Then, the signal is further passed to an inverter chain which performs the slew to jitter conversion, causing 16% unit interval (equal to ideal bit width (UI)) spread. Inverter chain output also limits the input drive strength capabilities guaranteeing more realistic system behaviour. Noise is added such that instantaneous fluctuation peaks reach $0.3V_{dd}$ and average rail variation equates to 10% of V_{dd}. The input signal comprising of the above modifications looks as shown in Fig. 5.

C. Throughput Validation

Throughput is validated using a Compliance Eye Mask (CEM) centered in the signal eye as shown in Fig. 6. The

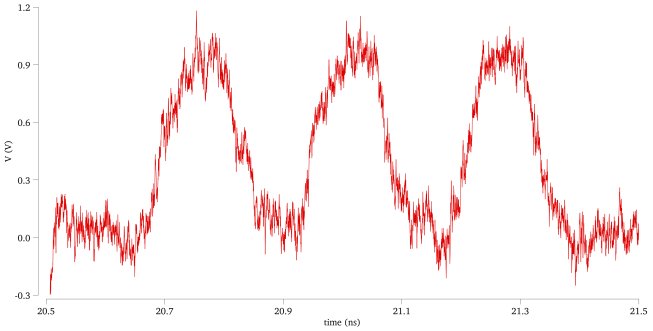


Fig. 5: System's input signal with limited drive strength after amplitude noise and jitter supper-imposition

minimum requirements are defined to be 200 mV voltage swing for a continuous time of 0.3 UI and a total opening of 0.5 UI; slew rate can be determined from the above [11]. The CEM shape is independent of DR as signal requirement fractions are constant - the only variable is UI. Optimal sizing is reached when CEM requirements are exactly met.

Chosen simulation length is lower than the required time to verify standard Bit Error Rate (BER) of 10^{-9}. However, as the same simulation time was applied for all design cases, it is assumed that the margin by which area has to be increased to comply with BER requirement is approximately the same for every system setup. Thereby, the acquired active area footprint ratios between the designs are deemed reliable.

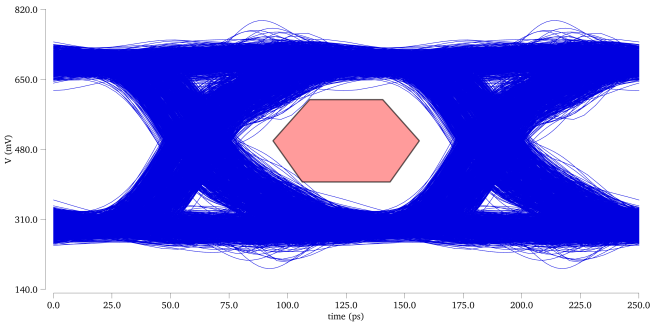


Fig. 6: Data throughput validation at RX input using CEM

III. Target Case Results and Discussion

For simulations, two different manufacturing flow devices were used: Replacement Metal Gate (RMG) and Gate First (GF). The gate of RMG devices is not thermally annealed, leaving the gate oxide unaffected resulting in a superior output drive strength. TX active area results for various device types can be seen in Fig. 7, where PL refers to planar RMG device.

It is worth noting that TX active area footprint can be directly compared in this particular case, as the same V_{dd} value of 1V for FinFET and planar 45 nm devices was used. With shared V_{dd}, total power must be equivalent among different device designs for the same topology - required current flow is approximately the same. For instance, all CTT designs

consume ≈8 mW, while SLVS dissipates ≈4.3 mW (1.86 times lower than CTT). 1.2 mW of the power is consumed by the RX, hence, SLVS to CTT topology power consumption ratio is actually 2.2.

The use of a 1V power supply represents characterization conditions for the FinFET compact model, ensuring highest accuracy comparison. Nonetheless, it leads to substantial over-voltage, which would significantly limit the lifespan of TX circuitry [20].

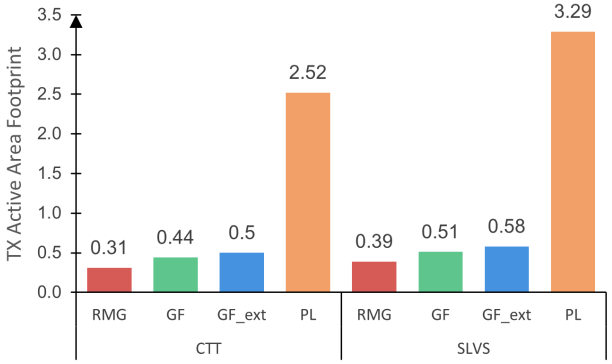


Fig. 7: Active Area Footprint of CTT and SLVS designs for various device types: replacement metal gate (RMG), gate first (GF), GF extended (GF_{ext}) and planar RMG (PL)

Fig. 7 shows that RMG performs better than GF and GF_{ext} by 42% and 61% in CTT, while the ratio drops by 12% for SLVS topology. The discrepancy in area is caused by SLVS sensitivity to PFET and NFET effective current differences - for RMG devices a slightly higher gap is present than in GF devices. Thus, pull-down (or pull-up in case an inverting buffer is used) transistors have to be sized larger such that V_{cm} is set at half V_{dd} and the next stage drive strength is balanced. Note, GF_{ext} refers to GF devices with extended channel length due to manufacturing adjustments.

When comparing planar 45 nm RMG designs against FinFET RMG, a staggering ≈8 times reduction for both CTT and SLVS was observed. The scaling factor is representative for the setup used and is composed of the following subfactors: 1.7^2 comes from device width (W) and length (L) scaling as $\frac{W}{L}$ remains roughly the same; 1.5 is caused by elimination of charge mobility differences in PFET and NFET; a factor of ≈2 comes from the 3D fin implementation. For the nominal V_{dd} cases, area gains from switching to FinFET would drop to ≈4-5 times, while power savings would equate to 20%. FinFET nominal V_{dd} is 0.8V.

TX active area footprint of SLVS in comparison to CTT is only 1.16-1.26 times larger, implying that the ratio has significantly reduced from conventional scaling of 2, which is observed for low-speed systems. Therrfore, irrespective of whether nominal or characterization V_{dd} has been used, SLVS is more superior topology at DR of 8 Gb/s than CTT, since its area-power product is roughly 2 times lower. However, the factor of 2 only relates to TX active footprint area as total area ratio is likely to increase due to necessity of 2 external signal paths for DS.

IV. Sensitivity Analysis

As signal requirements are just met for target case design, rather than looking at performance shifts upon variations in system parameters, area compensation/relaxation to keep iso-performance conditions was investigated instead. Target case results were used as reference for benchmarking.

A. Design Sensitivity against Data Rate

DR sweep range is selected to be ∈[2, 8] Gb/s as shown in Fig. 8 - this enables a comparison of obtained results with current state-of-the-art solutions. As SLVS is less susceptible to noise and crosstalk, it was predicted to exhibit linear dependence of TX active area footprint with respect to DR, while CTT should increase non-linearly [12]. Both claims were verified as shown in Fig. 8. As TX active area is directly proportional to the required current (by factor $\frac{W}{L}$), power curve trends are alike to area ones and thus is omitted here. Only the total power of RMG based designs is provided in Fig.9 b), where the gap between CTT and SLVS is seen to grow steadily.

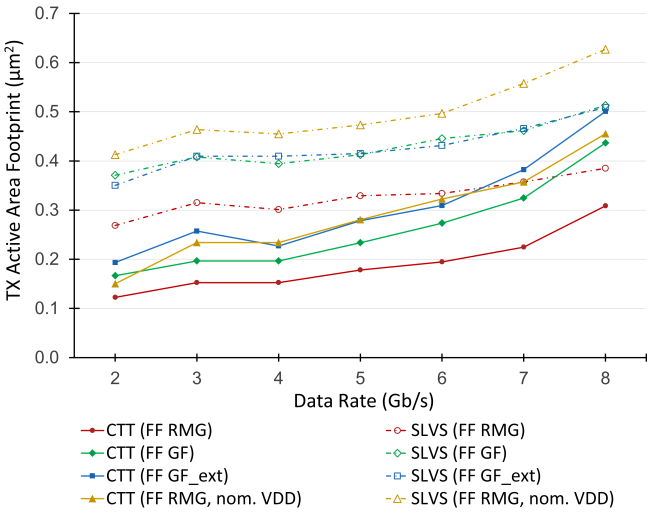


Fig. 8: Active area footprint of TX vs Data Rate

The gap between areas of CTT and SLVS designs is largely negated when DR increases several times. For instance, the area ratio reduces from a factor of 2.25 to a mere 1.26 when comparing RMG device 2 Gb/s and 8 Gb/s design scenarios. Thereby, it comes as no surprise that SES is used as the conventional signal topology for current commercial products reaching 2 Gb/s.

Fig. 9 a) depicts a normalized increase of area for both SLVS and CTT in planar and FinFET technologies. Here, it is clearly visible that SLVS will catch up to CTT for DR slightly beyond 8 Gb/s asserting full dominance over CTT in both active area footprint and power. One can conclude that CTT will fail to provide sufficient quality signal processing at

DR slightly above 8 Gb/s, since its TX active area tends to grow more rapidly for high-speed systems.

The sizing technique applied in this work is validated since planar and FinFET area characteristics almost overlap as shown in Fig. 9 a). Both devices are from FET family thus should provide comparable results when a common reference point is used. Also, the relative ratio between various device designs with respect to DR remains largely constant except for CTT RMG, which can be attributed to a very high degree of optimisation for the design in question.

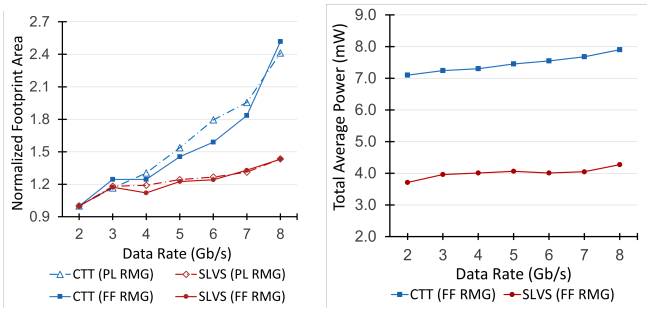


Fig. 9: a) Normalized TX active area footprint vs DR for FinFET (FF) and Planar (PL) replacement metal gate (RMG) devices b) Total power variation vs Data Rate for FinFET devices

B. Design Sensitivity against Transmission Line Length

Similarly as for DR sweep, CTT is predicted to suffer more from variations in TL length due to higher amount of crosstalk induced on the signal line [12]. This was confirmed as seen in Fig. 10 a), where CTT is depicted to exceed SLVS area at 5.5 cm TL length. At 5.5 cm, study of CTT has been discontinued since the design reaches limit length $\in$[6, 7] cm where no comprehensible output is achieved for active area footprint twice the size of SLVS. TX area growth of SLVS remains roughly linear even beyond 7 cm range due its high immunity to noise and crosstalk as it is directly proportional to L_{mut} increase.

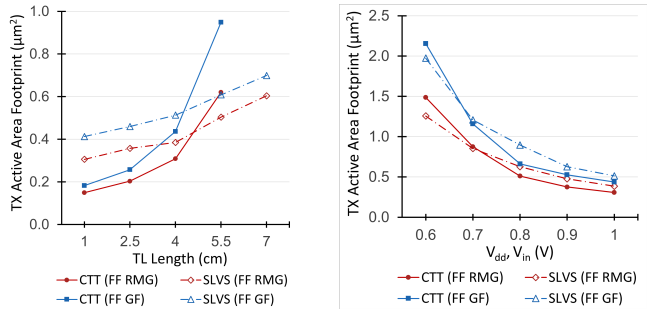


Fig. 10: TX active area footprint variation due to a) TL length sweep b) voltage shift for CTT and SLVS (FinFET)

C. Design Sensitivity against Voltage

Voltage variation can be realized in 3 different ways: case 1 is equivalent technology scaling where both V_{dd} and input voltage (V_{in}) are scaled; case 2 represents V_{in} skewing and case 3 corresponds to V_{dd} skewing where only the respective voltages are changed. Area variations for equivalent technology scaling are provided in Fig. 10 b), where a quadratic trend is expected for both CTT and SLVS designs. However, it can be immediately noticed that SLVS actually benefits from going to lower voltages in comparison to CTT. This is caused by SLVS design specifics: with reducing V_{dd}, H-bridge transistors that are usually located in the linear region tend to shift towards saturation conditions, since V_{cm} reduces at half the rate V_{dd} does. This in turn increases transition slew steepness pushing the signal eye further from CEM, if TERM is increased at the same time. Thereby, current flow can be slightly lowered for higher optimisation causing lower impact on the area increase. A similar approach does not work for CTT as parallel termination is used, implying a different RX input impedance transformation.

Average dynamic power trends also perfectly indicate that SLVS benefits from lowered system voltage. When looking at Fig. 11 a), the slope of SLVS power trend is steeper than that of CTT if equivalent technology scaling is performed. Linear dynamic power trends are expected, since $P_{dyn} \propto V_{dd} \cdot I_d$, where P_{dyn} is the dynamic power and I_d is the drain current of NFET transistor.

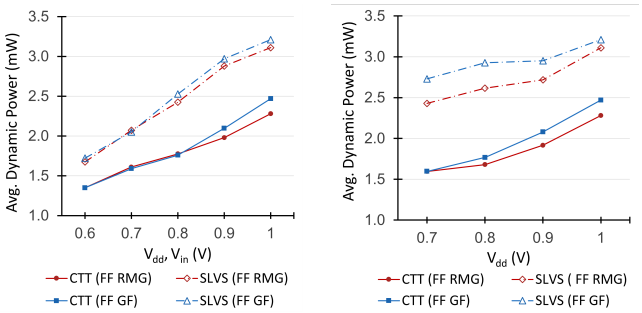


Fig. 11: Dynamic power shifts due to variation in a) V_{in} and V_{dd} b) V_{dd} for CTT and SLVS (FinFET)

When V_{in} or V_{dd} is varied separately, the situation reverses. Due to floating output V_{cm}, SLVS design suffers significant signal skewing as indicated in Fig. 12. As buffer pull-up and pull-down strength can be hardly balanced due to discrete W of FinFET, drive strength incompatibilities are propagated to the H-bridge which cannot fully rectify them. CTT is less susceptible to voltage skewing due to strongly set output V_{cm} and significant number of TX stages which tends to rebalance the signal.

The SLVS's requirement for higher current consumption if only V_{dd} undergoes changes can be deduced from the average dynamic power trend for 3rd case depicted in Fig. 11 b). CTT power on the other hand sees barely any fluctuations when comparing case 1 vs case 3. Thereby, for V_{dd} skewing SLVS

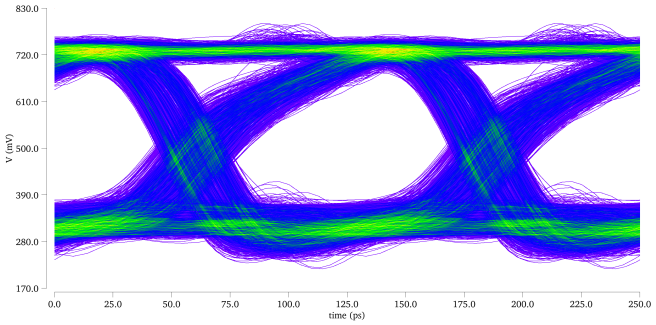


Fig. 12: SLVS signal eye skew when V_{in} is lowered by 300 mV

to CTT TX active area footprint ratio actually becomes worse when compared to target case results - see Fig. 13 a).

Even for relatively large changes in V_{in} almost no variations in either area or power were predicted, since only the first stage signal characteristics were altered. When looking at Fig. 13 b), CTT area is seen to be relatively constant for the entire simulation range. SLVS, on the other hand, exhibits deviations at even low variations of V_{in} due to both floating TERM and lack of stages for signal correction. Therefore, voltage variation results are inconclusive and case specific, since neither of the topologies can be said to be distinctly superior.

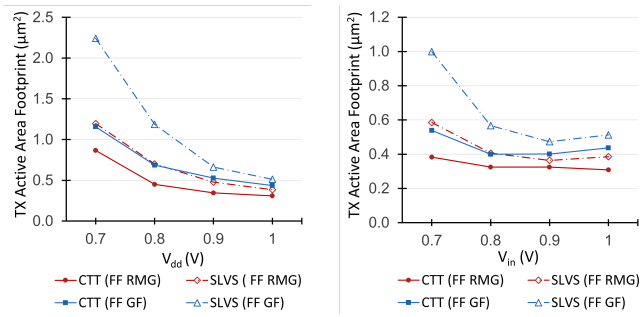


Fig. 13: TX active area footprint variation due to changes in a) V_{in} b) V_{dd} for CTT and SLVS (FinFET)

V. Conclusions

In this work a 3D NAND compatible I/O reaching 8 Gb/s was designed in 14 nm FinFET technology. It was found that both SES and DS schemes employing 14nm FinFET RMG devices lead to a ≈8 times reduction of TX active area footprint in comparison to thermally stable planar 45 nm RMG designs in iso-performance conditions and with shared V_{dd}. Also, designs using RMG devices attain 35% lower TX active area footprint than GF device implementations on average. Sensitivity analysis investigation showed that CTT cannot meet the defined signal quality requirements past 5.5 cm TL length, where it also exceeded SLVS size. A DR sweep perfectly indicated why SLVS RMG topology is the preferred option for 8 Gb/s NAND I/O as SLVS provides 1.74 times lower area-power product than CTT. The DR sweep also confirmed why SES is the current 2 Gb/s I/O standard. Lastly, voltage variation exploration returned case specific results, which have to be investigated further to determine the exact impact of number of cascaded TX stages.

References

[1] D. Efnusheva, A. Cholakoska, and A. Tentov, "A Survey of Different Approaches for Overcoming the Processor-memory Bottleneck," International Journal of Computer Science and Information Technology, vol. 9, no. 2, pp. 151–163, 2017.

[2] C.L.P Chen and C.Y Zhang. "Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data." Information sciences 275 (2014): 314-347.

[3] S. Inaba, "3D Flash Memory for Data-intensive Applications," in 2018 IEEE International Memory Workshop (IMW), IEEE, 2018, pp. 1–4.

[4] S. S. Rizvi and T.-S. Chung, "Flash SSD vs HDD: High Performance Oriented Modern Embedded and Multimedia Storage Systems," in 2010 2nd International Conference on Computer Engineering and Technology, IEEE, vol. 7, 2010, pp. V7–297.

[5] J. Cho, D. C. Kang, J. Park, et al., "30.3 a 512Gb 3b/Cell 7 th-Generation 3D-NAND Flash Memory with 184MB/s Write Throughput and 2.0 Gb/s Interface," in 2021 IEEE International Solid-State Circuits Conference (ISSCC), IEEE, vol. 64, 2021, pp. 426–428.

[6] K. Parat and C. Dennison, "A Floating Gate Based 3D NAND Technology with CMOS Under Array," in 2015 IEEE International Electron Devices Meeting (IEDM), IEEE, 2015, pp. 3–3.

[7] A. Spessot, et al. "80nm Tall Thermally Stable Cost Effective FinFETs for Advanced DRAM Periphery Devices for AI/ML and Automotive Applications." 2021 Jpn. J. Appl. Phys. 60 SBBB06.

[8] E. Capogreco et al., "FinFETs with Thermally Stable RMG Gate Stack for Future DRAM Peripheral Circuits," 2022 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2022, pp. 26.2.1-26.2.4.

[9] R. Ritzenthaler et al., "High Performance Thermally Resistant FinFETs DRAM Peripheral CMOS FinFETs with VTH Tunability for Future Memories," 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022, pp. 306-307.

[10] A. Spessot, S. M. Salahuddin, R. Escobar, et al., "Thermally Stable, Packaged Aware LV HKMG Platforms Benchmark to Enable Low Power I/O for Next 3D NAND Generations," in 2022 IEEE International Memory Workshop (IMW), IEEE, 2022, pp. 1–4.

[11] ONFI, Open NAND Flash Interface Specifications. 2022

[12] K. C. Yong, W. C. Song, B. E. Cheah, and M. F. Ain, "Signaling Analysis of Inter-chip I/O Package Routing for Multi-chip Package," in 2012 4th Asia Symposium on Quality Electronic Design (ASQED), IEEE, 2012, pp. 243–248.

[13] S. H. Hall, G. W. Hall, J. A. McCall, et al., High-speed Digital System Design: A Handbook of Interconnect Theory and Design Practices. Wiley New York, 2000.

[14] A. Athavale and C. Christensen, "High-speed Serial I/O Made Simple," Xilinx inc, vol. 4, 2005.

[15] M. Saint-Laurent and M. Swaminathan, "Impact of Power-supply Noise on Timing in High-frequency Microprocessors," IEEE Transactions on Advanced Packaging, vol. 27, no. 1, pp. 135–144, 2004.

[16] F. Yuan, CMOS Current-Mode Circuits for Data Communications. Springer Science & Business Media, 2007.

[17] I. S. Association et al., "IEEE Standard for Low-voltage Differential Signals (LVDS) for Scalable Coherent Interface (SCI)," IEEE Std 1596.3-1996, 1996.

[18] B. Razavi, Design of Analog CMOS Integrated Circuits. McGraw-Hill Education, 2000

[19] V. H. Vega-Gonzalez, R. Torres-Torres, and A. S. Sanchez, "Analysis of the Electrical Performance of Multi-coupled High-speed Interconnects for SoP," in 2009 52nd IEEE International Midwest Symposium on Circuits and Systems, IEEE, 2009, pp. 1030–1033.

[20] G. F. Taylor, T. Arabi, H. Greub, R. Muyshondt, A. Manthe, and P. Aminzadeh, "Reliability and Performance Tradeoffs in the Design of on-chip Power Delivery and Interconnects," in IEEE 8th Topical Meeting on Electrical Performance of Electronic Packaging (Cat. No. 99TH8412), IEEE, 1999, pp. 49–52.