HIGH-SPEED GLOBAL ON-CHIP INTERCONNECTS AND TRANSCEIVERS

Eisse Mensink

Samenstelling promotiecommissie:

Voorzitter:	prof. dr. C. Hoede	Universiteit Twente, EWI
Secretaris:	prof. dr. ir. A. J. Mouthaan	Universiteit Twente, EWI
Promotor:	prof. ir. A. J. M. van Tuijl	Universiteit Twente, EWI
Assistent Promotor:	dr. ing. E. A. M. Klumperink	Universiteit Twente, EWI
Referent:	dr. ir. J. P. M. van Lammeren	NXP, Nijmegen
Leden:	prof. dr. ir. B. Nauta	Universiteit Twente, EWI
	prof. dr. ir. W. Dehaene	KU Leuven
	prof. dr. ir. R. H. J. M. Otten	TU Eindhoven
	prof. dr. ir. G. J. M. Smit	Universiteit Twente, EWI

HIGH-SPEED GLOBAL ON-CHIP INTERCONNECTS AND TRANSCEIVERS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. W.H.M. Zijm,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 28 juni 2007 om 13.15 uur

door

Eisse Mensink
geboren op 10 januari 1979
te Almelo

Dit proefschrift is goedgekeurd door:

de promotor prof. ir. A. J. M. van Tuijl

de assistent promotor dr. ing. E. A. M. Klumperink

"of making many books there is no end; and much study is a weariness of the flesh."

Ecclesiastes 12:12 – The Bible

Contents

Chapter 1

Introduction

1.1 Communication over long distances

Communication is everywhere. People talk, write, use sms and msn, while animals warn each other for predators or try to attract a suitable partner. However, the success of communication can depend heavily on the distance between the communicators. As an illustration, imagine two people that are talking to each other. When close by, they are able to hear each other. However, if the distance between them increases, it will be harder for them to understand each other. At a certain distance they will have to shout and even further separated they cannot hear each other anymore. Then, other means are needed to talk with each other, for instance a telephone. The distance is simply too large for normal communication.

There is also communication between computers and many other electronic devices and also between different parts of an integrated circuit (IC or chip). And just as it was in the example above, also the communication on a chip becomes harder with increasing distance between the communicating parts. If two parts are close to each other, communication can take place over a short interconnect between them with voltage signals that are defined with respect to a common ground potential. However, the signals are attenuated and distorted by the interconnect and the longer the interconnect, the more attenuation and distortion is introduced. Therefore, if the two parts are placed further away from each other, they will need stronger drivers to overcome the attenuation and distortion. These stronger drivers can be compared with the shouting in the example above. For even longer interconnects, the

attenuation and distortion can be that large, that stronger drivers are not sufficient anymore for reliable communication and more elaborate techniques are necessary.

However, the attenuation and distortion do not only depend on the length of the interconnect, but also on the communication speed: the smaller the period of the transmitted symbol, the higher the attenuation and distortion will be. Thus, if an interconnect is called long, it is long relatively to the required speed.

The required speed depends of course on the application. But in general, the circuits in newer IC technologies become faster. Therefore, also the interconnects should become faster. However, we will see that there are certain interconnects that do not become faster for newer technologies. Although the actual length of these interconnects remains equal, compared to the increasing speed of the circuits their relative length becomes too large. These are the interconnects we will look at in this thesis.

1.2 Interconnects in CMOS technologies

There are many different technologies in which chips can be made. In this thesis, we restrict ourselves to CMOS (complementary metal oxide semiconductor) technologies. CMOS technologies are the most important technologies for very large scale integrated (VLSI) applications such as computers, digital signal processing, telecommunication, medical image processing, cryptography and digital control systems.

In a CMOS technology, transistors are fabricated in a doped silicon substrate, usually with a gate of polysilicon on top of a thin layer of oxide. In order to interconnect transistors, a stack of metal layers is available to the designer. An example is shown in figure 1.1.
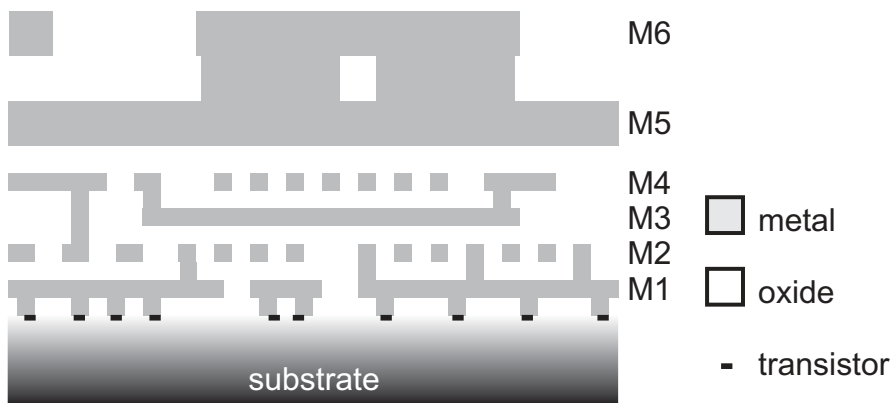
Figure 1.1: Example of the stack of metal layers in a CMOS chip

The figure shows part of the substrate with a number of metal layers on top of it, in this case six. Transistors are located at the top of the substrate. Through vias, it is possible to make connections with the metal layers. The space in between metal is filled with silicon

oxide or some other dielectric material. The permittivity of this material can differ for different metal layers. A passivation layer is usually on top of the metal stack.

In general, the highest metal layers are thicker than the lower metal layers. Often, the thick top metal layers are reserved for the power grid and clock routing. Interconnects between transistors, gates or logic blocks can then be made with the help of the other metal layers.

1.3 Interconnects and technology scaling

For over 30 years, the feature size of CMOS technology has shrunk to dimensions into the nanometer region nowadays. As a result of this continuous scaling, higher circuit speeds, lower power and larger packing densities of transistors are achieved.

The scaling of technology also affects interconnects. Both the thickness of the metal layers itself and the thickness of the oxide between the metal layers decrease with scaling. Also, the minimum width of an interconnect and the minimum spacing between two interconnects decrease.

When looking at the scaling of on-chip interconnects, it is important to distinguish between local and global interconnects [1]. The concept is shown in figure 1.2.
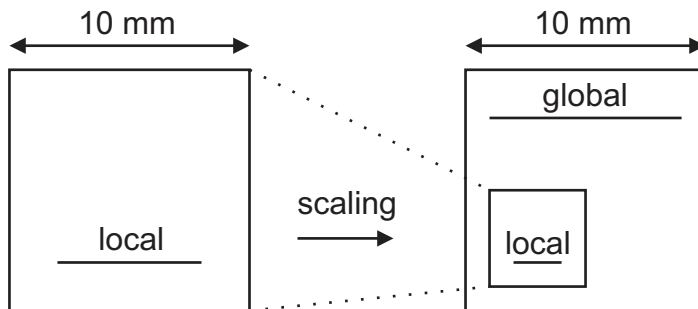
Figure 1.2: Interconnects and scaling: local and global interconnects [1].

Under scaling, a circuit with the same functionality will be smaller in a new technology. The interconnects of this circuit will also become shorter and these scaled interconnects are called local interconnects. However, as more functionality is packed on a chip, the total size of the chip remains roughly the same under scaling. Next to local interconnects, there will also be interconnects that span the entire chip. These interconnects do not scale in length and are called global interconnects.

In this thesis, we will mainly look at these global interconnects. In the next section, we will compare the speed of these global interconnects with the speed of circuits as a function of technology scaling.

1.4 Delay of global interconnects

In [1], the speed of circuits is measured as the delay through an inverter driving four identical copies of itself. This delay is called a "fanout-of-four inverter delay" or simply an FO4. The FO4 scales with technology, as shown in figure 1.3. The predicted FO4 and clock frequency ($1/(16 \cdot FO4)$ in [1]) are plotted against the drawn gate length (L_{drawn}), showing the increased speed of circuits for CMOS processes with decreasing gate lengths.
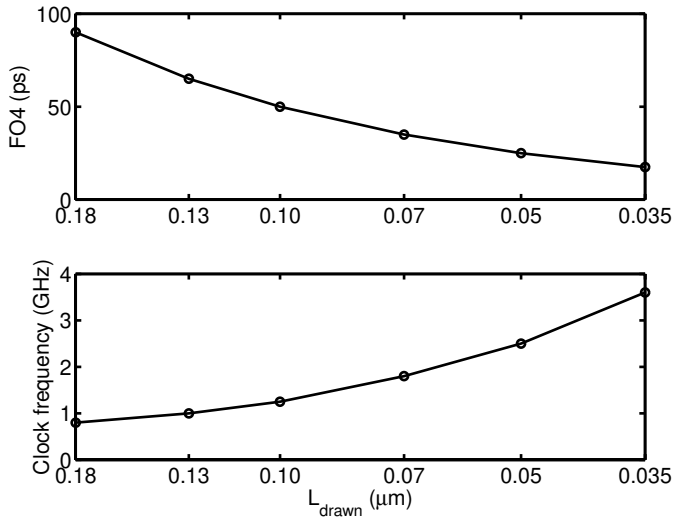
Figure 1.3: FO4 delay and clock frequency as a function of technology (drawn gate length) [1].

In order to also find the speed of the interconnects for these technology nodes, we model the interconnect with an infinite number of R'L'C'-sections:
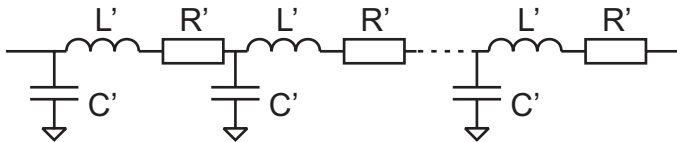
Figure 1.4: Distributed RLC model of an on-chip interconnect.

R' is the distributed resistance per unit length (Ω/m), L' the distributed inductance per unit length (H/m) and C' the distributed capacitance (F/m) per unit length. A more detailed description on interconnect modeling is given in chapter 2.

Table 1 gives some typical values for a practical interconnect and compares it with other types of wires [2].

	wire cross-section (m^2)	R' (Ω/m)	L' (nH/m)	C' (pF/m)
on-chip	0.5μ·0.5 μ	150k	600	200
PC board	150μ·50μ	5	300	100
twisted pair	~500μ diameter	0.08	400	40

Table 1.1: Typical values of R', L' and C' comparing a typical on-chip interconnect with board and twisted pair wires.

The table shows that although the distributed inductance and capacitance are of the same order of magnitude for all types of wires, the distributed resistance for on-chip interconnects is much larger. Because of this very large resistance, the distributed inductance can often be neglected for global on-chip interconnects. A metric for the delay of an interconnect then is [1]

$$t_D = FO4 + \frac{1}{2} \cdot R' \cdot C' \cdot l^2 \tag{1.1}$$

l is the length of the interconnect. The equation shows that the delay of an interconnect is proportional to the length squared. The delay is further determined by the distributed resistance and capacitance of the interconnect. (The FO4 comes from an inverter that drives the interconnect).

Both the distributed resistance and capacitance are plotted in figure 1.5 for a conservative and an aggressive prediction, depending on whether or not technological limitations are considered [1]. The values are for global interconnects with a pitch of $8 \cdot L_{drawn}$ [1]. The figure shows that as technology scales, the resistance increases dramatically, while the capacitance only decreases very slowly. Based on (1.1), the delay of interconnects increases tremendously. Note that technology solutions, such as low-k dielectrics and copper interconnects are already incorporated in the predictions.

Figure 1.6 shows the delay of a 10 mm long global interconnect relative to an FO4 delay as a function of technology. The large increase over six technology nodes shows that global interconnects cannot keep up with the ever increasing speed of digital gates. Local interconnects can keep up, because their length is scaled and interconnect delay is proportional to the length squared (see (1.1)). For comparison, figure 1.6 also shows the delay of an interconnect with a length that is scaled with Ldrawn.

1.5 Global interconnect problems

1.5.1 Introduction

The previous section showed that the distributed resistance of global interconnects increases dramatically for newer CMOS processes. As a consequence, the speed of global interconnects (measured as delay) decreases with respect to the faster digital gates. In this
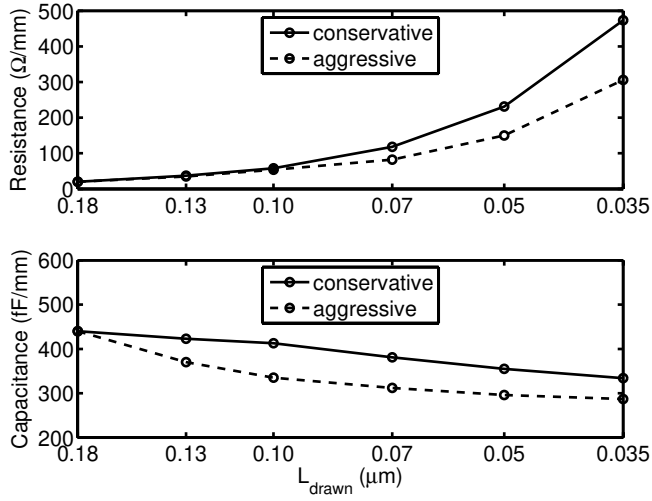
Figure 1.5: Resistance and capacitance of global interconnects as a function of technology [1].
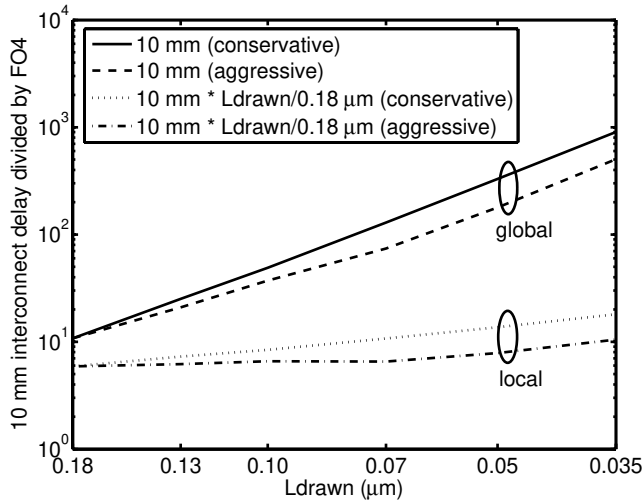
Figure 1.6: Global and local interconnect delay relative to an FO4 delay as a function of technology [1].

section, we will introduce global interconnects with respect to the achievable data rate and also to data integrity issues, area consumption and power consumption. But first, we will give a model for the transmitter and the receiver.

1.5.2 Transmitter and receiver model

In figure 1.7, a model is given for the transmitter and receiver of an interconnect.
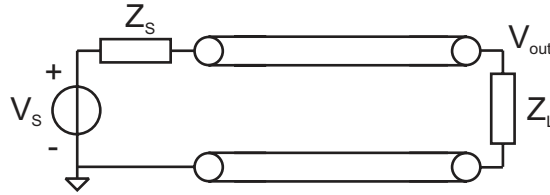
Figure 1.7: Model for global interconnects.

The voltage source, with voltage V_S and source impedance Z_S, models the transmitter. The load impedance Z_L models the receiver and V_{out} is the voltage at the receiver end of the interconnect. The symbol between Z_S and Z_L is the symbol that is used for an interconnect in this thesis.

In this thesis, we will make use of an example interconnect to illustrate the developed concepts. We will use the parameters of an 10 mm long interconnect in a 0.13 μm CMOS process with a width and spacing of both 0.4 μm. For this interconnect R' = 150 kΩ/m, L' = 400 nH/m and C' = 220 pF/m. If not stated otherwise, we will assume Z_S to be a resistance of 100 Ω and Z_L a capacitance of 100 fF. These impedances model inverters, which are conventionally used as both the transmitter and receiver

1.5.3 Achievable data rate

The achievable data rate is defined as the maximum number of bits that can be transmitted over the interconnect in a certain time period. In this section we will see how the large resistance and capacitance of interconnects in newer CMOS processes limit the achievable data rate.

The voltage source of figure 1.7 transmits symbols on the interconnect. In figure 1.8, this is shown for plain binary signaling. The duration of one symbol is T_S seconds. The voltage at the receiver (V_{out}) is also plotted in figure 1.8. The limited bandwidth of the interconnect is clearly seen in the slow rise and fall edges of this signal.

The performance of the interconnect is analyzed with the help of an eye-diagram. For this, the output signal is cut into pieces with a width of T_S and these pieces are all plotted on top of each other. Figure 1.9 shows the eye-diagram of V_{out}, using random data. The eye should be sufficiently open for reliable communication. The eye-opening is determined by an eye-height and an eye-width. Both eye-height and eye-width depend on the data rate. For determining the achievable data rate, we will calculate these eye-diagram properties as a function of data rate and see at which data rate the eye is closed. In order to do that, we will first calculate the transfer function from V_S to V_{out}. From this transfer function the impulse response is calculated. The convolution of this impulse response with the transmitted symbol shape gives the symbol response, from which it is possible to calculate the eye-diagram properties.
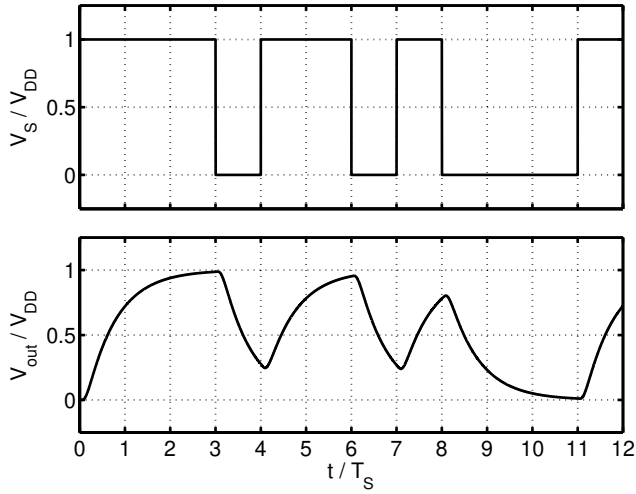
Figure 1.8: Input and output signals of 10 mm global interconnect.
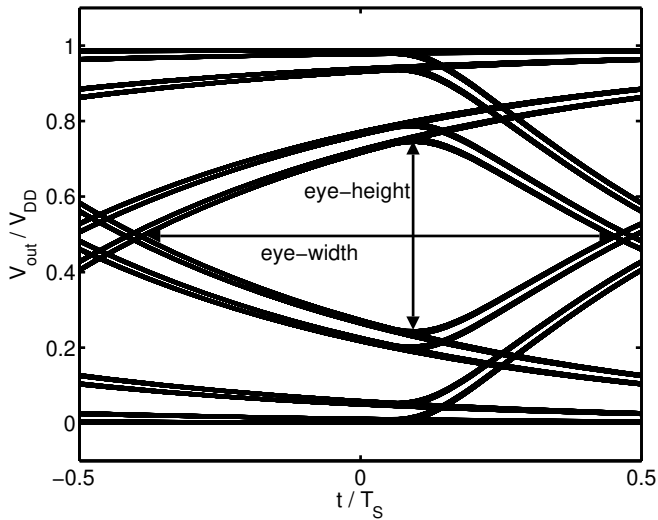
Figure 1.9: Eye-diagram of V_{out}.

The transfer function $H = V_{out}/V_S$ depends on both the interconnect parameters R', L' and C' (see figure 1.4) and on the termination impedances Z_S and Z_L (see figure 1.7). For the example interconnect from section 1.5.2 the transfer function is plotted in figure 1.10. Details on how to calculate this transfer function can be found in the next chapter.

0
−10
−20
−30
−40
−50
−60
−70
−80
−90
−100

$|H|$ (dB)

−3dB @ 95 MHz

10^6 10^7 10^8 10^9 10^{10}
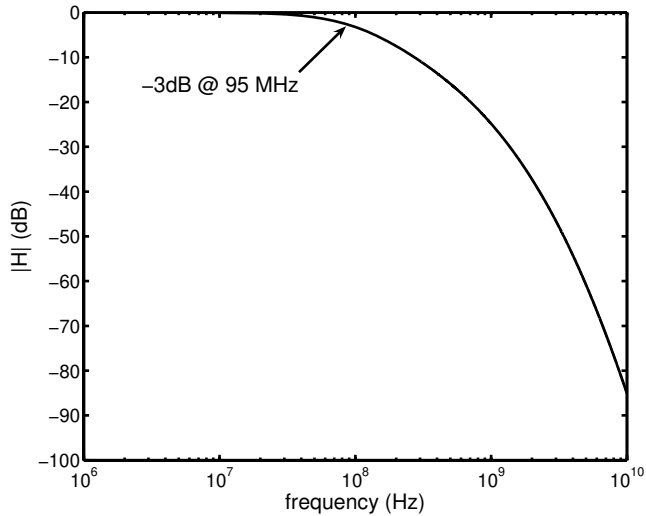
frequency (Hz)

Figure 1.10: Transfer function of a 10 mm global interconnect (see section 1.5.2 for parameters).

The transfer function shows that higher frequencies are more heavily attenuated. The high distributed resistance and capacitance give a -3dB bandwidth of only 95 MHz in this example. The limited bandwidth shows up as a long tail in the time domain. Figure 1.11 depicts the impulse response corresponding to the transfer function of figure 1.10.

The symbol response can be found by convolving the impulse response with the transmitted symbol shape. For plain binary signaling, the symbol response is shown in figure 1.12 (T_S = 2.5 ns). The long tail of the impulse response is also present in the symbol response. This long tail causes inter-symbol interference (ISI) between consecutive symbols, reducing the eye-opening.

Detection of a symbol takes place at a certain detection instant. A symbol time (T_S) later, the next symbol is detected. However, at this moment there is still signal left from the previous symbol, called inter-symbol interference. The maximum ISI is the summation of the absolute value of all ISI points. By subtracting this maximum ISI from the signal value at the detection instant, the worst-case eye-height at that detection instant is found. By calculating this eye-height for different detection instants, the optimum detection instant can be found where the eye-height is maximum. Also, the eye-width can be calculated as the time between the two detection instants where the eye-height is zero. The latency is equal to the optimum detection instant, as the symbol is transmitted at time is zero.

Figure 1.13 depicts the eye-diagram properties for different data rates. The data rate on the x-axis is normalized to the bandwidth (95 MHz for the example interconnect from section 1.5.2).
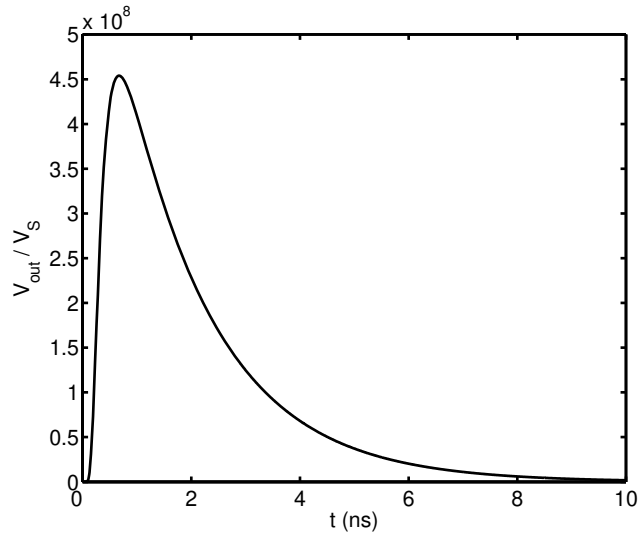
Figure 1.11: Impulse response of a 10 mm global interconnect (see section 1.5.2 for parameters).
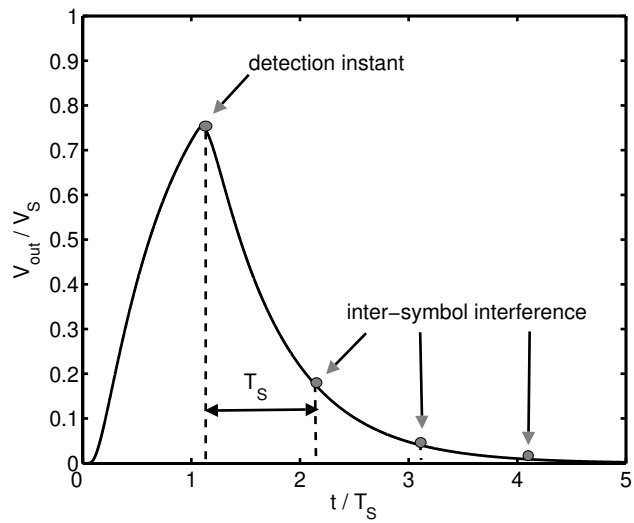
Figure 1.12: Symbol response with inter-symbol interference points.
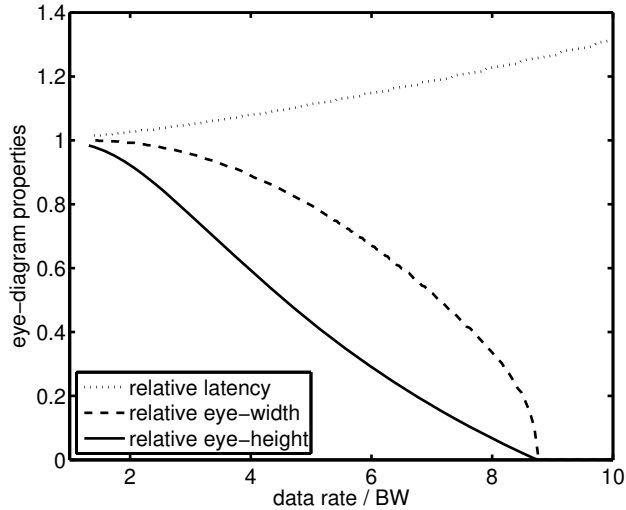
Figure 1.13: Eye-diagram properties as a function of data rate (b/s) divided by the -3 dB bandwidth (Hz) of the interconnect.

The figure shows that the eye-height, relative to V_S, decreases for higher data rates. Also the eye-width, relative to the symbol time, decreases for higher date rates. The latency, relative to the symbol time, increases. The eye-opening that is needed depends on the input swing that is required by the receiver and on error sources like offset and crosstalk. These noise sources will be treated in the next section.

In summary, the small bandwidth of global on-chip interconnects, caused by a high distributed resistance and distributed capacitance, limits the achievable data rate.

1.5.4 Data integrity

In the previous section, we have introduced the achievable data rate and have seen that it is limited by inter-symbol interference. In this section, we will look briefly at offset, crosstalk and noise sources like supply and substrate noise, which all degrade data integrity and further reduce the achievable data rate.

Offset results from mismatches in the circuits. We will represent all offsets by a single voltage source, as shown in figure 1.14, assuming that a circuit connected to point A has a high-ohmic input impedance. The eye-opening at the output of the interconnect should be large enough to tolerate this offset.

Another phenomenon that can degrade data integrity is crosstalk. A signal that is transmitted on an interconnect is not only seen at its own receiver, but also at the receiver of neighboring interconnects. The most important crosstalk mechanism is via the capacitance between two interconnects [3]. Figure 1.15 shows the capacitances of an interconnect.
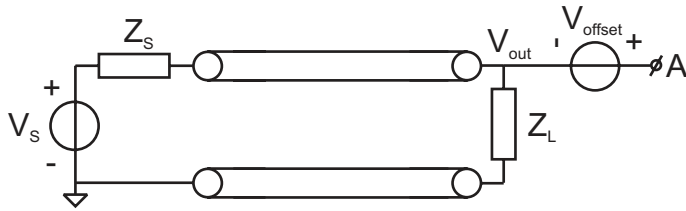
Figure 1.14: Offset is represented in a single voltage source.
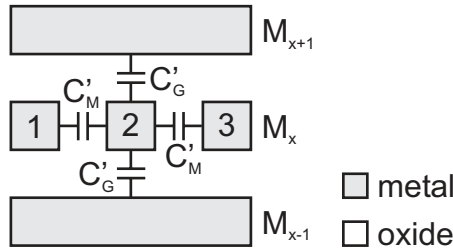
Figure 1.15: Capacitances to ground (C_G') and neighboring interconnects (C_M').

Assuming all metal except the middle interconnect (2) to be grounded, the earlier defined capacitance C' in figure 1.4 is equal to $2C_G'+2C_M'$.

C_M' is the capacitance between two interconnects and is mainly responsible for the crosstalk between them. Assuming now that the left interconnect (1) is also terminated with Z_S and Z_L, we can define a crosstalk voltage V_X at the end of this interconnect, as shown in figure 1.16.
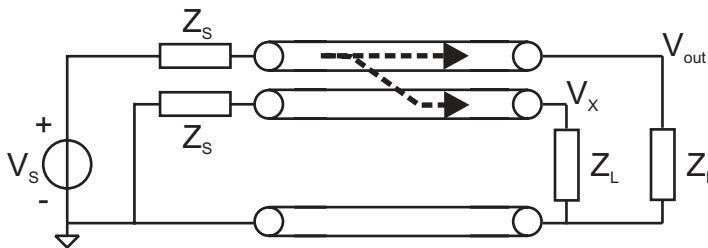
Figure 1.16: Definition of crosstalk voltage V_X.

The transfer functions $H = V_{out}/V_S$ and $H_X = V_X/V_S$ are plotted in figure 1.17. The parameters of the example interconnect from section 1.5.2 are used again with C_M' = 50 fF/mm.

Not only the limited bandwidth of H, but also the crosstalk will limit the achievable data rate. As was done in the previous section, we will look at the impulse and symbol responses. The impulse responses are in figure 1.18. Both V_{out} and V_X have a long tail. These long tails are also present in the symbol responses, which are shown in figure 1.19 for T_S = 2.5 ns.
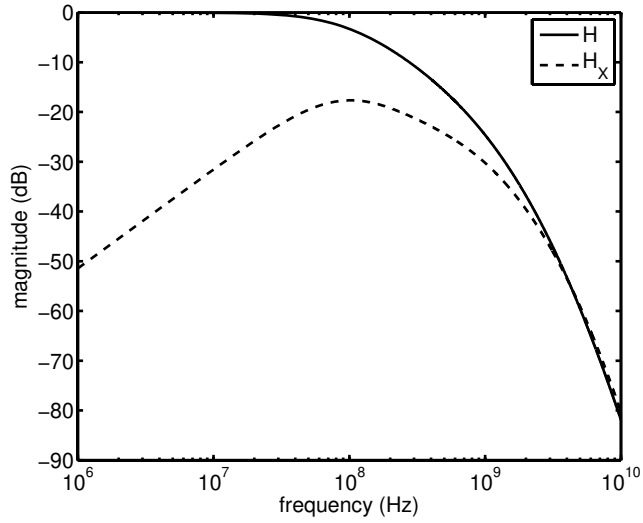
Figure 1.17: Transfer function and crosstalk transfer function (see section 1.5.2 for parameters).
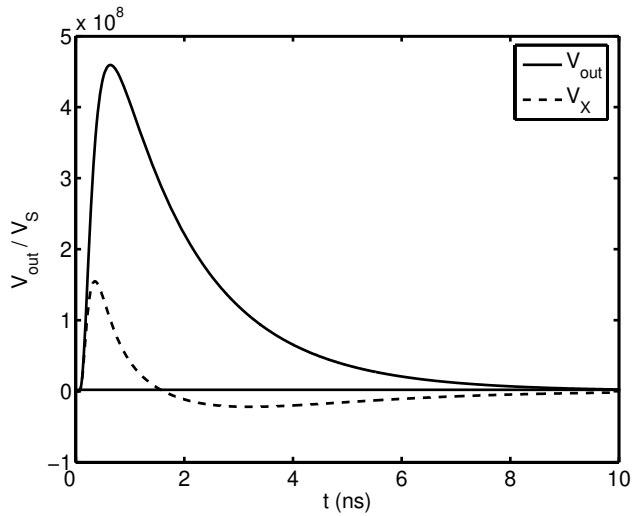
Figure 1.18: Impulse responses (see section 1.5.2 for parameters).
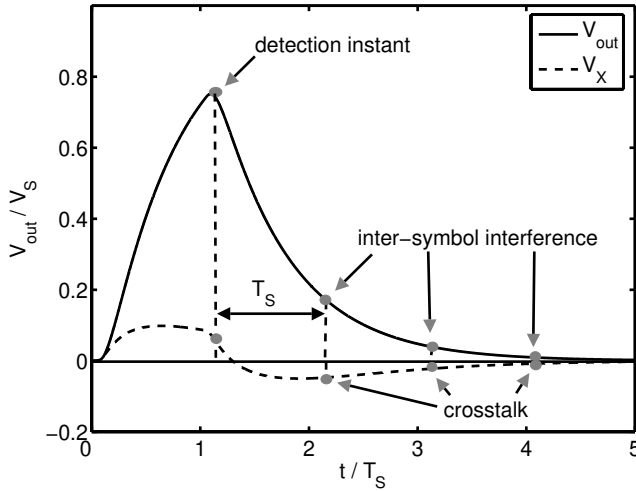
Figure 1.19: Symbol responses.

As explained in the previous section, the eye-height is calculated as the signal value at the detection instant minus all ISI points. However, the eye-height is even further reduced by the crosstalk. For calculating the worst-case eye-height, now both the absolute value of the ISI points and the absolute value of all crosstalk points (including the one at the detection instant) should be subtracted from the value at the detection instant. Therefore, the eye-diagram properties change with respect to the case when there is no crosstalk. Figure 1.20 shows the relative eye-height as a function of data rate with no crosstalk and with worst-case crosstalk on two neighboring interconnects. It is clear from the figure that the eye-height has decreased because of the crosstalk. Therefore, with crosstalk lower data rates are achievable.

There are still other noise sources besides neighbor-to-neighbor crosstalk like crosstalk from other metal layers, supply and substrate noise. Just as the eye-opening should be large enough to allow for offset, a noise margin is needed for these noise sources. However, by using differential interconnects (see chapter 4), most of these noise sources become common-mode and can be rejected by a well-designed receiver.

1.5.5 Chip area

We do not only want to have a high achievable data rate for our interconnects, but we also want to consume as less chip area as possible. In general, the area occupied by a global interconnect is much larger than the area occupied by its transmitter en receiver. For instance, the global interconnects that are discussed in this thesis (in 0.13 μm CMOS and 90 nm CMOS) have a total area of about 10^4 μm^2, while the transceivers only occupy an area in the order of 10^2 μm^2. Therefore, for minimal area, the dimensions of the interconnect must be optimized. Small interconnects might seem optimal, but they have a limited bandwidth. Very wide interconnects offer a high bandwidth, but also occupy large area.
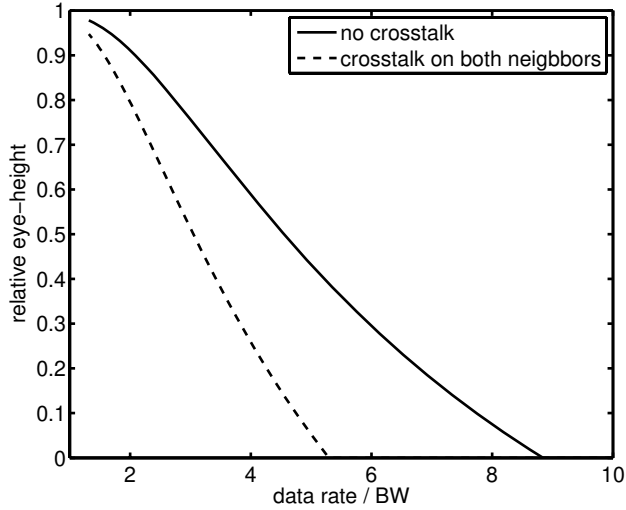
Figure 1.20: Eye-height with and without crosstalk.

In figure 1.21, three metal layers are drawn: M_{x-1}, M_x and M_{x+1}. The top and bottom metal layer are filled with metal and the global interconnects are placed in the middle metal layer. The height h and oxide thicknesses d_T and d_B are in general fixed by the process and often
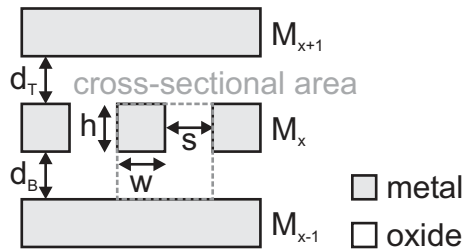
Figure 1.21: Interconnect dimensions.

$d_T \approx d_B \approx d$. A designer can only choose the width w of the interconnect and the spacing s between interconnects. We define the cross-sectional area as

$$area = (w+s) \cdot (h+d) \tag{1.2}$$

In this thesis, we will use optimal dimensions of the interconnect for which the bandwidth per cross-sectional area is largest. A bus with these optimized interconnects will have the highest aggregate data rate for a certain bus area.

1.5.6 Power consumption

Next to the desire to minimize chip area, also power consumption needs to be minimized. A simple model of power consumption is given by the following equation [4].

$$P_{dyn} = \frac{1}{2} \cdot p_{trans} \cdot C' \cdot l \cdot V_{swing} \cdot V_{DD} \cdot f_{clock} \tag{1.3}$$

With P_{dyn} the dynamic power consumption, p_{trans} the transition probability, C' the distributed capacitance, l the length of the interconnect, V_{swing} the voltage swing on the interconnect, V_{DD} the supply voltage and f_{clock} the clock frequency. The transition probability is the probability that a transition from high to low or from low to high is made. For random data, this number is 0.5. In this thesis, we will also use the term data activity, which is equal to the transition probability.

Next to this dynamic power consumption, there will in many cases also be static power consumption (P_{static}). For low data activities, there is not much dynamic power. In these cases, the total power is dominated by the static power. Therefore, both dynamic and static power consumption should be low.

Instead of looking just at the power consumption, we will look at the power consumption divided by the data rate, as we want to minimize power consumption at maximum data rate. This energy per bit (E/bit) can be plotted against p_{trans}, as shown in figure 1.22 for the example interconnect from section 1.5.2.

The more transitions, the more E/bit is dissipated. The overall E/bit should be as low as possible. Furthermore, it is desirable to have low 'static' energy consumption at zero data activity.
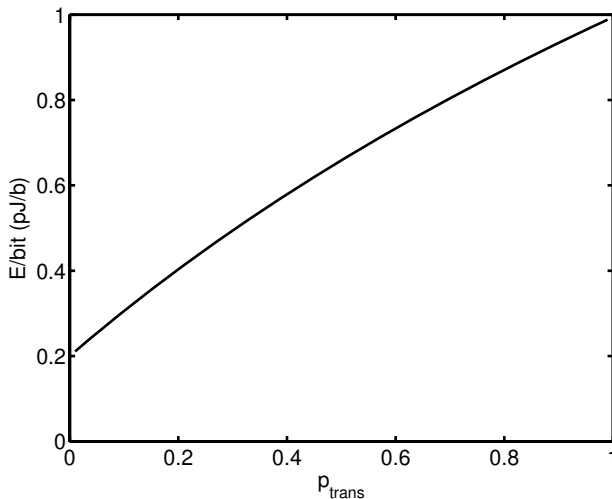
Figure 1.22: Example of an energy per bit curve as a function of transition probability.

1.6 Solutions in literature

1.6.1 Introduction

In the previous sections, we saw that global interconnects have a number of problems. First we saw that for newer technologies the delay of interconnects increases tremendously with respect to the delay of a circuit, due to an increasing distributed resistance. The high resistance and capacitance of interconnects in newer CMOS processes also limit the achievable data rate. We mentioned that offsets, supply noise and crosstalk can further limit the achievable data rate. Finally, we discussed area and power consumption, which should both be low.

In literature, a number of solutions are given to decrease the delay or increase the achievable data rate. Also solutions to lower the power consumption and decrease the influence of offset, supply noise and crosstalk can be found in literature, as summarized below.

1.6.2 Delay

The delay of global interconnects can be reduced by making the delay linear, instead of quadratic (see (1.1)), with length. This can be done by breaking up the interconnect into smaller segments, with a repeater, often implemented as an inverter, driving each segment. All segments have the same delay, as they are of equal length. Thus, a two times longer interconnect, having twice as much segments, will also have a two times larger delay.

By breaking up the interconnect into N segments, the delay would in principle be N times smaller. However, the repeaters add additional delay and a delay-optimal solution exists where both the segment length and driver strength can be calculated [5]. Unfortunately, the use of repeaters has some disadvantages. With plain non-clocked buffers as repeaters, delay variations due to crosstalk and due to process variations accumulate and limit the achievable data rate [6]. Furthermore, the number of repeaters needed in microprocessors in future technologies will increase tremendously [7]. Next to the (large) occupied area and placing problems, the power that is dissipated in these repeaters can become a large portion of the total power dissipation of the chip [8].

Another approach is to use 'smart' repeaters, where fewer repeaters are needed for the same total delay. Transition aware circuits, for example, detect a transition in an early phase. This can then be used to generate an output signal earlier [9] or accelerate or boost the signals on the interconnect [10, 11]. However, this will make the transceiver more susceptible to noise. Other 'smart' repeaters use current-mode sensing [6, 12-15] to decrease the delay of the interconnect up to a factor three compared to the conventional voltage mode sensing. However, this often comes at the cost of high static power consumption.

Also, the interconnect can be designed in such a way that is has transmission line behavior [16-24]. Instead of using interconnects at minimum pitch, extreme wide interconnects are used. In this way, the resistance becomes small and the interconnect is dominated by its inductance and capacitance. Near speed of light delays are possible, but at the cost of large interconnect area.

In summary, techniques to reduce delay exist, but come at the cost of data integrity, power or area.

1.6.3 Achievable data rate

The achievable data rate is limited because of the limited bandwidth of the interconnect. The bandwidth of the interconnect can be improved with an equalizer. An ideal equalizer has a transfer function such that $H_{interconnect} \cdot H_{equalizer} = 1$. In practice, this can be achieved only for a limited frequency band, but still, the bandwidth can be increased a lot.

In [25] switches are placed along the interconnect, that discharge the interconnect every clock period. However, this increases the already troublesome clock load and can cost considerable power consumption.

Equalization is also implemented in [13, 14, 26], referred to as dynamic overdriving or pre-emphasis equalization. The voltage swing on the interconnect is reduced, but at data transitions a temporarily higher voltage swing is transmitted. In this way, the data transitions are emphasized. However, a transmitter that is able to transmit both a high voltage and a low voltage in an efficient way is difficult to design. Furthermore, the low swing will make the transceiver more sensitive to noise sources.

Distributed loss compensation [27] uses negative capacitances that are placed along the interconnect, but has high static power consumption due to the bias current of the negative capacitances.

As was the case for delay, the achievable data rate can be improved, but again at the cost of data integrity, power or area.

1.6.4 Power consumption

For the standard repeater solution, a power reduction can be achieved by not optimizing for delay but for power. For a given delay penalty of 5%, a power reduction of 27% (180 nm technology node) up to 45% (50 nm technology node) can be achieved compared to the delay optimal solution [28].

Another way to decrease power consumption is the use of low-swing signaling [20, 25, 26, 29, 30]. Because of the low swing, the capacitance of the interconnect is not fully charged anymore. However, the lower swing makes the transceiver more susceptible to noise sources. Furthermore, methods to create low-swing signaling often use threshold voltage drops, which decrease the bandwidth and thus, the achievable data rate. Otherwise, separate low-voltage supplies are used, which is also not desirable.

Power can also be reduced by coding [31]. The coding targets to have as few data transitions as possible. However, coding schemes have often only limited power reduction for practical data activities.

1.6.5 Data integrity

Techniques like shielding can provide a solution for crosstalk problems. However, if low-swing signaling is used to reduce power consumption, noise margins become smaller [29].

Then, differential signaling is preferred. In differential systems, most noise sources appear as common-mode and can be rejected by the receiver. In this way, differential interconnects are for instance not sensitive to crosstalk from orthogonal crossing metal layers. However, differential interconnects do not solve the problem of neighbor-to-neighbor crosstalk.

In CMOS memory cells, twists in differential interconnect-pairs are already widely used to cancel crosstalk between bitlines [32-35]. Twists are also used on printed circuit boards [36, 37]. In general, many twists are placed at evenly-spaced intervals along the interconnects. To cancel neighbor-to-neighbor crosstalk, twists are also proposed for on-chip global interconnects [25, 38-41]. Again, many twists are placed at evenly-spaced intervals along the interconnects. However, the many vias needed to make the twists add to the already troublesome interconnect resistance. Moreover, each twist requires use of a second metal layer, which complicates global routing algorithms.

1.7 Challenges

From the solutions as found in literature, we can draw the conclusion that it is difficult to have a solution that does not only have a low delay or a high achievable data rate, but also low power and area consumption. Furthermore, low-swing techniques for low power consumption make data integrity issues more difficult.

In this thesis, the focus will be more on achievable data rate than on delay, because designers can get around the large delay of global interconnects by using pipelining. Moreover, solutions that increase the achievable data rate, often also decrease the delay.

While maximizing the achievable data rate, it is desirable to minimize chip area and power consumption. Also data integrity should be maintained. Crosstalk between interconnects, supply noise and substrate noise can degrade this data integrity and the interconnect should be robust against these noise sources. These issues are summarized in the central question of this thesis:

How can we design global interconnects and transmitter and receiver electronics in future IC technologies to
- *maximize data capacity,*
- *minimize chip area consumption,*
- *minimize power consumption while*
- *maintaining data integrity?*

The research that led to this thesis has been done in cooperation with Daniël Schinkel. His research includes topics like communication and equalization techniques both at the transmitter and the receiver and clocked sense amplifiers that are used to make a decision at the receiver [42]. The focus of this thesis will be on the modeling and design of the interconnect itself and the interactions of the interconnect with the transmitter and the receiver.

1.8 Scope of this thesis

The remainder of this thesis is organized as follows.

Chapter 2 describes the interconnect models that are used. From Maxwell's equations we find the response of interconnects to electromagnetic waves. This response results in a distributed model, described by s-parameters. With this distributed model transfer functions and power consumption can be calculated.

The design of the interconnects is discussed in chapter 3. First, the optimal dimensions of the interconnects are determined. Then, termination concepts are developed, both at the transmitter and the receiver side. These termination concepts are developed to increase the achievable data rate, while minimizing area and power. Other techniques for increasing the achievable data rate, resulting from the work of Daniël Schinkel [42], are briefly explained at the end of the chapter.

In chapter 4 the data integrity of on-chip transceivers is examined. First, offset mechanisms in the receiver are studied. After that, the interconnects are designed for minimal crosstalk. For neighbor-to-neighbor crosstalk, twists are used. The chapter describes how to calculate the optimal position of these twists, depending on the termination impedances that are used. The last section calculates the amount of crosstalk from aggressor interconnects that are not in the same, but in another metal layer than the victim interconnect.

Chapter 5 describes the implementation of the concepts, as developed in chapters 3 and 4. First, optimal interconnect dimensions are calculated for two different CMOS technologies. Then, two transceiver concepts are discussed that are designed in these technologies. The first transceiver achieves a high data rate and the second transceiver achieves both a data rate and low power consumption.

The conclusions of this thesis are found in chapter 6. Key answers to the central question of this thesis are given and the proposed transceivers of chapter 5 are evaluated.

Chapter 2

Interconnect models

2.1 Introduction

This chapter deals with the modeling of on-chip interconnects. The models that are treated in this chapter are used in the subsequent chapters for designing global interconnects and developing termination concepts. This chapter starts with a review of the location of interconnects in an integrated circuit in section 2.2. Section 2.3 discusses the response of interconnects to electromagnetic waves in terms of voltages and currents. This response results in a distributed model that is presented in section 2.4. In this section, it is shown how to calculate transfer functions and power consumption with the help of this distributed model. Section 2.5 expands the distributed model by introducing neighboring interconnects. In that section, the crosstalk transfer function is calculated. The relation between the distributed model and a lumped model is treated in section 2.6. Finally, in section 2.7 it is explained how we can find the parameters of the used models with the help of a 3D EM-field simulator [43].

2.2 Location of interconnects in an IC

In CMOS technology, transistors are fabricated in a doped silicon substrate usually with a gate of polysilicon on top of a thin layer of oxide. In order to interconnect transistors, a stack of metal layers is available to the designer. An example is shown in figure 2.1. The figure shows the cross-section of a chip with part of the substrate and six metal layers on top of it. Transistors are located at the top of the substrate and through vias, it is possible
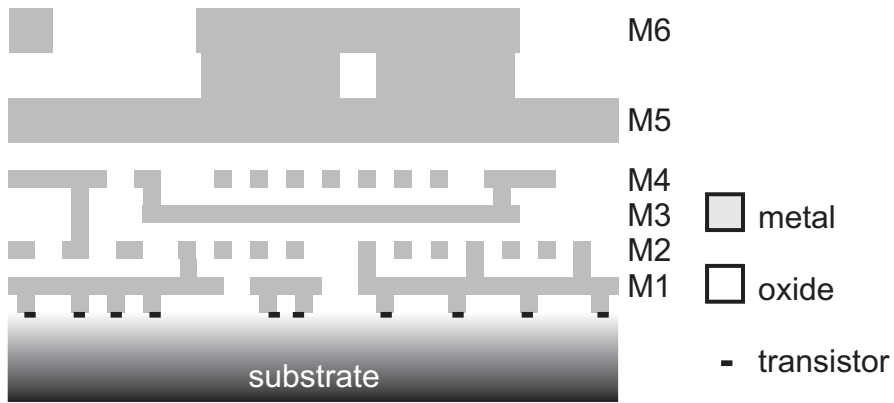
Figure 2.1: Example of the stack of metal layers in a CMOS chip.

to make connections with the metal layers. The space that is not filled with metal is filled with silicon oxide or another dielectric (preferably with a low dielectric constant). In general, the highest metal layers are thicker than the lower metal layers.

The best place for global interconnects would be in one of these thick top metal layers, as the distributed resistance of the interconnect is inversely proportional to its cross-section (as will be shown in (2.37)). In this thesis, however, we will locate them in one of the thinner metal layers. There are two reasons for this. The first is that the top metal layers are often reserved for the power and clock grid. The second reason has to do with the goal of our research. In future CMOS technologies, the high resistance and capacitance will provide a severely limited bandwidth (see chapter 1). Therefore, we would like to look at interconnects that have a limited bandwidth and it makes therefore sense to use the thinner metal layers.

Figure 2.2 shows a schematic view and defines the dimensions of an interconnect.
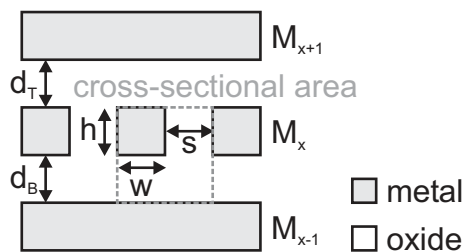
Figure 2.2: Interconnect dimensions.

We assume that a bus with global interconnects is placed in metal layer x, while metal layer x-1 and x+1 are filled with metal, to emulate high-density metal use.

22

2.3 Response to electromagnetic waves

2.3.1 Introduction

Suppose that one of the interconnects in metal layer x has a potential V(z) at a certain position z along the interconnect. If all other metal is connected to ground, the voltage difference between the interconnect and the other metal parts gives rise to an electric field density E. Suppose also a current I(z) that flows through the interconnect. This current then induces a magnetic flux density B. As both the E-field and the B-field have a direction, we will write them in bold case to denote that they are vectors. Figure 2.3 shows the direction of **E** and **B**.
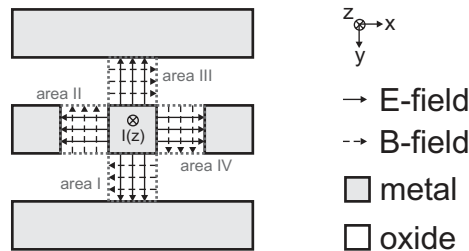
Figure 2.3: Direction of **E** and **B**.

For simplicity, we neglect fringe fields and suppose all fields to be concentrated in the areas I to IV. Due to the limited conductivity of the interconnect, there will also be an E-field inside the interconnect, which has the same direction as the current. For our analysis, we assume the E-field to be much smaller in the z-direction than in the x- and y-direction. This assumption can be made, since for the interconnects we will look at, the length (z-direction) is much larger than the spacing to other interconnects (x-direction) and than the oxide thickness (y-direction).

In this section, we calculate the voltage change in the z-direction (dV(z)/dz) as a function of I(z). We also calculate the current change in the z-direction (dI(z)/dz) as a function of V(z). With the help of these relations, we are able to construct a distributed model in the next section.

2.3.2 Definitions

First, some quantities are defined.

V: electric potential

The gradient of scalar V is defined as

$$\nabla V = \mathbf{a}_x \frac{\partial V}{\partial x} + \mathbf{a}_y \frac{\partial V}{\partial y} + \mathbf{a}_z \frac{\partial V}{\partial z} \tag{2.1}$$

with $\mathbf{a}_x$, $\mathbf{a}_y$ and $\mathbf{a}_z$ unit vectors in the directions x, y and z respectively.

E: electric field density
B: magnetic flux density
J: current density
A: vector magnetic potential

These are all vectors and have a component in the x-, y- and z-direction.
The curl of a vector **A** is defined as

$$\nabla \times \mathbf{A} = \mathbf{a}_x (\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}) + \mathbf{a}_y (\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}) + \mathbf{a}_z (\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y})$$

(2.2)

Furthermore, the following symbols are used:

σ: conductivity
μ: permeability
ε: permittivity.

Finally, we use the definition of potential [44]:

$$\nabla V = -\mathbf{E} - \frac{\partial \mathbf{A}}{\partial t}$$

(2.3)

2.3.3 dV(z)/dz as a function of I(z)

First we calculate dV(z)/dz as a function of I(z). We assume that the voltage inside the interconnect is constant in both the x- and y-direction and is only changing in the z-direction.

$$\frac{\partial V(z)}{\partial z} = -E_z - \frac{\partial A_z}{\partial t}$$

(2.4)

In order to find dV(z)/dz as a function of I(z), we have to express both E_z and A_z as a function of I(z). **E** is related to the current density **J** [44]:

$$\mathbf{J} = \sigma \cdot \mathbf{E}$$

(2.5)

and the current I(z) is given by

$$I(z) = J_z \cdot w \cdot h$$

(2.6)

Thus, (2.5) can be written as

$$\frac{I(z)}{w \cdot h} = \sigma_m \cdot E_z$$

(2.7)

24

with σ_m the conductivity of the metal. We have thus found E_z as a function of I(z). Now we come to A_z. **A** is related to **B** [44]:

$$\nabla \times \mathbf{A} = \mathbf{B} \tag{2.8}$$

A only has a component in the z-direction and (2.8) turns into

$$\mathbf{a}_x \cdot \frac{\partial A_z}{\partial y} - \mathbf{a}_y \cdot \frac{\partial A_z}{\partial x} = \mathbf{B} \tag{2.9}$$

There are four areas shown in figure 2.3 and (2.9) should be satisfied for all four areas. Thus:

$$\frac{\partial A_z}{\partial y} = B_x \rightarrow \begin{cases} A_z = -d \cdot B_{x,I} \\ A_z = d \cdot B_{x,III} \end{cases} \tag{2.10}$$

in areas I and III and

$$-\frac{\partial A_z}{\partial x} = B_y \rightarrow \begin{cases} A_z = -s \cdot B_{y,II} \\ A_z = s \cdot B_{y,IV} \end{cases} \tag{2.11}$$

in areas II and IV. Thus, the B-fields in the areas I and III are equal, but with opposite sign and the same is true for the B-fields in the areas II and IV. The relation between the B-fields in the areas I and II is

$$B_{x,I} = \frac{s}{d} \cdot B_{y,II} \tag{2.12}$$

Now we need to find either B_x or B_y as a function of I(z). This can be done with the help of one of Maxwell's equations [44]:

$$\nabla \times \frac{\mathbf{B}}{\mu} = \mathbf{J} + \frac{\partial(\varepsilon \cdot \mathbf{E})}{\partial t} \tag{2.13}$$

B has no component in the z-direction and the curl of **B** becomes

$$\nabla \times \mathbf{B} = -\mathbf{a}_x \frac{\partial B_y}{\partial z} + \mathbf{a}_y \frac{\partial B_x}{\partial z} + \mathbf{a}_z \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} \right) \tag{2.14}$$

In areas I to IV, **E** has no component in the z-direction. Therefore, (2.13) can be rewritten for the z-direction as

$$\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} = \mu \cdot J_z \tag{2.15}$$

Integrating over both w and h gives

$$-w \cdot B_{x,I} - h \cdot B_{y,II} + w \cdot B_{x,III} + h \cdot B_{y,IV} = -2 \cdot w \cdot B_{x,I} - 2 \cdot h \cdot B_{y,II} = \mu \cdot I(z) \tag{2.16}$$

Combining (2.12) and (2.16) gives expressions for the B-field in areas I and II:

$$B_{x,I} = -\mu \cdot I(z) \cdot \frac{s}{2 \cdot (w \cdot s + h \cdot d)} \tag{2.17}$$

$$B_{y,II} = -\mu \cdot I(z) \cdot \frac{d}{2 \cdot (w \cdot s + h \cdot d)} \tag{2.18}$$

Now the B-field is known as a function of I(z), also A_z can be calculated as a function of I(z).

$$A_z = -d \cdot B_{x,I} = \mu \cdot I(z) \cdot \frac{s \cdot d}{2 \cdot (w \cdot s + h \cdot d)} \tag{2.19}$$

With (2.7) and (2.19) we have both E_z and A_z as a function of I(z) and can rewrite (2.4) as

$$\frac{\partial V(z)}{\partial z} = -\frac{1}{\sigma_m \cdot w \cdot h} \cdot I(z) - \mu \cdot \frac{s \cdot d}{2 \cdot (w \cdot s + h \cdot d)} \cdot \frac{\partial I(z)}{\partial t} \tag{2.20}$$

or

$$\frac{\partial V(z)}{\partial z} = -R' \cdot I(z) - L' \cdot \frac{\partial I(z)}{\partial t} \tag{2.21}$$

2.3.4 dI(z)/dz as a function of V(z)

Now, we calculate dI(z)/dz as a function of V(z). We use Maxwell's equation (2.13) again. In (2.15) we looked at the z-direction, now we will look at the x- and y-direction. With the help of (2.14) we find

$$-\mathbf{a}_x \frac{\partial B_y}{\partial z} + \mathbf{a}_y \frac{\partial B_x}{\partial z} = \mu \cdot \mathbf{a}_x \cdot (J_x + \frac{\partial(\varepsilon \cdot E_x)}{\partial t}) + \mu \cdot \mathbf{a}_y \cdot (J_y + \frac{\partial(\varepsilon \cdot E_y)}{\partial t}) \tag{2.22}$$

which results in

$$-\frac{\partial B_y}{\partial z} = \mu \cdot J_x + \varepsilon \cdot \mu \cdot \frac{\partial E_x}{\partial t} \tag{2.23}$$

and

$$\frac{\partial B_x}{\partial z} = \mu \cdot J_y + \varepsilon \cdot \mu \cdot \frac{\partial E_y}{\partial t} \tag{2.24}$$

Expressions for B_x and B_y were already found in (2.17) and (2.18), which can be used to find

$$-\frac{\partial B_{y,II}}{\partial z} = \mu \cdot \frac{d}{2 \cdot (w \cdot s + h \cdot d)} \cdot \frac{\partial I_z}{\partial z} = \mu \cdot J_x + \varepsilon \cdot \mu \cdot \frac{\partial E_x}{\partial t} \tag{2.25}$$

$$\frac{\partial B_{x,I}}{\partial z} = -\mu \cdot \frac{s}{2 \cdot (w \cdot s + h \cdot d)} \cdot \frac{\partial I_z}{\partial z} = \mu \cdot J_y + \varepsilon \cdot \mu \cdot \frac{\partial E_y}{\partial t} \tag{2.26}$$

This can be written as

$$\frac{\partial I(z)}{\partial z} = \frac{2 \cdot (w \cdot s + h \cdot d)}{s \cdot d} \cdot (J_x \cdot s) + \frac{2 \cdot \varepsilon \cdot (w \cdot s + h \cdot d)}{s \cdot d} \cdot \frac{\partial (E_x \cdot s)}{\partial t} \tag{2.27}$$

$$\frac{\partial I(z)}{\partial z} = -\frac{2 \cdot (w \cdot s + h \cdot d)}{s \cdot d} \cdot (J_y \cdot d) - \varepsilon \cdot \frac{2 \cdot (w \cdot s + h \cdot d)}{s \cdot d} \cdot \frac{\partial (E_y \cdot d)}{\partial t} \tag{2.28}$$

To relate I(z) to a potential, we use (2.3) again. This equation can be rewritten for the x- and y-directions as

$$\frac{\partial V(z)}{\partial x} = -E_x - \frac{\partial A_x}{\partial t} \tag{2.29}$$

$$\frac{\partial V(z)}{\partial y} = -E_y - \frac{\partial A_y}{\partial t} \tag{2.30}$$

A only has a component in the z-direction, thus

$$V(z) = -E_x \cdot s \tag{2.31}$$

$$V(z) = E_y \cdot d \tag{2.32}$$

From (2.5), we also know

$$E_x = \frac{J_x}{\sigma_o} \rightarrow J_x \cdot s = -\frac{V(z)}{\sigma_o} \tag{2.33}$$

$$E_y = \frac{J_y}{\sigma_o} \rightarrow J_y \cdot d = \frac{V(z)}{\sigma_o} \tag{2.34}$$

with σ_o the conductivity of the oxide. Now, both (2.27) and (2.28) can be written as

$$\frac{\partial I(z)}{\partial z} = -\frac{2 \cdot \sigma_o \cdot (w \cdot s + h \cdot d)}{s \cdot d} \cdot V(z) - \frac{2 \cdot \varepsilon \cdot (w \cdot s + h \cdot d)}{s \cdot d} \cdot \frac{\partial V(z)}{\partial t} \tag{2.35}$$

or

$$\frac{\partial I(z)}{\partial z} = -G' \cdot V(z) - C' \cdot \frac{\partial V(z)}{\partial t} \tag{2.36}$$

2.4 Distributed model

2.4.1 Introduction

In the previous section, by looking at the electric and magnetic fields of an interconnect, two relations between voltage and current were found. In this section, we show the corresponding distributed model and we show the solution of these equations. After that, we explain how to calculate transfer functions of the interconnect and how to calculate the power consumption with the help of s-parameters.

2.4.2 Telegrapher's equations

The equations (2.21) and (2.36) are known as the Telegrapher's equations [44]. R' is the distributed resistance per unit length, L' the distributed inductance per unit length, G' the distributed conductance per unit length and C' the distributed capacitance per unit length of the interconnect. The corresponding distributed model is shown in figure 2.4.
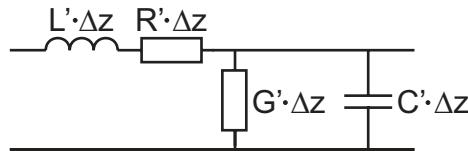
Figure 2.4: Equivalent circuit of a differential length Δz of an interconnect.

According to (2.20) and (2.35), the values of the distributed parameters for our interconnect are approximated by

$$R' = \frac{1}{w \cdot h \cdot \sigma_m} \quad (\Omega / m) \tag{2.37}$$

$$L' = \mu \cdot \left(\frac{s \cdot d}{2 \cdot (w \cdot s + h \cdot d)} \right) = \frac{\mu}{2} \cdot \left(\frac{d}{w} + \frac{s}{h} \right) \quad (H/m) \tag{2.38}$$

$$G' = \frac{2 \cdot \sigma_o \cdot (w \cdot s + h \cdot d)}{s \cdot d} = 2 \cdot \sigma_o \cdot \left(\frac{w}{d} + \frac{h}{s} \right) \quad (S/m) \tag{2.39}$$

$$C' = \frac{2 \cdot \varepsilon \cdot (w \cdot s + h \cdot d)}{s \cdot d} = 2 \cdot \varepsilon \cdot \left(\frac{w}{d} + \frac{h}{s} \right) \quad (F/m) \tag{2.40}$$

A solution for the Telegrapher's equations ((2.21) and (2.36)) is found in [44].

$$v(z,t) = \mathrm{Re}[V(z) \cdot e^{j \cdot \omega t}] \tag{2.41}$$

$$i(z,t) = \mathrm{Re}[I(z) \cdot e^{j \cdot \omega t}] \tag{2.42}$$

$$V(z) = V_0{}^p \cdot e^{-\gamma \cdot z} + V_0{}^n \cdot e^{\gamma \cdot z} \tag{2.43}$$

$$I(z) = I_0{}^p \cdot e^{-\gamma \cdot z} + I_0{}^n \cdot e^{\gamma \cdot z} \tag{2.44}$$

The p and n superscripts denote waves that travel in the positive z-direction and the negative z-direction respectively and

$$\gamma = \sqrt{(R' + j \cdot \omega \cdot L') \cdot (G' + j \cdot \omega \cdot C')} \tag{2.45}$$

The ratio of the voltage and current at any z for an infinitely long interconnect is independent of z and is called the characteristic impedance of the interconnect.

$$Z_c = \sqrt{\frac{R' + j \cdot \omega \cdot L'}{G' + j \cdot \omega \cdot C'}} \tag{2.46}$$

Of course, an interconnect is not infinitely long. If the interconnect is terminated with Z_c, then no reflections occur. Otherwise, if the interconnect is terminated with an impedance Z, reflections do occur and the ratio between reflected and incident wave is given by

$$\Gamma = \frac{Z - Z_c}{Z + Z_c} \tag{2.47}$$

2.4.3 Attenuation, reflections and distortion

An ideal interconnect would have R' = 1/G' = 0. Then, (2.45) and (2.46) can be written as

$$\gamma_{LC} = j \cdot \omega \cdot \sqrt{L' \cdot C'} \qquad (2.48)$$

$$Z_{c,LC} = \sqrt{\frac{L'}{C'}} \qquad (2.49)$$

The propagation constant has no real part. Therefore, the waves that are traveling in the interconnect are not attenuated. The characteristic impedance of the LC interconnect is real and by choosing a source and load impedance equal to this characteristic impedance, no reflections occur.

The group delay per unit length is the derivative of the imaginary part of γ with respect to ω:

$$\frac{\partial \operatorname{Im}(\gamma)}{\partial \omega} = \sqrt{L' \cdot C'} \qquad (2.50)$$

This group delay is independent of frequency, which means that every frequency component of a transmitted symbol arrives with the same delay, so the signal is not distorted.

However, global on-chip interconnects can have such a large R' that for the frequencies of interest, R' >> ωL' and not R', but L' can be ignored.

$$\gamma_{RC} = \sqrt{j \cdot \omega \cdot R' \cdot C'} = (1+j) \cdot \frac{1}{2} \cdot \sqrt{2} \cdot \sqrt{\omega \cdot R' \cdot C'} \qquad (2.51)$$

$$Z_{c,RC} = \sqrt{\frac{R'}{j \cdot \omega \cdot C'}} \qquad (2.52)$$

The propagation constant does not only have an imaginary, but also a real part. This means that the transmitted signal is attenuated along the interconnect. This attenuation is frequency dependent. The characteristic impedance also depends on frequency and it will be difficult to make a source and load impedance equal to this characteristic impedance. Therefore, reflections will occur. However, we will see later on (sections 3.3 and 3.4) that these reflections are actually beneficial.

Also, the group delay depends on frequency.

$$\frac{\partial \operatorname{Im}(\gamma)}{\partial \omega} = \frac{1}{4} \cdot \sqrt{2} \cdot \frac{R' \cdot C'}{\sqrt{\omega \cdot R' \cdot C'}} \qquad (2.53)$$

The frequency components of a transmitted symbol arrive with different delays. Therefore, the output signal is distorted.

The attenuation, reflections and distortion of RC dominated interconnects can be seen in the impulse response of the interconnect. Conventionally, an interconnect is driven by a strong driver and left open at the receiver end. For simplicity, we take $Z_S = 0$ and $Z_L = \infty$ and further use the parameters of the example interconnect from section 1.5.2. The impulse (1 V amplitude, 10 ps long) response is plotted in figure 2.5. This is done at different positions z along the interconnect, where $z = 1.00$ is at the end of the interconnect.
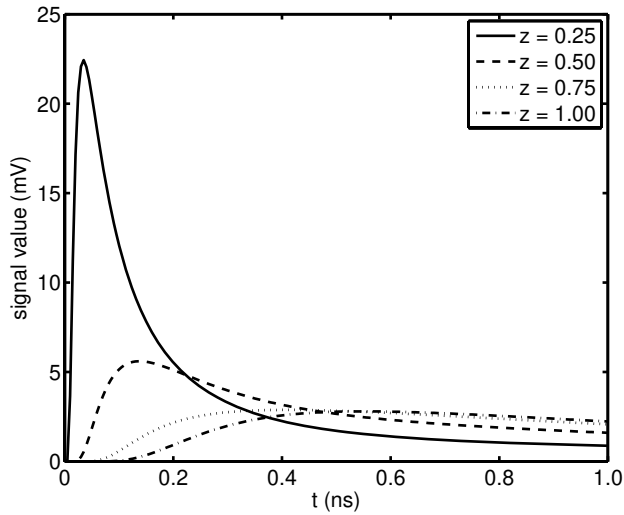
Figure 2.5: Impulse response for different positions along the interconnect with $Z_S = 0$ and $Z_L = \infty$.

The height of the response decreases for higher z (attenuation). Also, the pulse is widened (distorted). Note that the responses have a zero slope at starting time, which is due to LC behavior.

The shape of the impulse response at the different positions can be explained in two ways. The first way is to view the response as a diffusion process, where charge diffuses through the interconnect. A charge difference between two positions on the interconnect will give a current to counteract this difference. The larger the charge difference, the larger the current.

We can also explain the response with waves that are reflected many times by the termination impedances. The addition of all these reflected waves gives the shapes as seen in figure 2.5. These reflections can be made visible. In order to do that, we make use of two cases, an open and a shorted case.

Open case: $Z_S = 0$ and $Z_L = \infty$
Shorted case: $Z_S = Z_L = 0$.

From (2.47) we know that for $Z = 0$ the reflection coefficient is -1 and for $Z = \infty$ the reflection coefficient is +1. In the open case, the waves that are on the interconnect are

$$V_0{}^p + V_1{}^n - V_2{}^p - V_3{}^n + \ldots \tag{2.54}$$

V_i^p is a voltage wave traveling in the positive z-direction and V_i^n is a voltage wave traveling in the negative z-direction with i the number of reflections the wave has encountered. In the shorted case, the waves that are on the interconnect are

$$V_0{}^p - V_1{}^n + V_2{}^p - V_3{}^n + \ldots \tag{2.55}$$

By adding (2.54) and (2.55), we are able to calculate the wave that is traveling in the positive z-direction.

$$2 \cdot V_0{}^p - 2 \cdot V_3{}^n + \ldots \approx 2 \cdot V_0{}^p \tag{2.56}$$

Because of the large attenuation, V_3^n will be much smaller than V_0^p and can be neglected. In the same way, subtracting (2.55) from (2.54) gives the wave propagating in the negative z-direction.

$$2 \cdot V_1{}^n - 2 \cdot V_2{}^p + \ldots \approx 2 \cdot V_0{}^n \tag{2.57}$$

So, by adding or subtracting the responses of an open and a shorted case, we are able to plot the waves that are traveling in either the positive or the negative z-direction, as shown in figure 2.6.
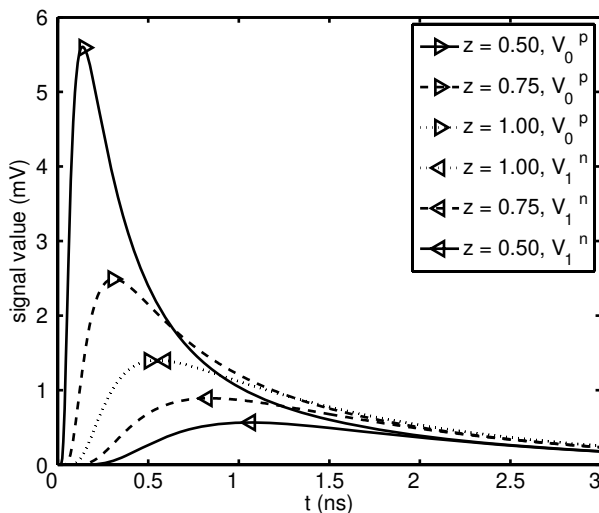
Figure 2.6: Impulse response for different positions along the interconnect decomposed in a wave traveling in the positive z-direction and a wave traveling in the negative z-direction.

2.4.4 S-parameters

The notion that the response of an RC interconnect can be thought of as a summation of waves traveling in both the positive and negative z-direction, is reflected in the s-parameter model of an interconnect.

The s-parameters of a two-port network are given by [45]:

$$
\begin{aligned}
b_1 &= s_{11} \cdot a_1 + s_{12} \cdot a_2 \\
b_2 &= s_{21} \cdot a_1 + s_{22} \cdot a_2
\end{aligned}
\tag{2.58}
$$

The incident (a) and scattered (b) waves are defined with respect to a reference impedance Z_0.

$$
a_k = \frac{V_k{}^p}{\sqrt{Z_0}} = \frac{1}{2} \cdot \left(\frac{V_k}{\sqrt{Z_0}} + \sqrt{Z_0} \cdot I_k \right)
\tag{2.59}
$$

$$
b_k = \frac{V_k{}^n}{\sqrt{Z_0}} = \frac{1}{2} \cdot \left(\frac{V_k}{\sqrt{Z_0}} - \sqrt{Z_0} \cdot I_k \right)
\tag{2.60}
$$

$V_k{}^p$ and $V_k{}^n$ are the incident and scattered voltage waves at port k, while V_k and I_k are the terminal voltage and current at port k.

If a_k and b_k are known, the terminal voltage and current can be calculated easily as

$$
V_k = \sqrt{Z_0} \cdot (a_k + b_k)
\tag{2.61}
$$

$$
I_k = \frac{a_k - b_k}{\sqrt{Z_0}}
\tag{2.62}
$$

The s-parameter model of an interconnect is shown in figure 2.7. The s-parameters give the relation between the incident waves a_1 and a_2 and the scattered waves b_1 and b_2. The values of the s-parameters are given by [45]:

$$
\begin{aligned}
S &= \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \\
&= \frac{1}{2 Z_0 Z_c \cosh(\gamma l) + (Z_c{}^2 + Z_0{}^2) \sinh(\gamma l)} \times \\
&\quad \begin{bmatrix} (Z_c{}^2 - Z_0{}^2) \sinh(\gamma l) & 2 Z_0 Z_c \\ 2 Z_0 Z_c & (Z_c{}^2 - Z_0{}^2) \sinh(\gamma l) \end{bmatrix}
\end{aligned}
\tag{2.63}
$$
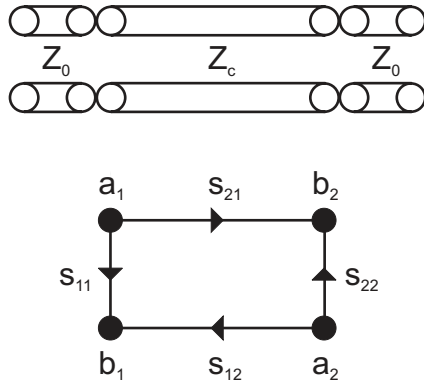
Figure 2.7: S-parameter model of an interconnect.

The propagation constant γ and characteristic impedance Z_c are defined in (2.45) and (2.46) respectively. Note that by choosing Z_0 equal to Z_c both s_{11} and s_{22} are zero and s_{12} and s_{21} reduce to

$$s_{21} = s_{12} = e^{-\gamma \cdot l} \tag{2.64}$$

The interconnect is on one side connected to a transmitter, modeled as a voltage source with source impedance Z_S. On the other side, the interconnect is connected to a receiver, modeled as a load impedance Z_L. The s-parameter model of the transmitter and receiver are [45]:
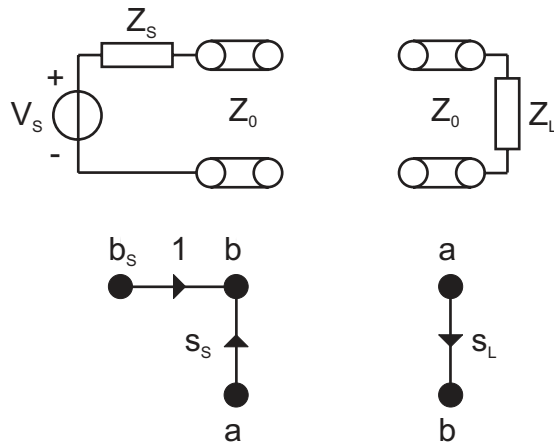
Figure 2.8: S-parameter model of transmitter and receiver.

The values of the s-parameters of the source and load impedance are

$$s_S = \frac{Z_S - Z_0}{Z_S + Z_0} \tag{2.65}$$

$$s_L = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{2.66}$$

and the source voltage appears in b_S:

$$b_S = V_S \cdot \frac{\sqrt{Z_0}}{Z_0 + Z_S} \tag{2.67}$$

Combining figures 2.7 and 2.8 and choosing $Z_0 = Z_c$, the model of figure 2.9 represents the interconnect connected between a transmitter and receiver.
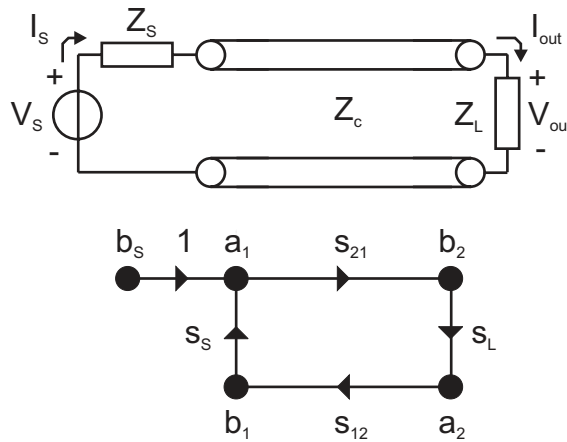
Figure 2.9: s-parameter model of an interconnect with transmitter and receiver.

2.4.5 Transfer functions

We will use the s-parameter model to calculate the transfer function V_{out}/V_S of the interconnect. In the following, Z_0 is always chosen equal to Z_c. Rewriting (2.67) gives the input voltage V_S of the interconnect.

$$V_S = b_S \cdot \frac{Z_c + Z_S}{\sqrt{Z_c}} \tag{2.68}$$

The output voltage V_{out} can be calculated with the help of (2.59) and (2.60) as

$$V_{out} = \sqrt{Z_c} \cdot (a_2 + b_2) \tag{2.69}$$

The transfer function from V_S to V_{out} can now be calculated as

$$H_{VV} = \frac{V_{out}}{V_S} = \frac{Z_c}{Z_c + Z_S} \cdot \left(\frac{a_2}{b_S} + \frac{b_2}{b_S}\right) \tag{2.70}$$

Both a_2/b_S and b_2/b_S can be found with the help of Mason's rule [45].

$$\frac{a_2}{b_S} = \frac{s_{21} \cdot s_L}{1 - s_{21} \cdot s_L \cdot s_{12} \cdot s_S} \tag{2.71}$$

$$\frac{b_2}{b_S} = \frac{s_{21}}{1 - s_{21} \cdot s_L \cdot s_{12} \cdot s_S} \tag{2.72}$$

The transfer function thus is

$$H_{VV} = \frac{V_{out}}{V_S} = \frac{Z_c}{Z_c + Z_S} \cdot \frac{s_{21} \cdot (1 + s_L)}{1 - s_{21} \cdot s_L \cdot s_{12} \cdot s_S} \tag{2.73}$$

with s_{21}, s_L, s_{12} and s_S as defined in the previous section.

Of course, also other transfer functions can be calculated.

$$H_{VI} = \frac{I_{out}}{V_S} = \frac{1}{Z_c + Z_S} \cdot \frac{s_{21} \cdot (1 - s_L)}{1 - s_{21} \cdot s_L \cdot s_{12} \cdot s_S} \tag{2.74}$$

$$H_{IV} = \frac{V_{out}}{I_S} = Z_c \cdot \frac{s_{21} \cdot (1 + s_L)}{1 - s_{21} \cdot s_L \cdot s_{12}} \tag{2.75}$$

$$H_{II} = \frac{I_{out}}{I_S} = \frac{s_{21} \cdot (1 - s_L)}{1 - s_{21} \cdot s_L \cdot s_{12}} \tag{2.76}$$

As an example, these four transfer functions are plotted in figure 2.10 for the example interconnect from section 1.5.2. The transfer function H_{VV} is the same as we saw already in figure 1.10. For high frequencies, the magnitude of H_{VV} drops and the bandwidth is limited. The transfer function H_{VI} shows that both for low and for high frequencies, no current is flowing through the load capacitance. For low frequencies, the obvious reason for this is that a capacitance has high impedance for low frequencies, while for high frequencies, the voltage across the load capacitance is small. The transfer function H_{IV} first drops with -20 dB/decade due to integration of the current in the interconnect capacitance. For higher frequencies, the drop rate becomes much more than -20 dB/decade. H_{II} has the same shape as H_{VV}. Interestingly, the bandwidth of H_{II} is larger than the bandwidth of H_{VV}.
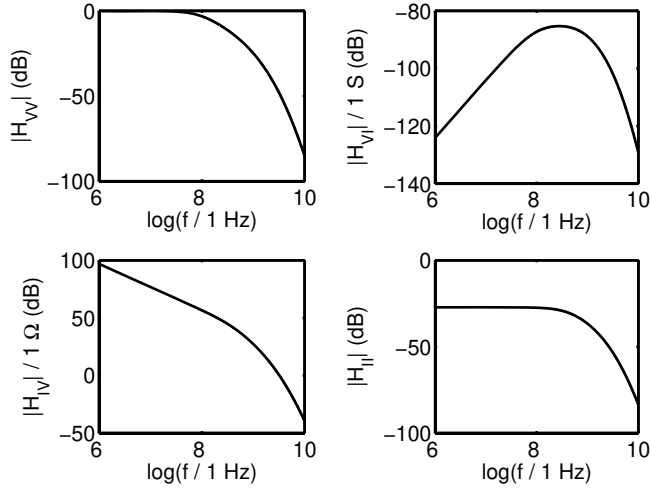
Figure 2.10: Transfer functions of the example interconnect from section 1.5.2.

2.4.6 Power consumption

In order to calculate the power consumption, the current I_S that is drawn from the voltage source V_S needs to be found. I_S/V_S can be calculated with the help of the s-parameter model. With (2.59) and (2.60), we can find

$$I_S = \frac{a_1 - b_1}{\sqrt{Z_c}} \tag{2.77}$$

and with (2.67)

$$Y_{in} = \frac{I_S}{V_S} = \frac{1}{Z_c + Z_S} \cdot \frac{a_1 - b_1}{b_S} \tag{2.78}$$

With Mason's rule [45]

$$Y_{in} = \frac{1}{Z_c + Z_S} \cdot \frac{1 - s_{21} \cdot s_L \cdot s_{12}}{1 - s_{21} \cdot s_L \cdot s_{12} \cdot s_S} \tag{2.79}$$

In order to find the power, Y_{in} has to be multiplied by the power spectral density (PSD) of the input voltage signal. The PSD of the input voltage signal can be calculated as follows.

We start with a 'binary (zero-mean) Markov source' which is characterized by:

$$\left.\begin{array}{l}\Pr(a_{n+1} \neq a_n) = p_{trans} \\ \Pr(a_{n+1} = a_n) = 1 - p_{trans}\end{array}\right\} \quad a = \{1, -1\} \tag{2.80}$$

Pr(x) is the probability that x is true and p_{trans} is the transition probability. The PSD of such a Markov source can be written as [46]

$$S_{MM}(f) = \frac{1}{T_S} \cdot \left| H_M(e^{j \cdot 2 \cdot \pi \cdot f \cdot T_S}) \right|^2 \cdot (1 - (1 - 2 \cdot p_{trans})^2) \tag{2.81}$$

with

$$H_M(e^{j \cdot 2 \cdot \pi \cdot f \cdot T_S}) = \frac{1}{e^{j \cdot 2 \cdot \pi \cdot f \cdot T_S} - (1 - 2 \cdot p_{trans})} \tag{2.82}$$

and T_S the symbol time. By multiplying $S_{MM}(f)$ with $|H_S(f)|^2$, the PSD of the input voltage signal is found. $H_S(f)$ is the Fourier transfer of the symbol response.

$$S_{SS}(f) = |H_s(f)|^2 \cdot \frac{1}{T_S} \cdot \left| H_m(e^{j \cdot 2 \cdot \pi \cdot f \cdot T_S}) \right|^2 \cdot (1 - (1 - 2 \cdot p_{trans})^2) \tag{2.83}$$

The power that is consumed in the interconnect and its termination impedances can be found by multiplying the PSD of the input voltage signal by Y_{in} and integrating over frequency.

$$P = \int_{f=-\infty}^{\infty} real\left(\frac{S_{SS}(f)}{Z_{in}(f)}\right) \cdot df = 2 \cdot \int_0^{\infty} S_{SS}(f) \cdot \frac{Y_{in}(f) \cdot Y_{in}(f)^*}{real(Y_{in}(f))} \cdot df \tag{2.84}$$

As an illustration, the power consumption of the example interconnect from section 1.5.2 is shown in figure 2.11. As we are interested in the highest achievable data rate for minimum power consumption, the power consumption is divided by the data rate. This energy per bit is plotted for plain binary signaling against the transition probability for different data rates.

The energy that is dissipated per bit is larger for a higher transition probability. This makes sense, as with more transitions, the interconnect is charged and discharged more often. For high data rates and high transition probabilities, the receiver end of the interconnect is not fully charged and discharged anymore. Therefore, the energy per bit at high transition probabilities is lower for high data rates. Note, that for highest information density, the transition probability should be 0.5 [47].
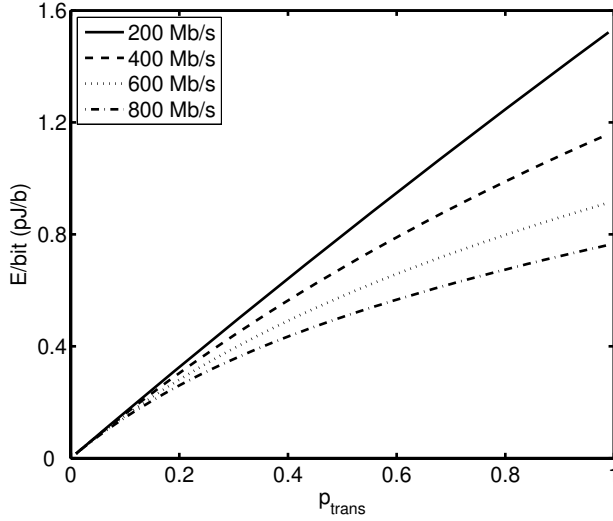
Figure 2.11: Energy per bit as a function of transition probability and data rate.

2.4.7 Additional power consumption in receiver

In the previous section, the power calculations have assumed signaling with normalized bipolar signals (-1 V and 1 V). However, a transceiver in CMOS will in general operate between the supply voltages V_{DD} and GND. Fortunately, from a power consumption perspective, the two schematics in figure 2.12 are equivalent. As the probability of the transmitter being connected to V_{DD} is equal to the probability of the transmitter being connected to GND, the net power delivered by the (conceptual) voltage source $V_{DD}/2$ is zero.

However in general, the practical implementation of this voltage source gives additional power consumption. If we assume that there is no resistance in the GND line, we can place the conceptual voltage source below the load impedance and in a practical receiver, this may be implemented as shown in figure 2.13. With this implementation, additional current is flowing from V_{DD}, equal to

$$I_{DD} = \frac{V_{DD}}{4 \cdot \mathrm{Re}(Z_L)} - \frac{I_{out}}{2} \tag{2.85}$$

For equal probabilities of a positive and a negative I_{out}, the extra power delivered by V_{DD} is

$$P_{RX} = \frac{{V_{DD}}^2}{4 \cdot \mathrm{Re}(Z_L)} \tag{2.86}$$

This power has to be added to the power as calculated with (2.84), if the implementation of figure 2.13 is used.
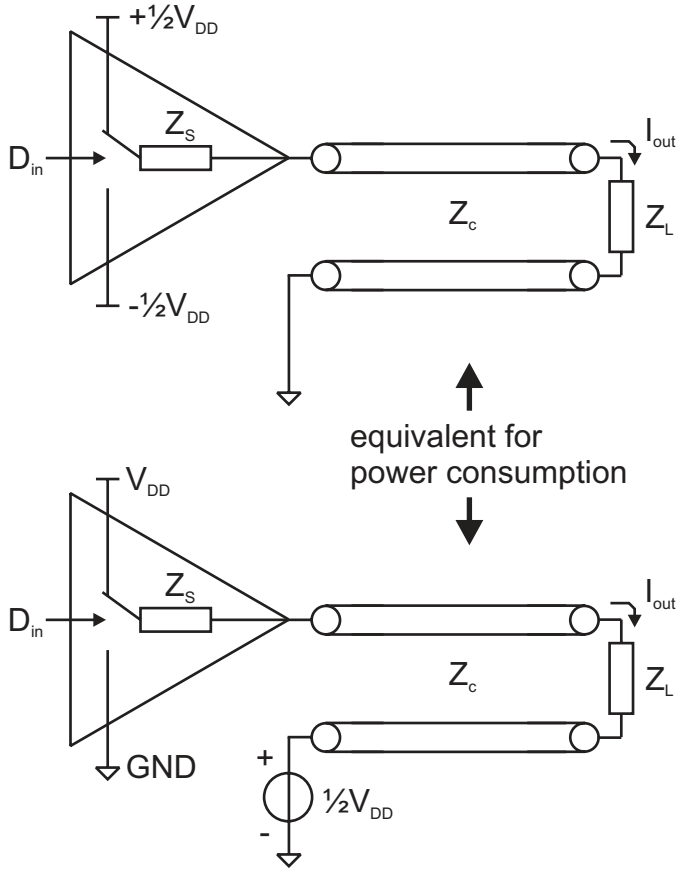
Figure 2.12: Power consumption is equivalent for transceivers operating between $+V_{DD}/2$ and $-V_{DD}/2$ or between GND and V_{DD}.
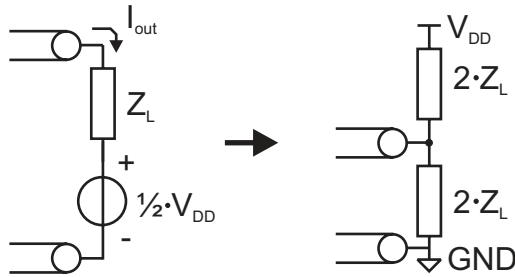
Figure 2.13: A practical implementation of the receiver with impedances to both supply voltages.

2.5 Expanded distributed model

Until now, we have used a distributed model for a single interconnect, with all other metal connected to ground. However, often interconnects have neighboring interconnects running in parallel, for instance in a bus structure (see interconnect 1, 2 and 3 in figure 2.14).
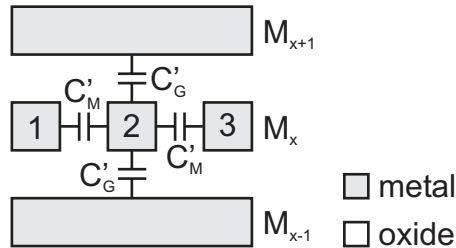
Figure 2.14: Capacitances to ground (C_G') and neighboring interconnects (C_M').

The distributed model of figure 2.4 can be expanded to the model of figure 2.15. Two interconnects are shown (for instance 1 and 2), with a ground line in between. Note however, that in the physical structure this ground line is above and below the interconnect (see figure 2.14).
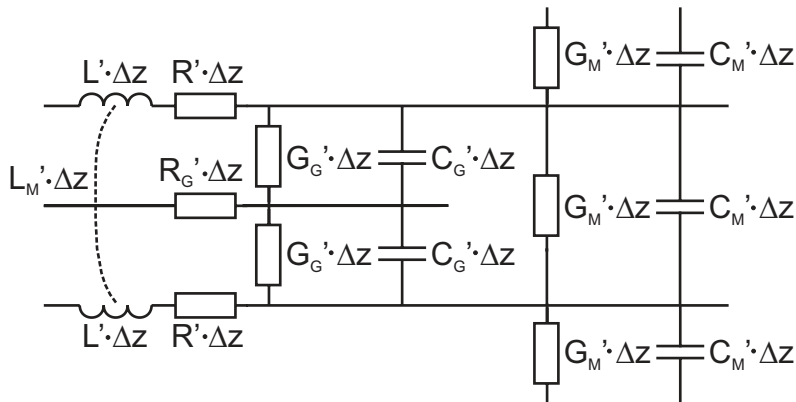
Figure 2.15: Distributed model for interconnect with parallel neighboring interconnects

Both interconnects have a distributed resistance R' and a distributed inductance L' and the ground line has a distributed resistance R_G'. There is also a distributed mutual inductance L_M' between the two interconnects. The distributed capacitance C_G' and the distributed conductance G_G' are connected to the ground line, while the distributed mutual capacitance C_M' and distributed mutual conductance G_M' are connected to neighboring interconnects.

Because of the mutual capacitance, inductance and conductance and the resistance of the common ground, there will be crosstalk between two neighboring interconnects. Figure 2.16 shows two neighboring interconnects.
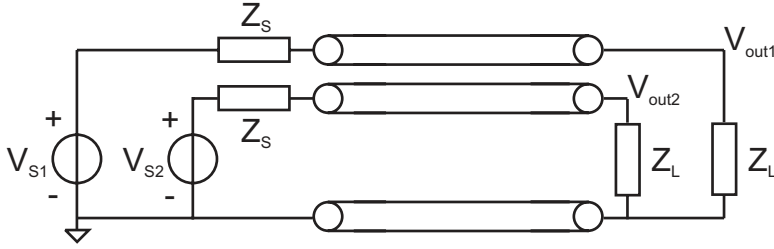
Figure 2.16: Two neighboring interconnects.

We assume that both interconnects are equal. V_{out1} and V_{out2} are a function of both V_{S1} and V_{S2}.

$$V_{out1} = H \cdot V_{S1} + H_X \cdot V_{S2}$$
$$V_{out2} = H \cdot V_{S2} + H_X \cdot V_{S1}$$
(2.87)

In this section, we calculate the transfer function H and the crosstalk transfer function H_X. In order to do that, we will use a modal analysis. We will look at two cases.

<u>Even case:</u> $V_{S1} = V_{S2}$.
<u>Odd case:</u> $V_{S1} = -V_{S2}$.

In the even case,

$$V_{out1} = (H + H_X) \cdot V_{S1} = H_{even} \cdot V_{S1}$$
$$V_{out2} = (H + H_X) \cdot V_{S2} = H_{even} \cdot V_{S2}$$
(2.88)

while in the odd case

$$V_{out1} = (H - H_X) \cdot V_{S1} = H_{odd} \cdot V_{S1}$$
$$V_{out2} = (H - H_X) \cdot V_{S2} = H_{odd} \cdot V_{S2}$$
(2.89)

If we are able to calculate H_{even} and H_{odd}, we can use

$$H = \frac{1}{2} \cdot (H_{even} + H_{odd})$$
$$H_X = \frac{1}{2} \cdot (H_{even} - H_{odd})$$
(2.90)

to calculate H and H_X. The transfer functions H_{even} and H_{odd} can be calculated in almost the same way as was done as in section 2.4.5. Only, the parameters R', L', C' and G' should be replaced by R_{even}', L_{even}', C_{even}' and G_{even}' in the even case and by R_{odd}', L_{odd}', C_{odd}' and G_{odd}' in the odd case.

42

These even and odd parameters differ from the standard RLCG parameters. In the even case, the mutual capacitance and mutual conductance to the neighboring interconnect is not seen, since the voltage across these mutual impedances is zero. In the odd case on the other hand, the voltage across these impedances is effectively doubled, thus also doubling the effective mutual capacitance and conductance (Miller multiplication). The mutual inductance should be added to the inductance of the interconnect for the even case and subtracted for the odd case. The resistance of the ground line is only seen in the even case, as for the odd case all current is returned in the neighboring interconnect.

These considerations are summarized in the equations below. For simplicity, we assume that the C_M' and the G_M' to other neighboring interconnects can be placed in parallel with C_G' and G_G'. Then, in even mode

$$R_{even}' = R' + 2 \cdot R_G' \tag{2.91}$$
$$L_{even}' = L' + L_M' \tag{2.92}$$
$$C_{even}' = 2 \cdot C_G' + C_M' \tag{2.93}$$
$$G_{even}' = 2 \cdot G_G' + G_M' \tag{2.94}$$

while in odd mode:

$$R_{odd}' = R' \tag{2.95}$$
$$L_{odd}' = L' - L_M' \tag{2.96}$$
$$C_{odd}' = 2 \cdot C_G' + 3 \cdot C_M' \tag{2.97}$$
$$G_{odd}' = 2 \cdot G_G' + 3 \cdot G_M' \tag{2.98}$$

By using these parameters with the equations of section 2.4.5, both H_{even} and H_{odd} can be calculated. With H_{even} and H_{odd} in turn, the transfer function H of the interconnect and the crosstalk transfer function to a neighboring interconnect H_X can be calculated.

As an example, these two transfer functions are plotted in figure 2.17 for the example interconnect from section 1.5.2. For low frequencies, the magnitude of the crosstalk transfer function is small, while for high frequencies, the crosstalk transfer function H_X approaches the transfer function H.

2.6 Lumped versus distributed models

If we want to simulate an interconnect in a circuit simulator, we cannot use the distributed model as described before, because we cannot use an infinite number of distributed elements. Instead, we can use a lumped model with n elements that have a value of $l \cdot R'/n$, $l \cdot L'/n$, $l \cdot C'/n$ and $l \cdot G'/n$ with l the length of the interconnect. The question can be asked: how many elements do we need for an accurate simulation?

Figure 2.18 shows the transfer function of an interconnect with lumped elements for different values of n. For this figure, $l \cdot R' = 1.5$ kΩ, $l \cdot L' = 0$ H, $l \cdot C' = 2.2$ pF, $l \cdot G' = 0$ S, $Z_S = 0$ Ω and $Z_L = \infty$ Ω.
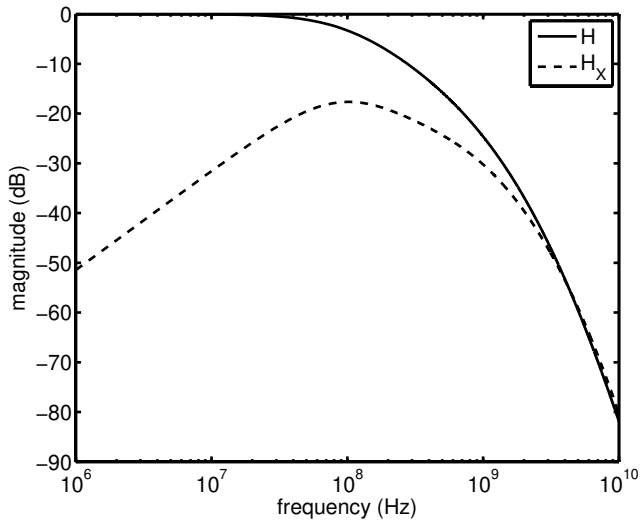
Figure 2.17: Transfer function and crosstalk transfer function.
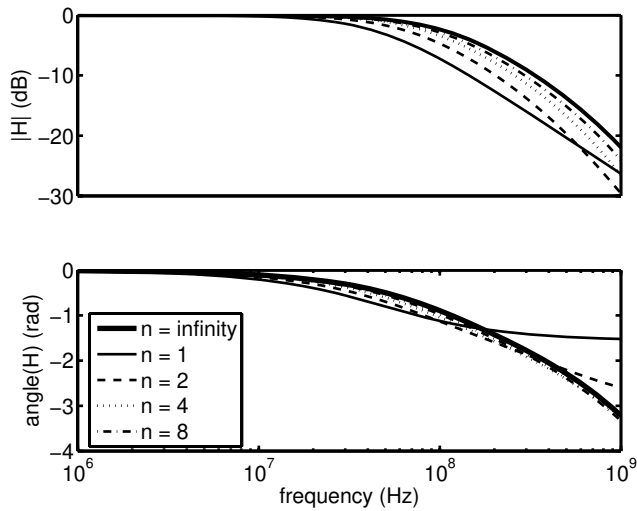
Figure 2.18: Transfer functions for lumped elements model for different number of elements.

As can be seen in the figure, the higher n is, the better the transfer function matches to the distributed transfer function (n = infinity).

However, the bandwidth of the lumped elements model is always too small. In order to correct for this, we can use element values of $a \cdot l \cdot R'/n$ and $a \cdot l \cdot C'/n$. The value of a can then be tuned to have the first pole at the correct position, as shown in figure 2.19.
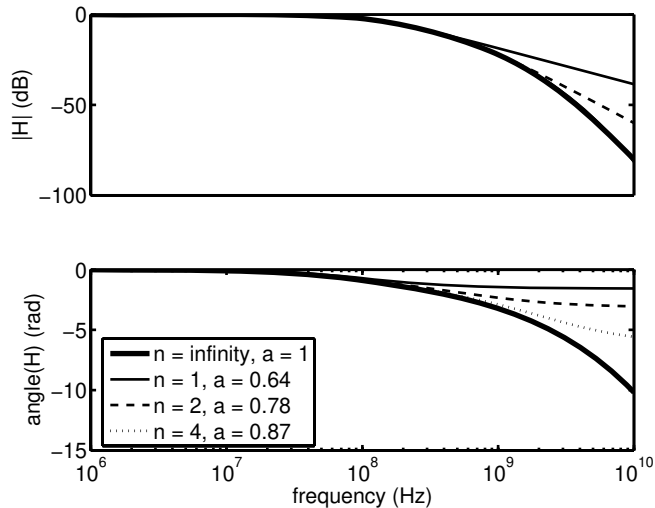
Figure 2.19: Transfer functions for lumped elements model for different number of elements with elements scaled with a factor a.

The lines with n = infinity and n = 4 in the figure are on top of each other. The figure shows the dominant first-order RC behavior of an interconnect. The transfer function can already be approximated with a single RC-section with only 0.5 dB error at three times the bandwidth. Note, that the factor a also depends on Z_S and Z_L [48].

2.7 3D EM-field simulator

With the results of the previous sections, we are able to calculate (crosstalk) transfer functions and power consumption. However, in order to do this we need to know the values of the parameters of figure 2.15. We could use the analytical expressions of section 2.4.2, but they are not very accurate as they neglect fringe fields and assume all fields to be concentrated in four areas. In order to find more accurate values, a 3D EM-field simulator [43] is used. This simulator calculates the electric and magnetic fields of a structure that can be drawn in the program. The simulator outputs the s-parameters of the drawn structure. Figure 2.20 shows an example structure.

The s-parameters from the simulator can be mapped on the model of figure 2.15. Therefore, we write the propagation constant and the characteristic impedance of section 2.4.4 as a function of the s-parameters. The relation between γ and Z_c and the s-parameters s_{11} and s_{21} are [49]:
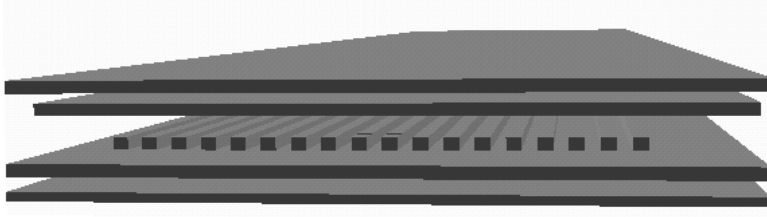
Figure 2.20: Example structure of interconnects in the 3D EM-field simulator [43].

$$e^{-\gamma l} = \left(\frac{1 - s_{11}^2 + s_{21}^2}{2 \cdot s_{21}} \pm K \right)^{-1} \tag{2.99}$$

$$K = \sqrt{\frac{(s_{11}^2 - s_{21}^2 + 1)^2 - (2 \cdot s_{11})^2}{(2 \cdot s_{21})^2}} \tag{2.100}$$

$$Z_C = \sqrt{Z_0^2 \cdot \frac{(1 + s_{11})^2 - s_{21}^2}{(1 - s_{11})^2 - s_{21}^2}} \cdot \tag{2.101}$$

The parameters R', L', C' and G' are found by using (2.45) and (2.46):

$$R' = \mathrm{Re}(\gamma \cdot Z_c) \tag{2.102}$$

$$L' = \frac{\mathrm{Im}(\gamma \cdot Z_c)}{\omega} \tag{2.103}$$

$$C' = \frac{\mathrm{Im}(\dfrac{\gamma}{Z_c})}{\omega} \tag{2.104}$$

$$G' = \mathrm{Re}(\frac{\gamma}{Z_c}) \tag{2.105}$$

In this way, the RLCG parameters are found. By doing an even and an odd mode simulation of the drawn interconnect structure, also the mutual parameters R_G', L_M', C_M' and G_M' can be found with the help of (2.91) - (2.98).

2.8 Summary

− The interconnects are modeled with a distributed resistance, inductance, conductance and capacitance. With the help of s-parameters, transfer functions and power consumption can be calculated.
− Crosstalk can be modeled by adding a distributed ground resistance, mutual inductance, mutual conductance and mutual capacitance to the general model. By doing an even and odd mode analysis, crosstalk transfer functions can be found.
− For simulation in a circuit simulator, a lumped model can be used. A single RC-section already models the interconnect accurately as the interconnect has a dominant first-order pole.
− The distributed parameters can be found by mapping the results of a 3D EM-field simulator on the s-parameter model.

Chapter 3

Interconnect design and termination concepts

3.1 Introduction

In the previous chapter, we discussed the modeling of global interconnects. Now, we come to the design of the interconnects. We design the interconnects for maximum achievable data rate and data integrity and for minimum power and optimized area consumption, as discussed in chapter 1. In order to calculate transfer functions, eye-diagram properties and power consumption, we use the distributed models of chapter 2. First-order models are given to explain the results intuitively.

In the next section, we find the optimal dimensions for the interconnects. After that, we examine how to terminate the interconnects. Termination concepts are given, at the transmitter side in section 3.3 and at the receiver side in section 3.4. In section 3.5, both termination concepts are compared to each other. Section 3.6 discusses the choice of the parameters of these concepts and the sensitivity to variations in these parameters. Finally, in section 3.7 some alternative solutions are discussed.

3.2 Optimal dimensions

3.2.1 Introduction
As already shown in section 1.5.5, an interconnect has certain physical dimensions. Figure 3.1 shows these dimensions again.
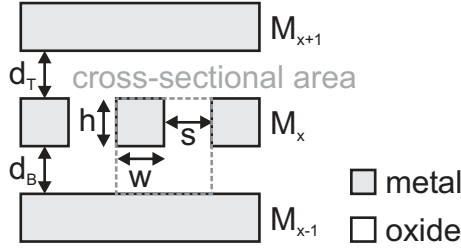
Figure 3.1: Interconnect dimensions.

A designer cannot choose the height h and oxide thicknesses d_T and d_B, but he can choose the width w and spacing s. In this section, we will find optimal values for these two dimensions.

3.2.2 LC or RC

In section 2.4.3, we discussed the possibility of an LC and an RC dominated global interconnect. In the first case, the resistance can be neglected, while in the second case the inductance can be neglected. We saw that only in the RC case, there was attenuation and distortion. Therefore, the LC case would be favorable. Thus we ask the question: can we choose the dimensions of the interconnect in such a way, that the inductance dominates over the resistance?

The resistance can be ignored if the condition $\omega \cdot L' \gg R'$ is fulfilled for frequencies above the RC bandwidth of the interconnect. In other words,

$$\frac{R'}{L'} < \frac{2}{l^2 \cdot R' \cdot C'}$$

(3.1)

where for the bandwidth, (3.19) is used.

Both R'/L' and the bandwidth are plotted in figures 3.2 and 3.3 for a 10 mm long interconnect as a function of the width w and spacing s. For R', L' and C' (2.37), (2.38) and (2.40) are used. The height h and oxide thickness d are 0.35 μm and 0.36 μm respectively, the conductance of the metal $\sigma_m = 4.75 \cdot 10^7$ (m·Ω)$^{-1}$, the relative permittivity $\varepsilon_r = 3.7$ and the relative permeability $\mu_r = 1$. These are the values for an interconnect in metal layer four in a 0.13 μm CMOS process, one of the processes that is used in this project.

With both w and s small, the R'/L'-frequency is the highest. At the same time, the bandwidth is the lowest. In order to have an LC dominated interconnect, we have to use a larger w or s. By increasing s, the resistance stays equal, the inductance is increasing and the capacitance is decreasing. However, both the inductance and capacitance values saturate at a certain value. Therefore, for very large s, the R'/L'-frequency is not decreasing and the bandwidth is not increasing anymore. By increasing w, both the resistance and inductance decrease, while the capacitance increases. First, the resistance is decreasing faster than both the inductance decreases and the capacitance increases. But for higher values of w, they will all decrease or increase with the same speed and the R'/L'-frequency and the
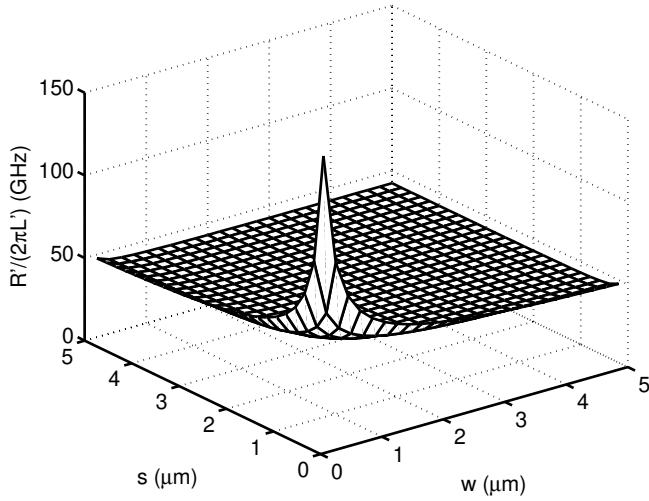
Figure 3.2: R'/L' as a function of the dimensions w and s.
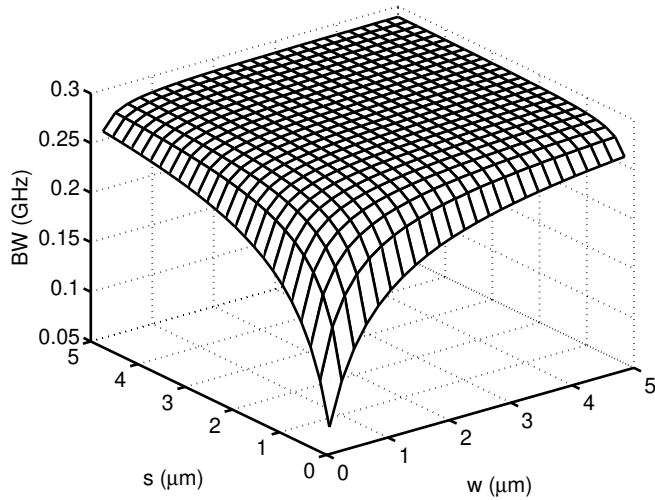
Figure 3.3: Bandwidth as a function of the dimensions w and s.

bandwidth become constant. Thus, increasing w or s only helps for a limited range. After that, both the R'/L'-frequency and the bandwidth saturate.

The final values of the R'/L'-frequency and the bandwidth can be calculated with the help of (2.37), (2.38) and (2.40) and letting w or s go to infinity:

$$\left. \frac{R'}{L'} \right|_{w \to \infty} = \frac{2}{h \cdot d \cdot \sigma_m \cdot \mu} \tag{3.2}$$

$$BW|_{w \to \infty} = \frac{h \cdot d \cdot \sigma_m}{l^2 \cdot \varepsilon} \tag{3.3}$$

In figures 3.2 and 3.3, the values of the R'/L'-frequency and the bandwidth for w→∞ are 42 GHz and 291 MHz respectively. Thus, the bandwidth is always lower than the R'/L'-frequency and the interconnect will show RC behavior, even for large w and s. This is also shown with the transfer function in figure 3.4.
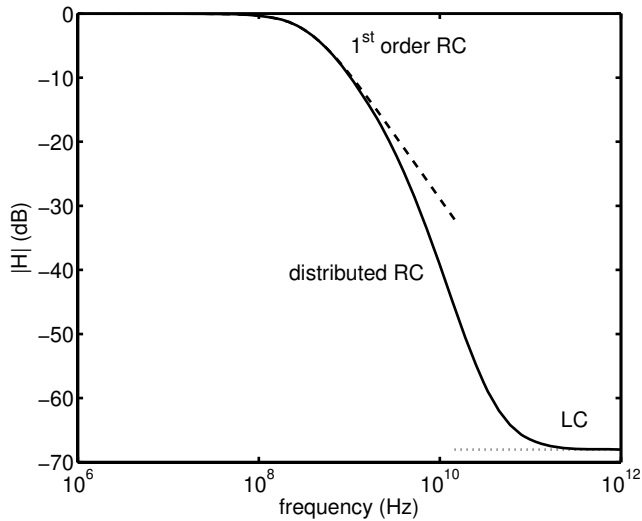
Figure 3.4: Transfer function for w→∞, showing three regions.

The transfer function has three regions. In the first region, the transfer function is dominated by 1st order RC behavior. The second region shows a large roll-off due to the distributed RC and in the third region, LC behavior dominates. However, because the R'/L'-frequency is much larger than the bandwidth, the attenuation is already large in this LC region. Therefore, even for an extremely large w or s, this example interconnect will be dominated by (first-order) RC behavior.

For smaller lengths, the bandwidth increases with $1/l^2$. The R'/L'-frequency does not depend on l and both the R'/L'-frequency and the bandwidth are equal for a length of 0.8 mm. In order to have LC behavior, the interconnect should have a length below this 0.8 mm.

The only way for longer lengths to have an LC dominated interconnect, is to increase h and d. Thus, thick top level metal for a large h and one or more metal layers below should have no metal at all for a large d. As an example, if h = d, h should be increased from 0.35 μm to

1.2 μm in order to have the R'/L'-frequency equal to the bandwidth in the example above. Still, a very large w and s are assumed.

In summary, for long (global) interconnects we can only make an interconnect that is dominated by LC behavior by using very large dimensions (width, spacing, oxide thickness and height)

3.2.3 Optimized dimensions

In the previous section, we have seen that creating an interconnect with LC behavior will cost considerable area. Not only do we have to use very thick metal layers, which could be reserved for the power and clock grid, but we should also keep metal layers below the interconnect free of metal to create a large dielectric thickness (d in figure 3.1). Together with a large width and spacing, the cross-sectional area will be large.

Therefore, in this thesis we will only look at RC-dominated interconnects, that have a much smaller cross-sectional area consumption. As discussed in section 1.5.5, we will use optimal dimensions for which the bandwidth per cross-sectional area is largest. A bus with these optimized interconnects will have the highest aggregate data rate for a certain bus area.

The bandwidth per cross-sectional area can be expressed as a function of the dimensions (see figure 3.1) [6]. The bandwidth is inversely proportional to R' and C'.

$$BW \propto \frac{1}{R' \cdot C'} \tag{3.4}$$

The cross-sectional area is given by (w+s)·(h+d)

$$\frac{BW}{area} \propto \frac{1}{R' \cdot C'} \cdot \frac{1}{(w+s) \cdot (h+d)} \tag{3.5}$$

Equations for R' and C' are given in (2.37) and (2.40) respectively. These can now be used to find the bandwidth per cross-sectional area.

$$\frac{BW}{area} \propto \frac{1}{\dfrac{w+s}{h} + \dfrac{w+s}{d} + \dfrac{h+d}{w} + \dfrac{h+d}{s}} \tag{3.6}$$

The bandwidth per cross-sectional area is maximized if all partial derivatives are zero. This is the case for w = s = h = d. So ideally, all dimensions should be chosen equal. However, in practice the optimum can shift a bit: h, d_T and d_B are fixed by the process and are not necessarily equal. Also, fringe capacitances play a role and if differential signaling is used, the capacitances to the other differential half is effectively doubled due to the Miller effect. We will come back to this in chapter 5, when we discuss the dimensions of the interconnects on our test chips.

Instead of dimensioning the interconnects for maximum bandwidth per area, we can also dimension the interconnect for minimum energy per bit. For this, we assume that the E/bit is proportional to the capacitance of the interconnect.

$$\frac{E}{bit} \propto C'$$

$$\frac{E}{bit} \propto \frac{w}{d} + \frac{h}{s}$$

(3.7)

For minimum energy per bit, the width of the interconnect should be minimized, while the spacing between the interconnects should be maximized. In practice this means to have minimum width interconnects with twice or three times the minimum spacing. A larger spacing only has a small power gain for large area consumption.

3.3 Source impedance

3.3.1 Introduction

Once the dimensions of the interconnect are determined, we can look at the termination of the interconnect. For this, we look at the impedances Z_S and Z_L (see figure 1.7). The question that is to be answered is: can we increase the achievable data rate with the help of these impedances? We will see that this is indeed possible. Moreover, also the power consumption can be decreased. In this section, we look at the source impedance, while the load impedance is treated in the next section. For now, we use $Z_L = 100$ fF.

3.3.2 Ideal source impedance

For LC transmission lines, both source and load impedance are in general chosen equal to the characteristic impedance Z_c. This is done to avoid reflections. However, for RC-dominated interconnects it is better not to terminate with Z_c. Figure 3.5 compares the transfer function of the example interconnect from section 1.5.2 for $Z_S = Z_c$ and $Z_S = 100\ \Omega$.

Interestingly, the bandwidth with $Z_S = 100\ \Omega$ is larger than with $Z_S = Z_c$. Thus, the bandwidth is increased by choosing a termination impedance that is not equal to Z_c. So, if Z_c is not the best termination impedance for RC-dominated interconnects, what would be the ideal source impedance?

An ideal source impedance would result in the distortionless transfer function

$$H_{ideal} = A \cdot e^{-j \cdot \omega \cdot \sqrt{L' \cdot C'} \cdot l}$$

(3.8)

This ideal transfer function has a constant attenuation A for all frequencies and a linear phase-frequency relation. By comparing the ideal transfer function with the transfer function of an interconnect (see section 2.4.5), we can calculate the ideal Z_S for a given Z_L. This is done for the example interconnect from section 1.5.2 for two values of A in figure 3.6 ($Z_L = 100$ fF).
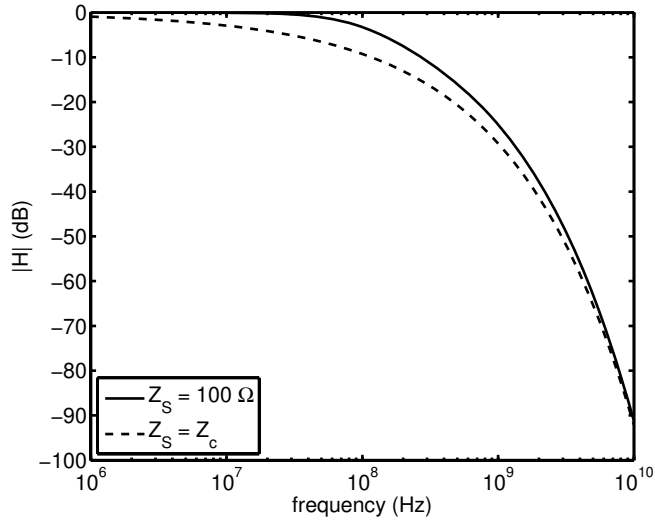
Figure 3.5: Transfer function of the example interconnect from section 1.5.2 for $Z_S = 100$ Ω and $Z_S = Z_c$.
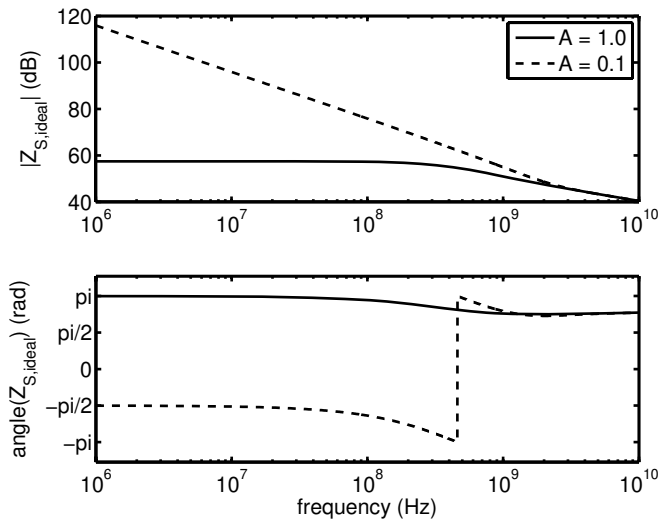
Figure 3.6: Ideal source impedance for the example interconnect from section 1.5.2.

In order to have a flat transfer function with amplitude $A = 1$, the source impedance has to be a negative resistance over a large frequency range. This will be difficult to implement. On the other hand, the ideal Z_S with $A = 0.1$ is very comparable to a capacitance, in this example of 250 fF. Thus, a capacitive transmitter, as shown in figure 3.7, will give a more ideal transfer function.
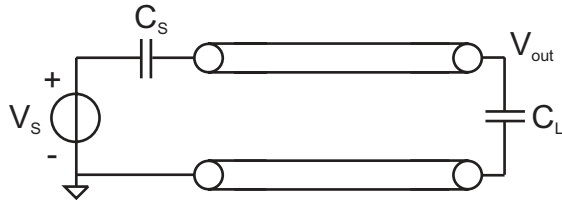
Figure 3.7: Capacitive transmitter circuit.

The transfer functions for different values of the capacitance are given in figure 3.8. The values of C_S are given as a fraction of $C_T = C' \cdot l$, which is the total capacitance of the interconnect.
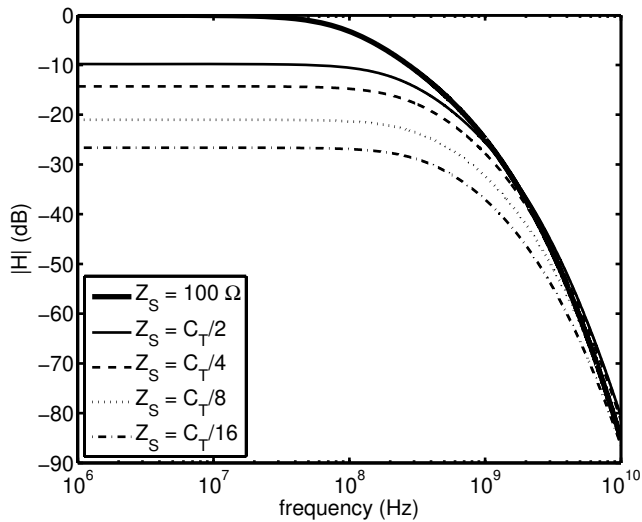
Figure 3.8: Transfer functions of the example interconnect from section 1.5.2 with a series capacitance as source impedance.

The figure shows that the series capacitance C_S in figure 3.7 indeed increases the bandwidth of the interconnect. Furthermore, the smaller C_S is chosen, the higher the bandwidth becomes.

Note that the voltage swing at V_{out} is also decreased. So C_S does not only increase the bandwidth, but also creates a low voltage swing on the interconnect. This low voltage swing reduces power consumption. Thus, the capacitive source impedance has two advantages: it both increases the bandwidth and decreases power consumption.

3.3.3 First-order model

As was discussed in section 2.6, the interconnect can be modeled with a first-order RC network, where both $R_T = l \cdot R'$ and $C_T = l \cdot C'$ are scaled with a factor a. The transfer function with zero source impedance and infinite load impedance is:

$$H\big|_{Z_S=0} = \frac{1}{1 + j \cdot \omega \cdot R_T \cdot C_T \cdot a^2} \tag{3.9}$$

With a capacitive source impedance C_S, the transfer function changes to

$$H\big|_{Z_S=C_S} = \frac{\dfrac{C_S}{C_S + C_T}}{1 + j \cdot \omega \cdot R_T \cdot C_T \cdot a^2 \cdot \dfrac{C_S}{C_S + C_T}} \tag{3.10}$$

The smaller the capacitance C_S, the higher the bandwidth. In practice however, if C_S is becoming very small, the bandwidth is not increasing anymore. This is because the transfer function of an RC-dominated interconnect has only a limited region, where the transfer function can be approximated with a first-order model (see figure 3.4).

The voltage swing on the interconnect also depends on C_S. The lower C_S, the lower the voltage swing. This lower voltage swing decreases the power consumption.

3.3.4 Eye-diagram properties

Figures 3.9, 3.10 and 3.11 give the absolute eye-height, relative eye-width and the relative latency for the example interconnect from section 1.5.2 as a function of data rate for different values of C_S. The figures show the increased bandwidth by using a capacitive source impedance. Both the eye-height and eye-width start to drop at higher data rates. Also, the relative latency decreases. The absolute eye-height for low data rates depends on the capacitive divider ratio $\dfrac{C_S}{C_S + C_T}$: for lower values of capacitance C_S, the voltage swing on the interconnect is smaller.

3.3.5 Power consumption.

The low voltage swing on the interconnect reduces the power consumption. Figure 3.12 shows the energy per bit of the example interconnect from section 1.5.2 as a function of data rate with a data activity of 0.5. The figure shows that the lower C_S is chosen, the lower the power consumption becomes.
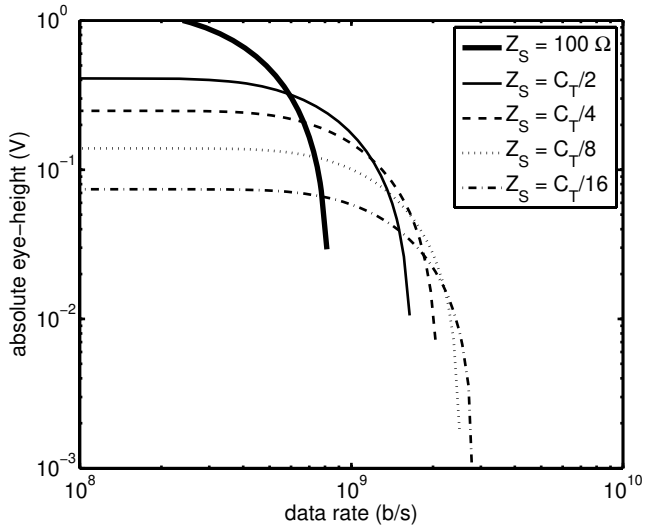
Figure 3.9: Absolute eye-height of the example interconnect from section 1.5.2 as a function of data rate for different values of Z_S.
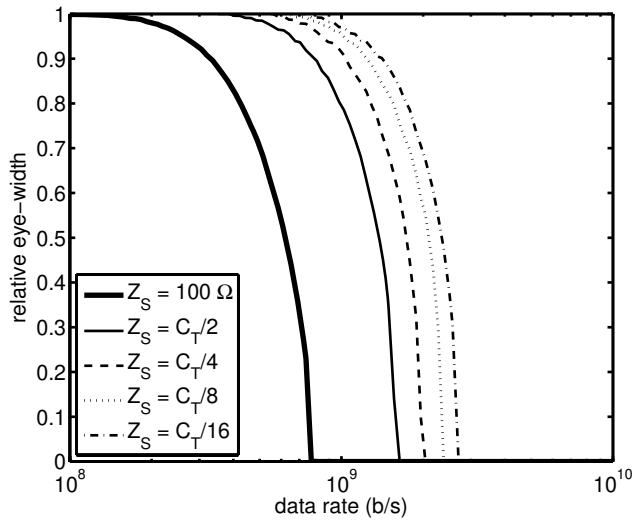
Figure 3.10: Relative eye-width of the example interconnect from section 1.5.2 as a function of data rate for different values of Z_S. The eye-width is relative to the symbol time.
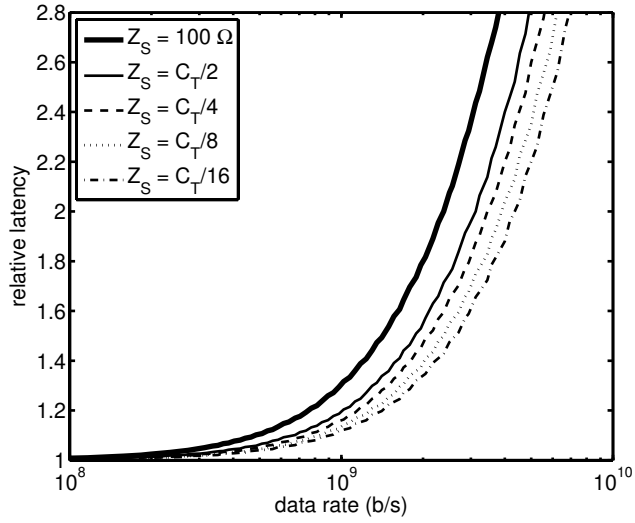
Figure 3.11: Relative latency of the example interconnect from section 1.5.2 as a function of data rate for different values of Z_S. The latency is relative to the symbol time.
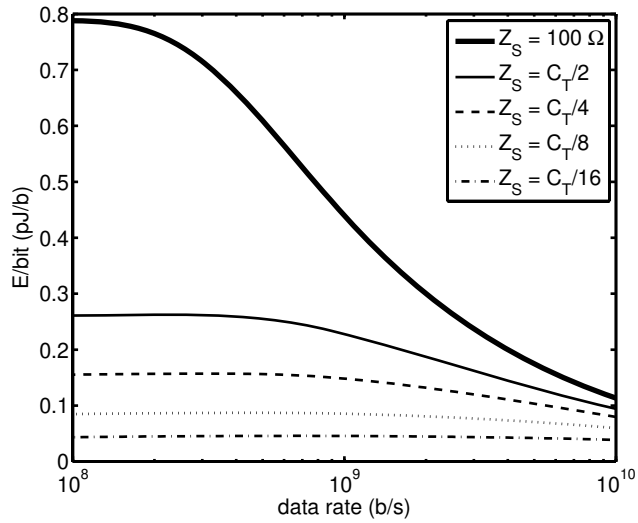
Figure 3.12: Power consumption of the example interconnect from section 1.5.2 as a function of data rate for different values of Z_S.

3.3.6 Biasing circuit

The operation of the circuit of figure 3.7 can be described as follows. When a data transition occurs, current is flowing through C_S, charging the line capacitance C_T to a certain voltage $V_{DC} + \frac{1}{2} \cdot V_{swing}$. Only if a next transition occurs, the line is discharged again to $V_{DC} - \frac{1}{2} \cdot V_{swing}$. V_{swing} is determined by $C_S/(C_S+C_T)$. V_{DC} should be a constant, but in the circuit of figure 3.7 this voltage is ill-defined as there is no DC path to one of the supplies. If there is some parasitic leakage path, charge can be removed from the interconnect and the voltage V_{DC} will change, resulting in false decisions by the receiver. In order to solve this problem, some means of controlling the voltage V_{DC} on the interconnect is needed. Different possibilities to define this voltage exist, e.g. via a switched capacitors network, but few solutions take very small chip area and power dissipation. A very simple effective way to solve the problem is to use a load resistor R_L connected to a bias voltage V_{bias} and a current source with transconductance G_M (controlled by V_S) as shown in figure 3.13.
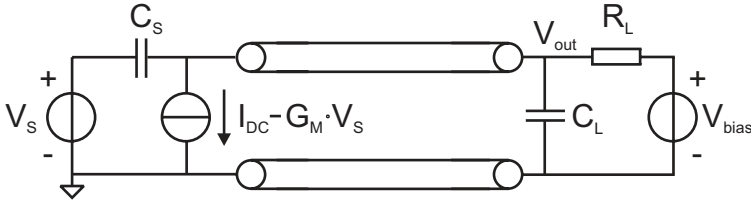
Figure 3.13: DC biasing of the capacitive transmitter circuit.

A constant current I_{DC} can be used to set V_{DC} to a suitable value.

If we assume that the interconnect can be approximated by a first-order RC model with a resistance of $a \cdot R_T$ and a capacitance of $a \cdot C_T$ (see section 2.6), the transfer function of figure 3.13 from V_S to V_{out} is

$$\frac{V_{out}}{V_S} = \frac{j \cdot \omega \cdot C_S \cdot R_L + G_M \cdot R_L}{1 + j \cdot \omega \cdot ((a \cdot R_T + R_L) \cdot C_S + a \cdot R_L \cdot C_T) + (j \cdot \omega)^2 \cdot a^2 \cdot R_T \cdot R_L \cdot C_S \cdot C_T} \quad (3.11)$$

This transfer function had two poles and a zero. In order to get a first-order RC response, the C_S has to be chosen as

$$C_S = \frac{a \cdot G_M \cdot R_L \cdot C_T \cdot (1 + a \cdot G_M \cdot R_T)}{1 - G_M \cdot (a \cdot R_T + R_L)} \quad (3.12)$$

For small G_M, this equation can be reduced to

$$\frac{C_S}{G_M} = a \cdot R_L \cdot C_T \quad (3.13)$$

The two time constants, C_S/G_M and $a \cdot R_L \cdot C_T$, have to be matched. Then, the bandwidth and DC gain are

$$BW = \frac{1}{a^2 \cdot R_T \cdot C_T} \cdot \frac{1}{G_M \cdot R_L} \quad [rad/s] \tag{3.14}$$

$$A = G_M \cdot R_L \tag{3.15}$$

The resistor R_L is chosen as high as possible to minimize power consumption, but low enough to absorb the worst case leakage current without significant voltage change at V_{out}.

By having the time constants C_S/G_M and $a \cdot R_L \cdot C_T$ equal, the transfer function of the circuit in figure 3.13 resembles the transfer function of the circuit in figure 3.7. But in contrast to the circuit in figure 3.7, the low frequency part is now determined by G_M and R_L and only the high frequency part by C_S and the interconnect. The circuit of figure 3.13 is named a capacitive pre-emphasis transmitter, since although the maximum voltage swing (low frequencies) at V_{out} is determined by G_M and R_L, every transition (high frequencies) is 'emphasized' by the transmitter by injecting a current via capacitance C_S. If a small G_M and a large R_L are chosen, the static current is kept small and the power consumption is also comparable to the circuit of figure 3.7.

3.4 Load impedance

3.4.1 Introduction

The previous section examined the best termination at the transmitter side. In this section, we will look at the receiver side. We again use the parameters of the example interconnect from section 1.5.2 and assume that the source impedance is 100 Ω.

3.4.2 Ideal load impedance

Similar to the previous section, we calculate the load impedance that gives an ideal transfer function, as given in (3.8), with magnitude A for all frequencies and a linear phase characteristic. Figure 3.15 shows this ideal load impedance. The ideal transfer function with A = 1 is only possible with a negative capacitance as load impedance. Note that the ideal source impedance was a negative resistance for A = 1.

For A = 0.1, the ideal load impedance can be approximated by a resistance. Thus, the schematic of figure 3.14 would give a bandwidth improvement.
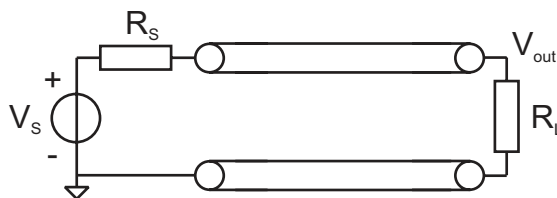
Figure 3.14: Resistive load.

The transfer function for different values of R_L is given in figure 3.16.
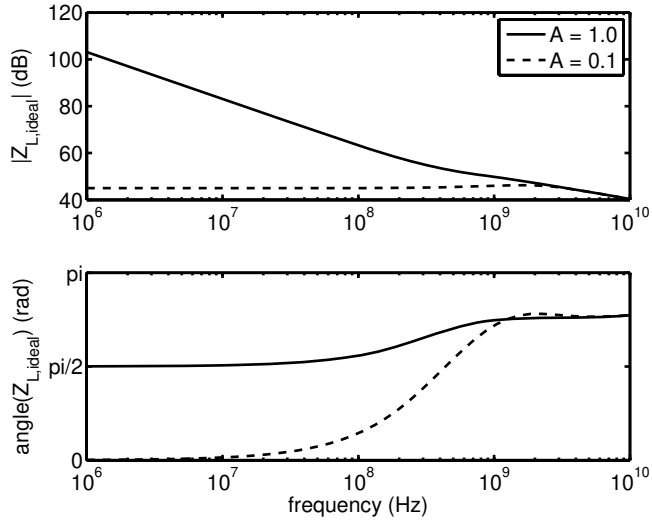
Figure 3.15: Ideal load impedance of the example interconnect from section 1.5.2.
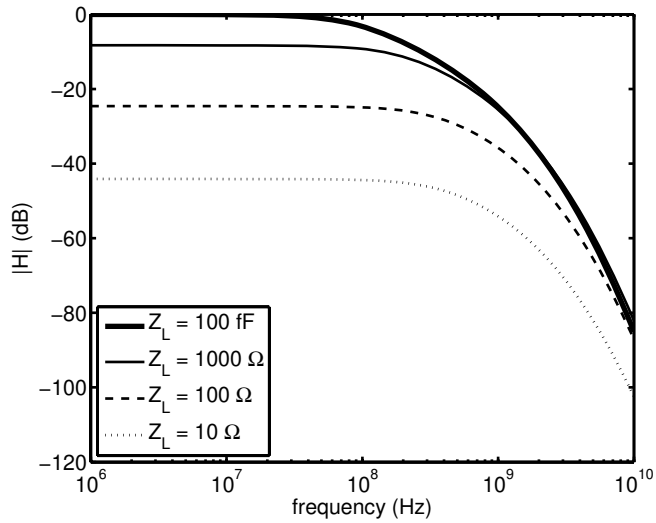
Figure 3.16: Transfer function of the example interconnect from section 1.5.2 for different load resistances.

As was the case for a capacitive source impedance, the resistive load impedance increases the bandwidth. Again, the bandwidth improvement gives a smaller swing at V_{out}.

3.4.3 First-order model for delay and bandwidth

In order to calculate a first-order model of the delay of an interconnect, dominated by RC behavior, the method of [12] can be used. In this paper, a ramp function is used as input signal, as shown in figure 3.17.
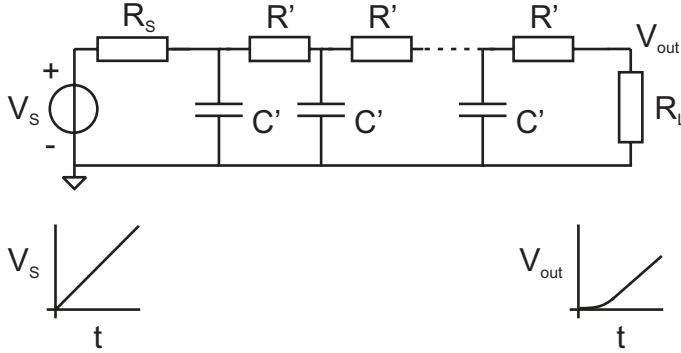
Figure 3.17: An RC-dominated interconnect with a ramp function as input signal.

After a certain delay, the output will also be a ramp function, but with a less steep slope. An expression for the output voltage can rather easily be obtained by using the fact that a ramp voltage across a resistor gives also a ramp current through this resistor and a ramp voltage across a capacitor gives a constant current through the capacitor.

The result for $V_S = p_v \cdot t$ is

$$V_{out} = \frac{R_L}{R_S + R_T + R_L} \cdot p_v \cdot (t - \frac{R_T \cdot C_T}{2} \cdot \frac{R_S + \dfrac{R_T}{3} + R_L \cdot (1 + \dfrac{2 \cdot R_S}{R_T})}{R_S + R_T + R_L})$$ (3.16)

where $R_T = l \cdot R'$ and $C_T = l \cdot C'$. This result shows the DC gain A from V_S to V_{out} to be

$$A = \frac{R_L}{R_S + R_T + R_L}$$ (3.17)

and the delay d

$$d = \frac{R_T \cdot C_T}{2} \cdot \frac{R_S + \dfrac{R_T}{3} + R_L \cdot (1 + \dfrac{2 \cdot R_S}{R_T})}{R_S + R_T + R_L}$$ (3.18)

The delay depends not only on the total resistance and capacitance of the interconnect, but also on the termination resistances R_S and R_L.

The bandwidth of the interconnect can be approximated with the inverse of the delay. Note that for $R_S = 0$ and $R_L = \infty$, the bandwidth is given by

$$BW = \frac{1}{0.5 \cdot R_T \cdot C_T} \quad [rad / s] \tag{3.19}$$

In section 2.6, we have already seen that an interconnect that is dominated by RC behavior can be approximated with a first-order model. There, a factor $a = 0.64$ was used to match the pole to the first pole of the distributed model. With the simple model, we find a slightly higher factor $a = \frac{1}{2} \cdot \sqrt{2}$.

By not choosing $R_L = \infty$, but giving R_L a lower value, the bandwidth increases. The maximum bandwidth is obtained if $R_L = 0$.

$$BW\big|_{R_L=0} = \frac{1}{0.5 \cdot R_T \cdot C_T} \cdot \frac{R_S + R_T}{R_S + \dfrac{R_T}{3}} \tag{3.20}$$

The bandwidth gain, compared to $R_L = \infty$, is equal to

$$BWgain = \frac{R_S + R_T}{R_S + \dfrac{R_T}{3}} \tag{3.21}$$

Thus, for a small source impedance, the maximum bandwidth gain is three. Of course, for $R_L = 0$, the voltage swing is also zero. However, one can look at the current and convert this into a voltage with a transimpedance amplifier (see section 4.2). Terminating an interconnect with a low-ohmic termination is also referred to as current-sensing.

3.4.4 Eye-diagram properties

The transfer functions of figure 3.16 result in the eye-diagram properties of figures 3.18, 3.19 and 3.20. The lower the load resistance, the larger the relative eye-opening and the smaller the latency. Thus, just as with a capacitive transmitter, a higher data rate is achievable with resistive termination.

The bandwidth improving effect of the resistive termination can be fully explained with the s-parameter model of chapter 2. However, as already mentioned in section 2.4.3, we can also describe the interconnect with a diffusion process. The charge that is put on the distributed capacitance at the beginning of the interconnect diffuses to the other end of the interconnect. The speed of the diffusion process depends on the derivative of the charge with respect to distance. The bandwidth-increasing effect of a low-ohmic load resistance can now be explained in the following way. With a high-ohmic load, charge can only be put on and removed from the interconnect at the transmitter side. All charge has to diffuse back to the low-ohmic transmitter and it takes a long time before the interconnect is fully discharged. With also a low-ohmic point at the receiver side, the charge has two drains and the interconnect is discharged much faster.
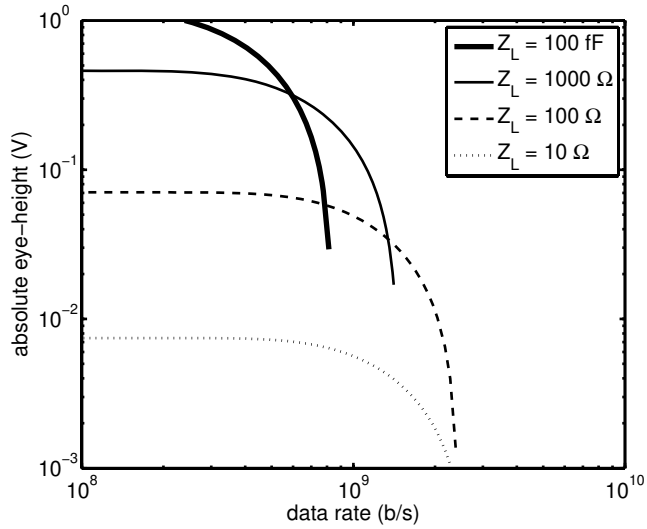
Figure 3.18: Absolute eye-height of the example interconnect from section 1.5.2 as a function of data rate for different values of Z_L.
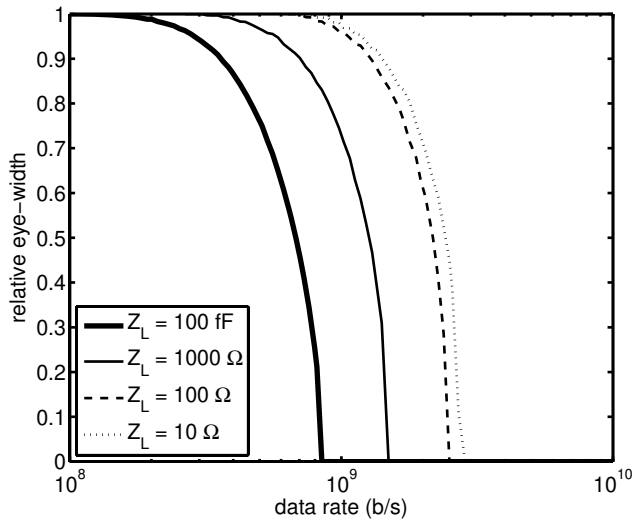
Figure 3.19: Relative eye-width of the example interconnect from section 1.5.2 as a function of data rate for different values of Z_L. The eye-width is relative to the symbol time.
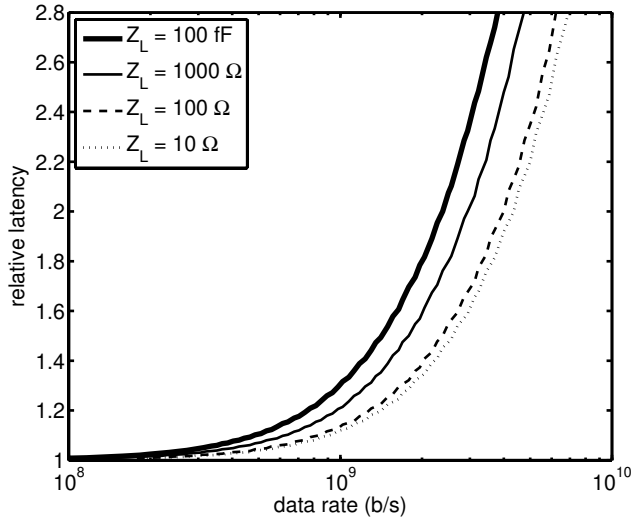
Figure 3.20: Relative latency of the example interconnect from section 1.5.2 as a function of data rate for different values of Z_L. The latency is relative to the symbol time.

3.4.5 Power consumption

The power consumption for different load impedances is shown in figure 3.21. The energy per bit is plotted for a transition probability of 0.5.
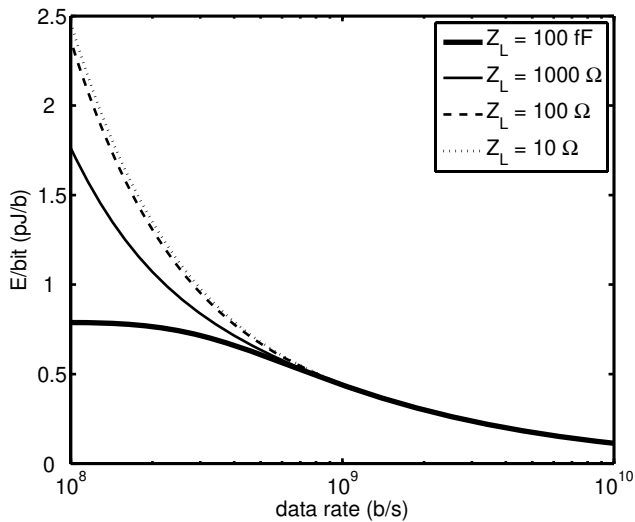
Figure 3.21: Power consumption of the example interconnect from section 1.5.2 as a function of data rate for different values of Z_L.

Note that the additional power to create the resistive termination (see section 2.4.7) is not included. Still, the E/bit is larger for resistive termination than for a conventional scheme with $Z_L = 100$ fF. The reason for the high power consumption (at low data rates) of the resistive termination scheme is high static power consumption, as we will see in the next section.

3.4.6 First-order model for power consumption

A first-order model for power consumption was already given in section 1.5.6. It is repeated here.

$$P_{dyn} = \frac{1}{2} \cdot p_{trans} \cdot C_T \cdot V_{swing} \cdot V_{DD} \cdot f_{clock} \qquad (3.22)$$

P_{dyn} is the dynamic power consumption, p_{trans} the transition probability, C_T the total distributed capacitance, V_{swing} the voltage swing on the interconnect, V_{DD} the supply voltage and f_{clock} the clock frequency.

Unfortunately, this equation is only useful if the entire interconnect is charged up to a voltage V_{swing}. For $Z_L = \infty$ and low data rates this is true. However, if Z_L is a small resistor, the voltage swing will not be constant over the length of the interconnect. Therefore, a better power model would use

$$V_{swing}(z) = \frac{(1-z) \cdot R_T + R_L}{R_S + R_T + R_L} \cdot V_{DD} \qquad (3.23)$$

with z the position along the interconnect: $z = 0$ is at the beginning of the interconnect and $z = 1$ at the end. The dynamic power consumption is found by integrating the square of $V_{swing}(z)$ over the entire length of the interconnect.

$$\int_{z=0}^{1} (V_{swing}(z))^2 \cdot dz = \frac{\frac{1}{3} \cdot R_T{}^2 + R_T \cdot R_L + R_L{}^2}{(R_S + R_T + R_L)^2} \cdot V_{DD}{}^2 \qquad (3.24)$$

$$P_{dyn} = \frac{1}{2} \cdot p_{trans} \cdot C_T \cdot \frac{\frac{1}{3} \cdot R_T{}^2 + R_T \cdot R_L + R_L{}^2}{(R_S + R_T + R_L)^2} \cdot V_{DD}{}^2 \cdot f_{clock} \qquad (3.25)$$

For static power consumption, the following equation can be used

$$P_{static} = \frac{V_{DD}{}^2}{4 \cdot (R_S + R_T + R_L)} \qquad (3.26)$$

We can compare this model with the distributed model of section 2.4.6. In figure 3.22, this is done for $V_{DD} = 1.2$ V, $R_T = 1.5$ kΩ, $C_T = 2.2$ pF, $Z_S = 100$ Ω and two different values for Z_L and the data rate.
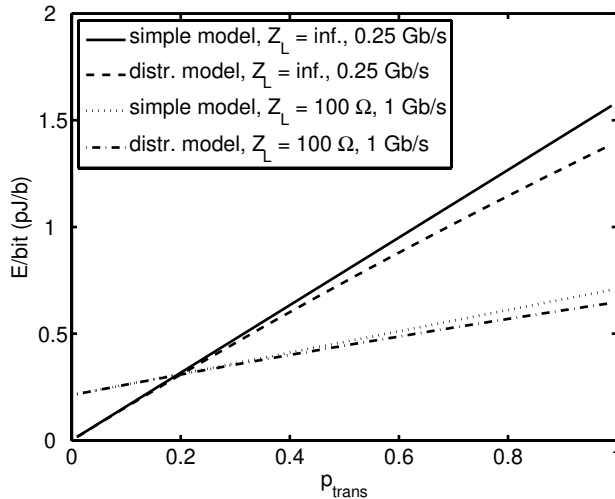
Figure 3.22: Comparison simple model with distributed model for power consumption.

The figure shows that the simple model of this section only slightly overestimates the power consumption for high transition probabilities. The limited bandwidth reduces the swing on the interconnect and thus the power consumption and this effect is only reflected in the distributed power consumption model.

With the help of the first-order model of this section, we are able to explain the higher power consumption of the resistive termination scheme. The dynamic power consumption is given in (3.25). This equation states that for $R_S = R_L = 0$, the dynamic power consumption is three times smaller than for $R_S = 0$ and $R_L = \infty$ (i.e. a small capacitor). So, a resistive termination has smaller dynamic power than a conventional termination scheme, with a maximum power consumption gain of three. However, the increase in static power, given in (3.26), is equal to the decrease in dynamic power for the data activity as used in figure 3.21. For low data rates the increase in static power consumption is even larger than the decrease in dynamic power consumption. So, for resistive termination the bandwidth gain comes together with large static power consumption.

3.4.7 Inductive termination

The low resistance of the load attenuates the low-frequency transfer more than the high frequency transfer and thus acts as an equalizer. We can even further equalize by placing an inductor in series with the resistance. In this way, the attenuation at high frequencies will be lower. Note that inductive peaking is also used in wideband amplifier designs [50]. The schematic looks like:
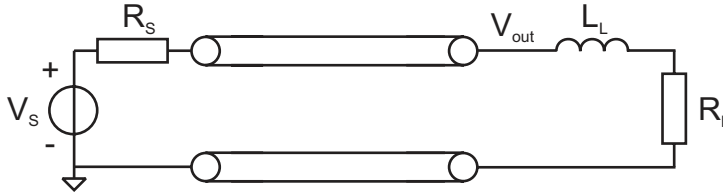
Figure 3.23: Resistive and inductive load.

For the example interconnect from section 1.5.2, the optimal value for the inductance is about 37 nH. The transfer function of figure 3.24 shows the large improvement in bandwidth.
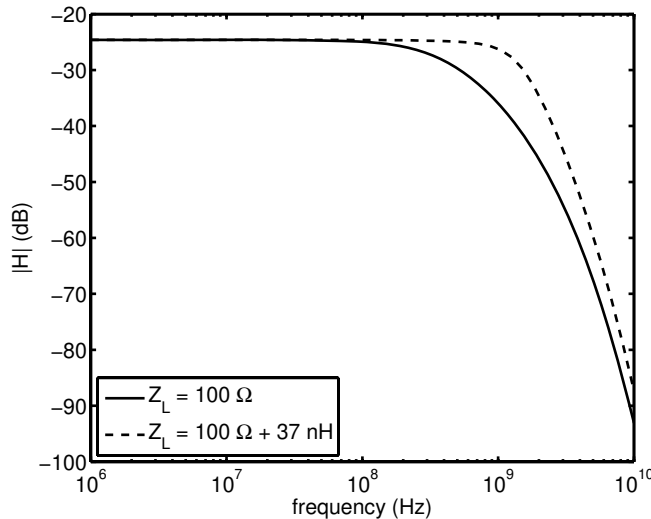
Figure 3.24: Transfer function of the example interconnect from section 1.5.2 for a load impedance that has a resistor and inductor in series.

Unfortunately, implementing an inductance of 37 nH with on-chip metal layers would require impractically large area. However, by using the circuit implementation of [42], it is possible to create such a scheme on-chip. This circuit uses a transconductance to implement the low-ohmic load resistance. The output voltage of the interconnect is low-pass filtered and the output of this filter is used to turn off the transconductance at high frequencies. In this way a load impedance is created that is low-ohmic for low frequencies and high-ohmic for high frequencies.

3.5 Comparison

In section 3.3, we concluded that a capacitive source impedance not only increases the bandwidth, but also decreases the power consumption. The resistive termination of section

3.4 also increases the bandwidth. Unfortunately, the static power consumption is much higher for this scheme. But also the dynamic power consumption is larger.

In order to understand the difference between the schemes in terms of dynamic power consumption, figure 3.25 plots the step response against time for different positions along the interconnect. The beginning of the interconnect is at $z = 0$ and the end at $z = 1$. The parameters of the example interconnect from section 1.5.2 are used again. For the capacitive source impedance case $Z_S = 275$ fF and $Z_L = 100$ fF, while for the resistive load impedance case $Z_S = 100$ Ω and $Z_L = 200$ Ω. The components have been chosen such that both schemes have the same final value for the output voltage at $z = 1$.
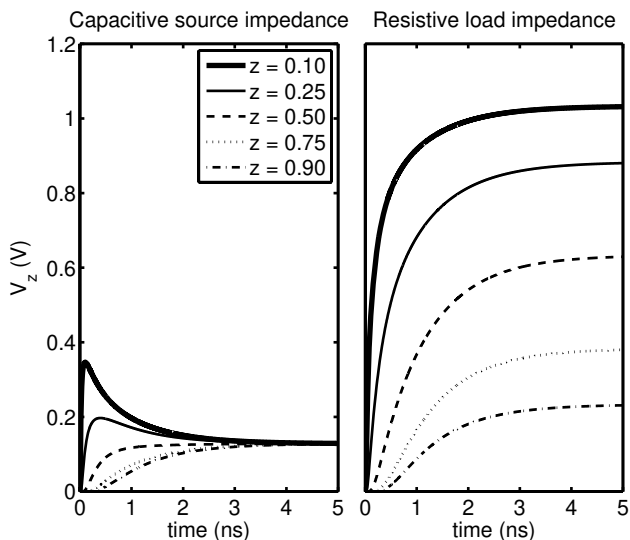
Figure 3.25: Step response of the example interconnect from section 1.5.2 as a function of time for different positions along the interconnect.

The figure shows that with a capacitive source impedance the entire interconnect is charged to the same low voltage. The overshoot at the beginning of the interconnect gives the bandwidth-increment. Because the entire interconnect is only charged to a small voltage, the power consumption is low. This is not true for a resistive load impedance. For small z the interconnect is charged to almost V_{DD} and only for a z close to one, the interconnect has a small swing. Due to this larger swing at the beginning of the interconnect, the scheme with a resistive load impedance has much larger dynamic power consumption than the scheme with a capacitive source impedance.

3.6 Parameter spread

3.6.1 Introduction

We now further examine the termination concepts as given in the previous sections. First we look at the capacitive pre-emphasis transmitter, as given in figure 3.13. We look at how

to choose the different parameters and how parameter spread influences the eye-height. After that, we do the same for the resistive scheme. For the interconnect, we use the parameters of the example interconnect from section 1.5.2.

3.6.2 Capacitive pre-emphasis transmitter

The schematic of figure 3.13 has a few parameters that have to be chosen: C_S, G_M, R_L, I_{DC} and V_{bias}. The constant current I_{DC} and bias voltage V_{bias} are only used to set an appropriate bias voltage. We will come back to that in chapter 5, where circuit implementations of the concept are discussed. The choice of the other three parameters is treated in this section.

The bandwidth and DC gain for small values of G_M are given by

$$ BW = \frac{1}{a^2 \cdot R_T \cdot C_T} \cdot \frac{1}{G_M \cdot R_L} \quad [rad/s] \tag{3.27} $$

$$ A = G_M \cdot R_L \tag{3.28} $$

The product of G_M and R_L should be chosen in such a way, that the DC gain is large enough for the receiver. If G_M is chosen small and R_L large, then (3.27) holds and the bandwidth is fixed. However, we can also choose a larger G_M with a small R_L, as we know from section 3.4 that a small R_L will give a higher bandwidth. Unfortunately, this will also give larger static power consumption. So, a trade-off should be made.

In section 3.3.6, it was mentioned that there are two time constants in figure 3.13 that have to be matched: C_S/G_M and $a \cdot C_T \cdot R_L$.

$$ \frac{C_S}{G_M} = a \cdot R_L \cdot C_T \tag{3.29} $$

C_S is chosen in order to match the time constants. Figure 3.26 shows what happens to the transfer function if the time constants are not matched. G_M and R_L are chosen as 11 μS and 10 kΩ respectively and C_S is given values of 200 fF, 300 fF and 450 fF. The figure shows that a too high or a too low capacitance value gives a step in the transfer function. This causes a closure of the eye.

In order to examine the influence of all design parameters on the eye properties, figure 3.27 shows the eye-height as a function of parameter spread. This is done at a data rate of 1.25 Gb/s, where the eye-height is around 78 mV in the nominal case. The figure shows that both a smaller and a larger C_S decrease the eye-height. However, for a larger C_S the decrease is less. The influence of G_M and R_L is almost linear with the mismatch: 20% mismatch also gives 20% eye-closure. A smaller R_T or C_T increases, while a larger R_T or C_T decreases the eye-height. The reason for this is the change in bandwidth of the interconnect, which depends linearly on both R_T and C_T. However, the influence of C_T is somewhat different, because C_T is also part of the time constants that have to be matched, while R_T is not.
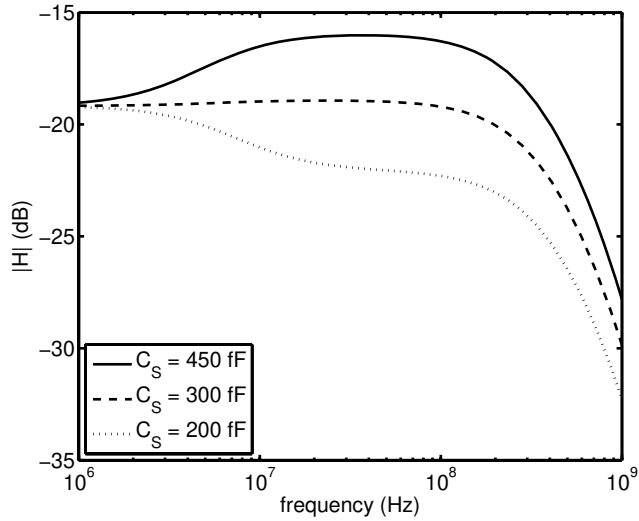
Figure 3.26: Transfer function with time constant mismatch due to C_S changes.
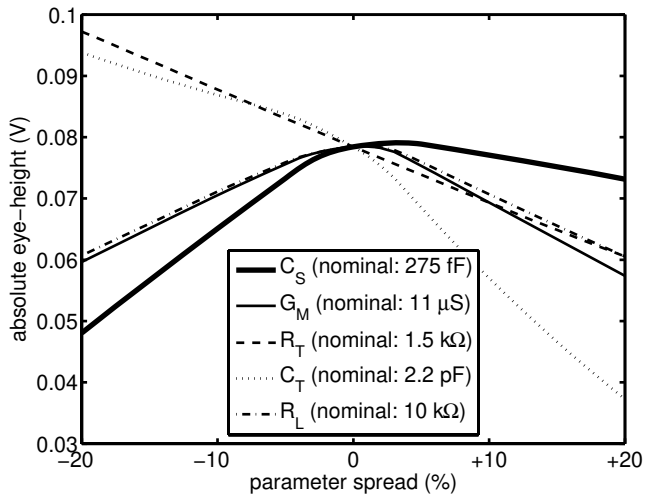
Figure 3.27: Influence of parameter spread on eye-height for the circuit of figure 3.13.

Next to parameter variation, another effect can limit the performance of the capacitive pre-emphasis transmitter. The voltage source V_S will have a source resistance R_S that is placed in series with C_S. The effect of this source resistance is shown in figure 3.28. As long as the source resistance is much smaller than the interconnect resistance R_T, the eye-height is not reduced.
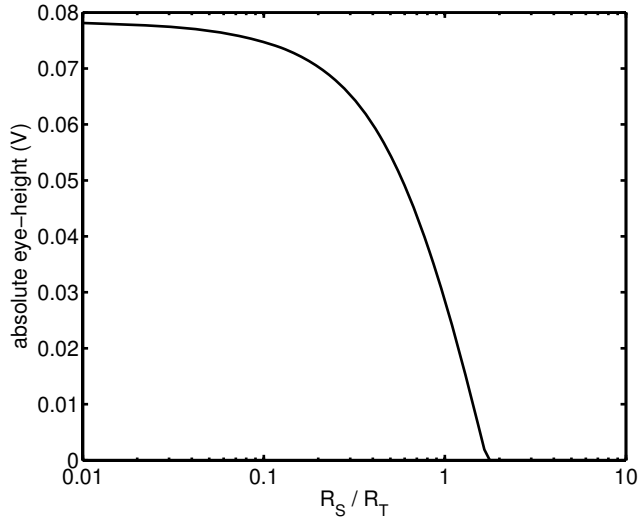
Figure 3.28: Influence of source resistance R_S on eye-height for the circuit of figure 3.13.

3.6.3 Resistive termination

For the resistive termination, as in figure 3.14, the bandwidth and DC gain are defined in section 3.4.3 and the power consumption in 3.4.5. For choosing the parameter values, a trade-off has to be made between bandwidth and static power consumption. A small load resistance gives a high bandwidth, but also high static power consumption.

The bandwidth for a conventional case with $R_S = 0$ and $R_L = \infty$ was already given in section 3.4.3, but is repeated here.

$$BW = \frac{1}{0.5 \cdot R_T \cdot C_T} \quad [rad / s] \tag{3.30}$$

The maximum bandwidth is achieved for $R_L = 0$ and is equal to

$$BW = \frac{3}{0.5 \cdot R_T \cdot C_T} \quad [rad / s] \tag{3.31}$$

Thus, the maximum bandwidth gain by using a low-ohmic resistive termination is three. This factor of three is achieved if both R_S and R_L are much smaller than R_T.

Again we will examine the influence of all design parameters on the eye properties. Figure 3.29 shows the eye-height as a function of parameter spread. This is done at a data rate of 1.25 Gb/s, where the eye-height is around 78 mV in the nominal case. No inductor is used.
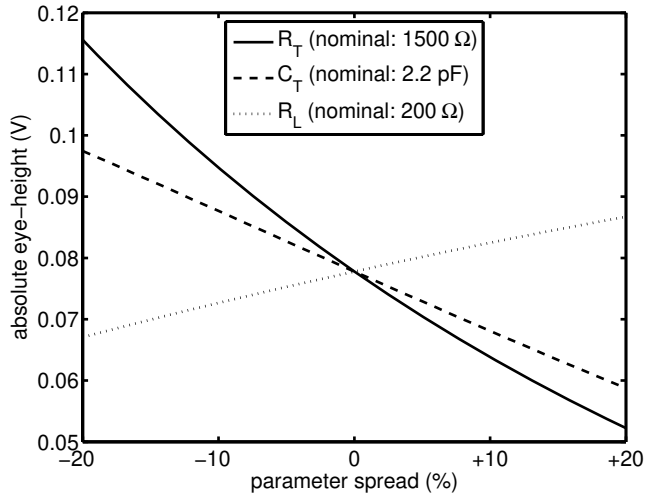
Figure 3.29: Influence of parameter spread on eye-height for the circuit of figure 3.14.

As was the case for the capacitive transmitter, the eye-height is most sensitive to R_T and C_T, because these parameters directly influence the bandwidth and thus the eye-height. Now, the influence of R_T is larger than of C_T, because the output swing depends on R_T and R_L. The influence of R_L is opposite to the influence of R_T.

The effect of the source resistance R_S is shown in figure 3.30.
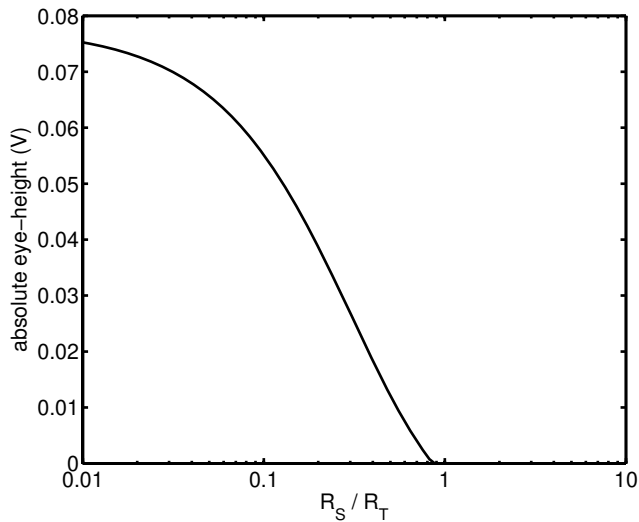
Figure 3.30: Influence of source resistance R_S on eye-height for the circuit of figure 3.14.

74

Again, the source resistance should be kept well below R_T. If R_S is larger, the bandwidth will be further limited and the eye-height drops.

3.7 Equalization techniques

3.7.1 Introduction

The previous sections describe termination concepts at both the transmitter and receiver side to increase the bandwidth of the interconnect. This section will briefly discuss two equalization techniques that can also be used to increase the achievable data rate. A more detailed discussion of both techniques can be found in [42]. First, an equalization scheme at the transmitter side is discussed and then an equalization scheme at the receiver side.

3.7.2 Pulse-width equalization

An equalization technique that can be used at the transmitter side is called pulse-width (PW) pre-emphasis equalization. Figure 3.31 explains the technique.
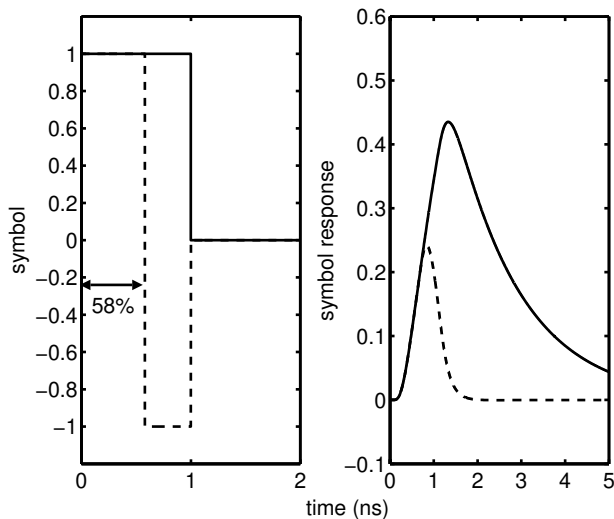
Figure 3.31: Conventional binary signaling (straight lines) compared to pulse-width pre-emphasis equalization (dashed lines) for the example interconnect of section 1.5.2.

The figure compares the technique with conventional binary signaling. With binary signaling, the symbol is one for the entire symbol time. In the example of the figure, the parameters of the example interconnect of section 1.5.2 are used and the symbol time is 1 ns. The symbol response has a characteristically long tail. With PW pre-emphasis equalization, the last part of the symbol is minus one. If the pulse-width, defined as the period that the symbol is plus one, is chosen well (58% in the example), the long tail is cancelled.

The optimal pulse-width depends on the time constant of the interconnect and on the data rate. Figure 3.32 plots the absolute eye-height at the output of the interconnect for different data rates. At each data rate, the optimal pulse-width is used, varying between 100% for low data rates and 52% at 3.5 Gb/s. Also added in the figure are the eye-heights of the capacitive transmitter of section 3.3.
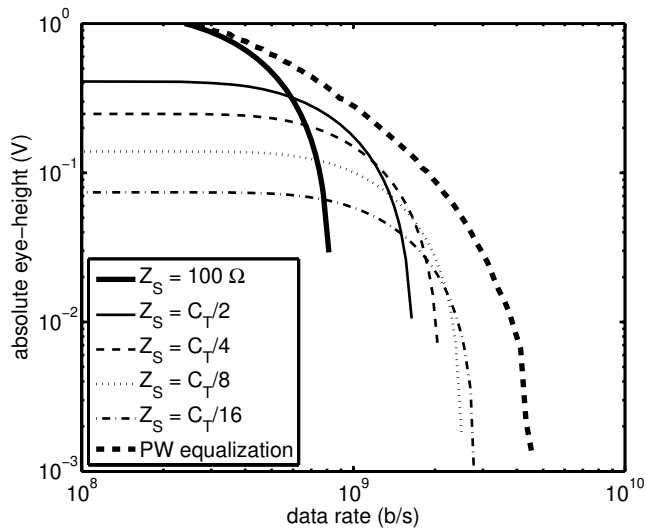
Figure 3.32: Absolute eye-height of the example interconnect of section 1.5.2 for conventional termination ($Z_S = 100 \ \Omega$, $Z_L = 100$ fF), capacitive transmitters and PW equalization.

We see from the figure, that with pulse-width pre-emphasis even higher data rates are possible than by using the capacitive transmitter. However, if we look at power consumption, the pulse-width pre-emphasis equalization is less beneficial, as shown in figure 3.33. Because with pulse-width equalization, the transmitter is switching every symbol period, the power consumption is even larger than in the conventional case. Furthermore, the power consumption will not drop for lower transition probabilities.

Concluding, with pulse-width pre-emphasis higher data rates are possible, but at the cost of static power consumption (see [42]). The technique can also be used simultaneously with other techniques as discussed in this thesis. A transceiver that both uses pulse-width equalization and a low-ohmic load resistance will be shown in section 5.3
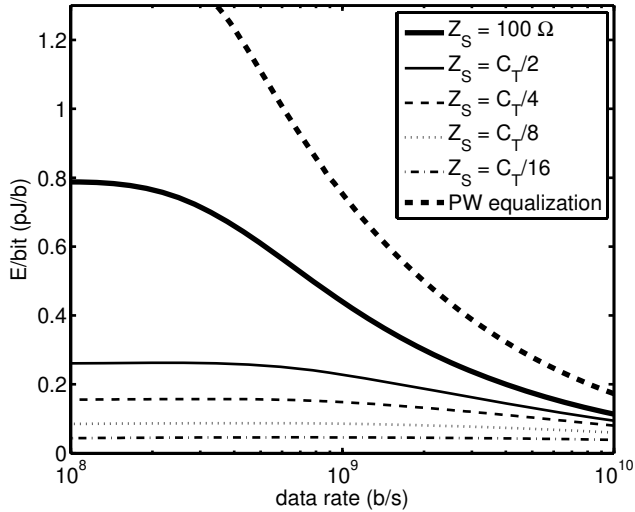
Figure 3.33: Power consumption of the example interconnect of section 1.5.2 for conventional termination ($Z_S = 100\ \Omega$, $Z_L = 100$ fF), capacitive transmitters and PW equalization ($p_{trans} = 0.5$)

3.7.3 Decision feedback equalization

Now we look at an equalization scheme at the receiver side. For that we use decision feedback equalization (DFE), as also used for off-chip communication [51]. The concept is shown in figure 3.34.
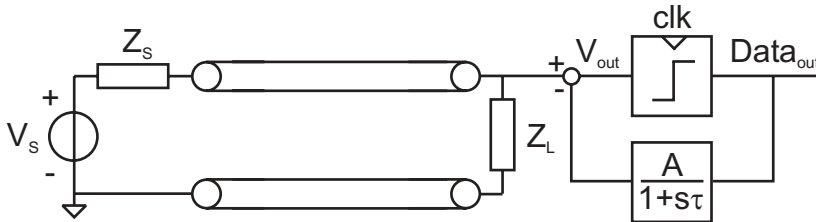
Figure 3.34: Concept of decision feedback equalization (DFE).

A clocked comparator is used to restore the interconnect output to full swing (Data$_{out}$). This signal is fed back to the interconnect output through a low-pass filter with time constant τ. If this time constant matches with the time constant of the interconnect, a large part of the long tail of the symbol response can be compensated. Thus, higher data rates can be achieved, as shown in figure 3.35. In this figure, Z_L is 100 fF for the DFE curve. For comparison, also the eye-heights of the low-ohmic terminations from section 3.4 are shown. Note, that the power that is consumed in the interconnect is equal for the conventional and the DFE case. The comparator of the DFE scheme is also needed to restore the low-swing output signals of the capacitively terminated (at the source) and resistively terminated (at
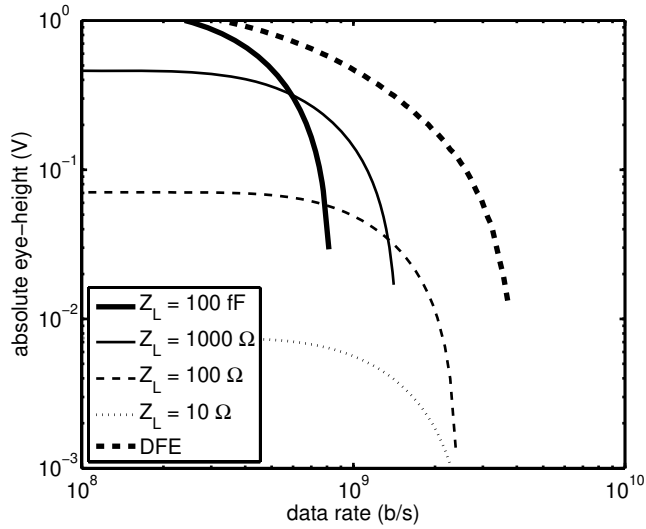
Figure 3.35: Absolute eye-height of conventional, resistive termination and DFE equalization.

the load) interconnects. The power of the additional low-pass filter can be made very low, as we will see in chapter 5.

In conclusion, the DFE equalization achieves a high data rate without adding much power consumption (see [42]). The technique can also be used simultaneously with other techniques that are discussed in this thesis. In section 5.5, we will discuss a transceiver that uses both a capacitive pre-emphasis transmitter and decision feedback equalization.

3.8 Summary

− Global interconnects are dominated by their resistance and capacitance. Only short (< 1 mm) or very thick and wide interconnects with a lot of spacing towards other metal are dominated by their inductance and capacitance.
− For highest bandwidth per area, all interconnect dimensions should be chosen equal.
− The ideal source impedance for an interconnect that has a small capacitive load impedance, resembles a series capacitance. By using a series capacitance as the source impedance, both the bandwidth increases and the power consumption decreases.
− With a small transconductance and large load resistance, the DC potential of a capacitively coupled transmitter can be defined, without increasing the static power consumption much.
− The ideal load impedance for an interconnect that has a small resistive source impedance, resembles a low-ohmic resistance. By having a load resistance much lower than the resistance of the interconnect, the bandwidth of the interconnect increases. Also, the dynamic power consumption decreases, but this effect is canceled by an increase in static power consumption.
− Other useful equalization techniques include pulse-width pre-emphasis equalization at the transmitter and decision feedback equalization at the receiver, possibly in combination with one or both termination techniques.

Chapter 4

Data integrity

4.1 Introduction

The previous chapter described the optimization of the dimensions of the interconnect to achieve minimum area consumption. Furthermore, termination concepts and equalization techniques were introduced to increase the achievable data rate. However, the higher achievable data rate comes at the cost of a low voltage swing at the receiver side, giving data integrity issues. These issues are discussed in this chapter.

First, we look at offset. For the resistive termination scheme, discussed in the previous chapter, a very low voltage swing is present at the receiver end of the interconnect. To have also low offset, in general, high power consumption would be needed in the receiver circuits. In order to have less power consumption, a receiver that senses the current instead of the voltage is discussed in the next section. After that, in section 4.3, crosstalk from neighboring interconnects is discussed. These are aggressor interconnects in the same metal layer that run in parallel to a victim interconnect. A solution for this crosstalk with twists is explained. Crosstalk from other interconnects in other metal layers is discussed in section 4.4.

4.2 Offset

In general, the termination concepts and equalization techniques, as described in the previous chapter, have a low voltage swing at the receiver side. In order to restore the

signals to full-swing we use a comparator (or sense amplifier based flip-flop), which achieves fast decisions due to a strong positive feedback. The comparator makes the decision for a "one" or a "zero" at one of the clock edges. In an application, a simple skew circuit or a source-synchronous approach could be used to generate the proper clock phase.

Unfortunately, the comparator will suffer from offset, as depicted in figure 4.1.
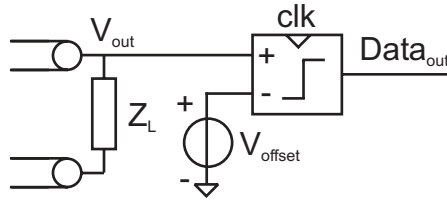
Figure 4.1: The receiver circuits have offset, represented with a single offset voltage source.

The offset voltage results from mismatches in the receiver circuits. We represent all offsets by a single offset voltage source at one of the inputs of the comparator. The eye-opening at the output of the interconnect should be large enough to tolerate this offset. Thus, if a very low voltage swing is used on the interconnect, the offset should also be very low. However, the power consumption of the comparator depends heavily on the amount of offset that is to be expected. If no offset compensation techniques are used, high power consumption is needed to have a low offset.

If we use the resistive termination concept of the previous chapter, the voltage swing at V_{out} will be low. Thus, the offset voltage should also be low, giving high power consumption in the comparator. We can alleviate the offset problem by using a transimpedance amplifier instead of a real termination resistor [48, 52], as shown in figure 4.2.
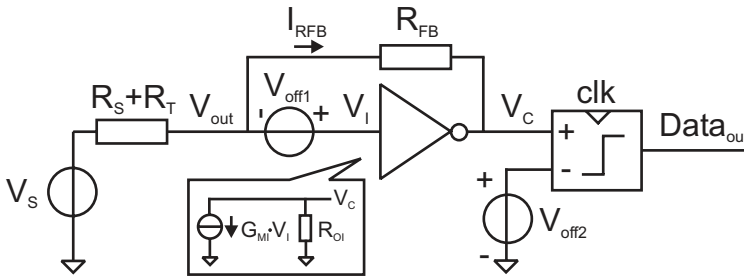
Figure 4.2: Offset voltage is added to the output voltage of the interconnect.

The load impedance Z_L is replaced with an inverter with transconductance G_{MI} and output resistance R_{OI} and with a feedback resistor R_{FB}. The resulting load resistance can be written as

$$R_{L,eff} = \frac{V_{out}}{I_{RFB}} = \frac{R_{FB} + R_{OI}}{1 + G_{MI} \cdot R_{OI}} \tag{4.1}$$

For $R_{OI} \rightarrow \infty$, $R_{L,eff} = 1/G_{MI}$.

Without offsets, the voltage V_{out} and V_C can be written as

$$\frac{V_{out}}{V_S} = \frac{R_{L,eff}}{R_S + R_T + R_{L,eff}} \tag{4.2}$$

$$\frac{V_C}{V_S} = \frac{R_{L,eff} - R_{FB}}{R_S + R_T + R_{L,eff}} \tag{4.3}$$

Thus, if R_{FB} is chosen much larger than $R_{L,eff}$, the signal at V_C is larger than at V_{out}. Due to this amplification from V_{out} to V_C, the influence of the offset of the comparator (V_{off2}) has become smaller. Therefore, the power in the comparator can be much lower. Of course, the transimpedance amplifier also consumes power, because in order to have a high G_{MI}, a high bias current is needed. Furthermore, the transimpedance amplifier has offset. The influence of this offset will be calculated next.

The transfer function of the offset voltage of the transimpedance amplifier (V_{off1}) to V_C is

$$\frac{V_C}{V_{off1}} = \frac{G_{MI} \cdot R_{OI} \cdot (R_S + R_T + R_{FB})}{R_S + R_T + R_{FB} + R_{OI} + G_{MI} \cdot R_{OI} \cdot (R_S + R_T)}$$

$$\xrightarrow{R_{out} \rightarrow \infty} \frac{R_S + R_T + R_{FB}}{R_S + R_T + \dfrac{1}{G_{MI}}} = \frac{R_S + R_T + R_{FB}}{R_S + R_T + R_{L,eff}} \tag{4.4}$$

Thus, if R_{FB} is much larger than R_T this offset voltage has approximately the same transfer to V_C as the interconnect input voltage V_S (see (4.3)). As for plain binary signaling the swing of V_S is V_{DD}, V_{off1} can be large. Note that the voltage swing at V_{out} is not important for the offset in this case.

4.3 Neighbor-to-neighbor crosstalk

4.3.1 Introduction
Next to offset issues, crosstalk also limits the performance of the interconnect. This section focuses on crosstalk between neighboring interconnects. A neighboring interconnect acts as an aggressor interconnect when it runs in parallel to a victim interconnect and is placed in the same metal layer. We only look at capacitive crosstalk, as this is the most dominant crosstalk source. However, the analysis of this section can easily be expanded to other forms of crosstalk (e.g. inductive). The distributed capacitance between two neighboring interconnects is called C_M', as shown in figure 4.3.

Section 1.5.4 already describes how crosstalk can be included in calculating the eye-height, while it was shown in section 2.5 how to calculate the crosstalk transfer function.
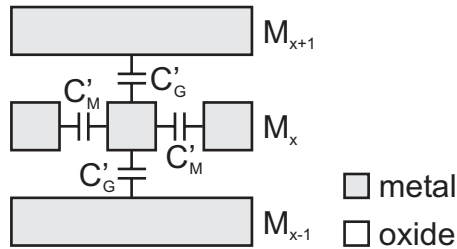
Figure 4.3: Capacitances to ground (C_G') and neighboring interconnects (C_M').

Figure 4.4 shows the eye-height of the example interconnect from section 1.5.2 with and without crosstalk from a neighboring interconnect.
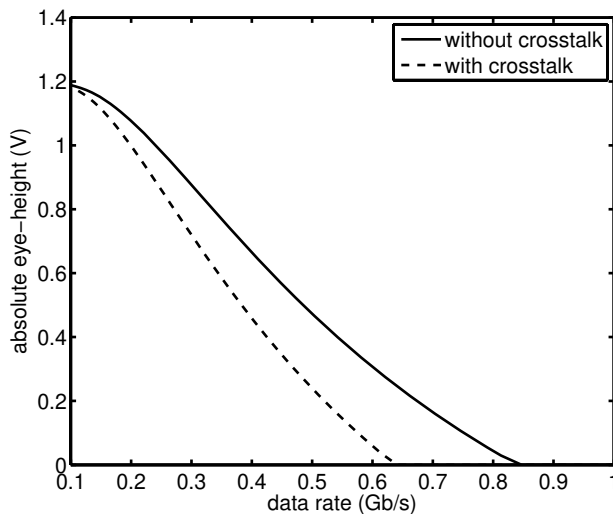
Figure 4.4: Absolute eye-height with and without crosstalk on a neighboring interconnect.

We see that the crosstalk, which comes from one neighbor, decreases the eye-height. If the other neighbor also injects crosstalk, the eye-closure will even be worse. Thus, because of crosstalk, lower data rates are achievable.

A number of solutions exist in literature, as described in section 1.6.5. The solution that is discussed in this section uses differential interconnects with twists. As we already want to use differential interconnects to be robust against other noise sources (see end of section 1.5.4), it is easy to make a twist in these interconnects. However, the vias that are needed to make a twist add resistance. Furthermore, the use of many twists makes the layout more difficult. Therefore, we use a minimum number of twists (only one or two). A single twist in a differential interconnect cancels differential mode crosstalk from its neighboring interconnects. By giving these neighboring interconnects a double twist, also common-mode crosstalk is canceled.

4.3.2 Crosstalk and twists

Figure 4.5 shows eight interconnects (line 1 to 8) forming four differential interconnects. Each differential interconnect has either one or two twists.
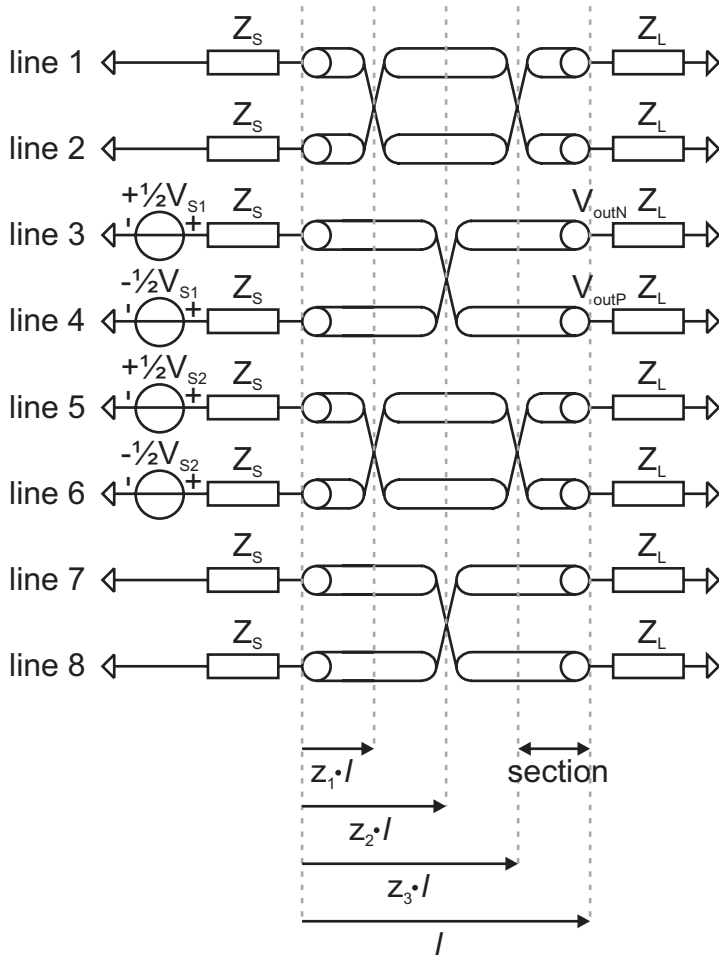
Figure 4.5: Four differential interconnects with one and two twists.

In the figure, lines 3 and 4 are driven differentially with a voltage V_{S1} and lines 5 and 6 are driven differentially with a voltage V_{S2}. In this section, we will calculate the following transfer functions:

$$H_{DM} = \left. \frac{V_{outP} - V_{outN}}{V_{S1}} \right|_{V_{S2}=0} \tag{4.5}$$

$$H_{XDM} = \left. \frac{V_{outP} - V_{outN}}{V_{S2}} \right|_{V_{S1}=0} \tag{4.6}$$

$$H_{XCM} = \left. \frac{1}{2} \cdot \frac{V_{outP} + V_{outN}}{V_{S2}} \right|_{V_{S1}=0} \tag{4.7}$$

These transfer functions are calculated as a function of the relative twist positions z_1, z_2 and z_3. This is done in order to be able to optimize these twist positions: we will use the transfer functions above to calculate the optimal twist positions for which both differential mode crosstalk (H_{XDM}) and common-mode crosstalk (H_{XCM}) are minimized. We will see that these optimal positions depend on the chosen termination impedances Z_S and Z_L.

In order to calculate the transfer functions above, we use a modal analysis with (differential) even and odd modes. First, we define four transfer functions:

$$H_{evenP} = \left. \frac{V_{outP}}{V_{S1}} \right|_{V_{S2}=V_{S1}} \tag{4.8}$$

$$H_{evenN} = \left. \frac{V_{outN}}{V_{S1}} \right|_{V_{S2}=V_{S1}} \tag{4.9}$$

$$H_{oddP} = \left. \frac{V_{outP}}{V_{S1}} \right|_{V_{S2}=-V_{S1}} \tag{4.10}$$

$$H_{oddN} = \left. \frac{V_{outN}}{V_{S1}} \right|_{V_{S2}=-V_{S1}} \tag{4.11}$$

With the help of these transfer functions, we are able to calculate

$$\left. \frac{V_{outP}}{V_{S1}} \right|_{V_{S2}=0} = \frac{1}{2} \cdot (H_{evenP} + H_{oddP}) \tag{4.12}$$

$$\left. \frac{V_{outN}}{V_{S1}} \right|_{V_{S2}=0} = \frac{1}{2} \cdot (H_{evenN} + H_{oddN}) \tag{4.13}$$

$$\left. \frac{V_{outP}}{V_{S2}} \right|_{V_{S1}=0} = \frac{1}{2} \cdot (H_{evenP} - H_{oddP}) \tag{4.14}$$

$$\left.\frac{V_{outN}}{V_{S2}}\right|_{V_{S1}=0} = \frac{1}{2}\cdot(H_{evenN}-H_{oddN}) \tag{4.15}$$

and with (4.5) - (4.7) H_{DM}, H_{XDM} and H_{XCM} can be calculated.

Thus, in order to calculate H_{DM}, H_{XDM} and H_{XCM} as a function of twist position, we need to find the four transfer functions H_{evenP}, H_{evenN}, H_{oddP} and H_{oddN}, also as a function of twist position. We can calculate these four transfer functions with the help of s-parameters. However, we cannot use the model of figure 2.9, because if we look at a certain interconnect, after each twist another interconnect becomes its neighboring interconnect. Thus, the twists divide the interconnect into four sections and we will use the s-parameter model of figure 4.6. Note that in the analysis, we do not include the resistance of the vias, which are needed to make the twists: since we only use one or two twists and by using multiple vias per twist, the total via resistance is small compared to the total resistance of the interconnect and can be neglected.
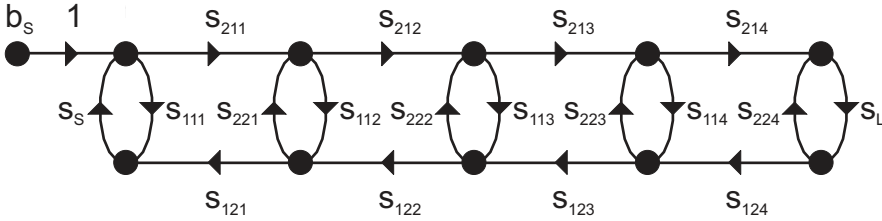
Figure 4.6: S-parameter model of the twisted scheme.

Each section in figure 4.6 has its own set of s-parameters s_{ijk}. In order to calculate these s-parameters, the length l_k, the characteristic impedance Z_{Ck} and the propagation constant γ_k have to be calculated for every section k. For simplicity, we will assume that the interconnect is dominated by RC behavior and that $L' = L_M' = 0$ and $G_G' = G_M' = 0$ (see figure 2.15). Then,

$$Z_{Ck} = \sqrt{\frac{R'}{j\omega(2C_G'+M_kC_M')}} \tag{4.16}$$

$$\gamma_k = \sqrt{j\omega R'(2C_G'+M_kC_M')} \;. \tag{4.17}$$

M_k is a Miller multiplication factor and depends on the signal that is on the neighboring interconnects.

As an example, let us look at the capacitance seen by line 3 in even mode ($V_{S2} = V_{S1}$). For the first section the capacitance to line 2 and for the second section the capacitance to line 1 are seen once (no signals on these lines). As the capacitance to line 4 is seen double (lines 3 and 4 are differentially driven), $M_1 = M_2 = 3$. For the third section, the capacitance to line 6 is seen double (the signal on line 6 has opposite sign) and the capacitance to line 4 is also

seen double, thus $M_3 = 4$. Finally, for the fourth section, the capacitance to line 5 is not seen (the signal on line 5 has equal sign) and again, the capacitance to line 4 is seen double. Therefore, $M_4 = 2$. All values of M_k for the lines 3 and 4 (both in even and odd mode) are shown in table 4.1. Also, the length l_k of every section is given.

k	M_k				l_k
	Line 3		Line 4		
	even	odd	even	odd	
1	3	3	4	2	$z_1 \cdot l$
2	3	3	2	4	$(z_2 - z_1) \cdot l$
3	4	2	3	3	$(z_3 - z_2) \cdot l$
4	2	4	3	3	$(1 - z_3) \cdot l$

Table 4.1: Values of M_k and l_k for the four sections of lines 3 and 4.

With the values for M_k and l_k known, the s-parameters s_{ijk} of every section k can be calculated by using (4.16), (4.17) and the equations of section 2.4.4. Note that in the equation for b_S, $V_S = +\frac{1}{2}V_{S1}$ for line 3 and $V_S = -\frac{1}{2}V_{S1}$ for line 4. Once the s-parameters of every section are known, the transfer functions H_{evenP}, H_{evenN}, H_{oddP} and H_{oddN} can be found by using Mason's Rule on figure 4.6. However, as these equations are very large, they are not reproduced here. Note, that as l_k depends on the twist positions z_1, z_2 and z_3, also the calculated transfer functions depend the twist positions.

With (4.6), H_{XDM} can be calculated. Figure 4.7 shows the differential crosstalk transfer function (H_{XDM}) as a function of z_2 for the example interconnect from section 1.5.2.
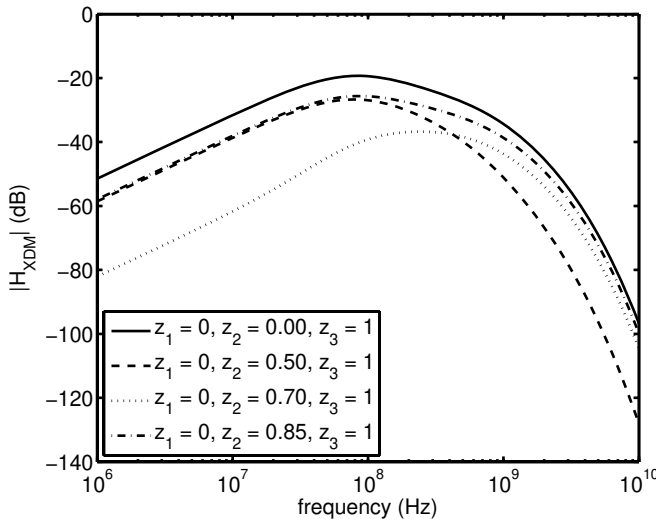
Figure 4.7: Differential crosstalk transfer function as a function of twist position z_2.

The figure shows that a twist at position z_2 ($z_2 \neq 0$) will reduce differential mode crosstalk. The amount of reduction depends on the position of the twist. Note, that for the conventional termination scheme used for this figure, the optimal position of z_2 is 0.7 for low frequencies, while it shifts to 0.5 for higher frequencies.

4.3.3 Optimal position of the single twist

Figure 4.7 indicates that there is an optimal position for the twist at z_2 for which differential mode crosstalk is minimized. In order to find this optimal position, we will look at the differential eye-opening at the output of the interconnect as a function of z_2. For now, $z_1 = 0$ and $z_3 = 1$ (so, these twists are not present yet).

Figure 4.8 shows the differential eye-height relative to the maximum eye-height as a function of z_2. This is not only done for the conventional termination scheme with $Z_S = 100$ Ω, $Z_L = 100$ fF and a data rate of 0.4 Gb/s, but also for a scheme with $Z_S = 100$ Ω and $Z_L = 100$ Ω. As with this last scheme higher data rates are achievable, the data rate for this scheme is 1.2 Gb/s.
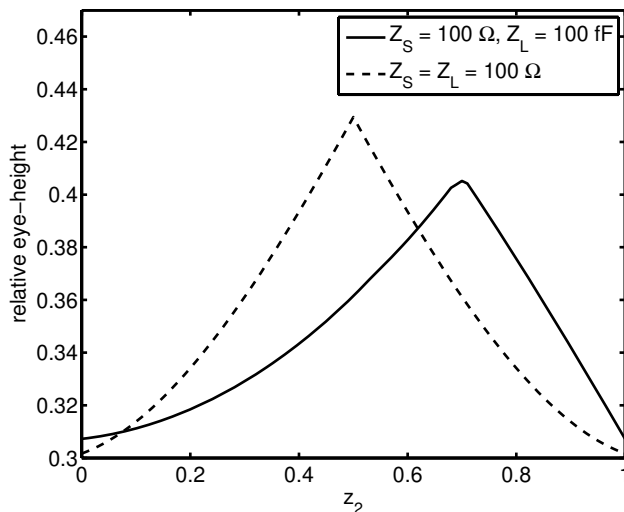
Figure 4.8: Relative differential eye-height as a function of twist position for two different termination schemes.

The figure shows that for the conventional scheme ($Z_S = 100$ Ω, $Z_L = 100$ fF), the optimal position of the twist is at 70%, while for a termination scheme with $Z_S = Z_L = 100$ Ω, the optimal position of the twist is at 50%. So, the optimal position of the twist depends on the termination impedances. In order to understand this result, we will look at a first-order model.

4.3.4 First-order model

A first-order model for crosstalk can be calculated with the same method as in section 3.4.3. With this method, the crosstalk voltage at the end of a neighboring interconnect is

$$V_X = p_v \cdot R_L \cdot \frac{R_T \cdot C_{MT}}{2} \cdot \frac{R_S + \dfrac{R_T}{3} + R_L \cdot (1 + \dfrac{2 \cdot R_S}{R_T})}{(R_S + R_T + R_L)^2} \tag{4.18}$$

with $C_{MT} = l \cdot C_M{}'$.

The crosstalk can also be calculated in another way. In figure 4.9, we first calculate the voltage on the aggressor interconnect (the upper interconnect in the figure) as a function of z. We will call this voltage $V_A(z)$ and $V_A(1) = V_{out}$.
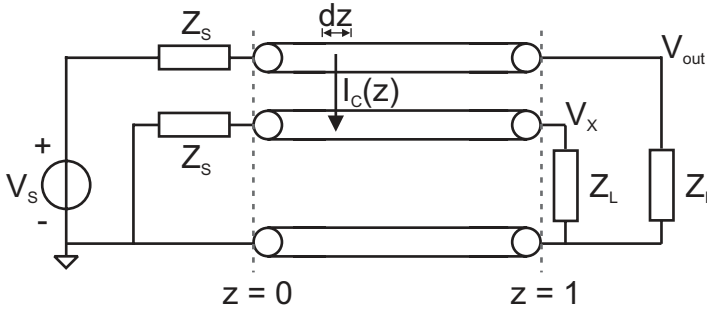
Figure 4.9: Crosstalk between two interconnects.

For low frequencies the voltage $V_A(z)$ simply shows a linear decrease along the line, given by a resistive divider:

$$\frac{V_A(z)}{V_S} = \frac{(1-z) \cdot R_T + R_L}{R_S + R_T + R_L} \tag{4.19}$$

Assuming that the crosstalk voltage at the victim line is much smaller, the voltage $V_A(z)$ is present across a small length dz of the total mutual capacitance C_{MT} and this generates at position z a current:

$$dI_C(z) = \frac{dV_A(z)}{dt} \cdot C_{MT} \cdot dz \tag{4.20}$$

This current sees a resistive divider and only part of the current will contribute to the voltage at the output of the victim line:

$$\frac{dV_X(z)}{dI_C(z)} = R_L \cdot \frac{R_S + z \cdot R_T}{R_S + R_T + R_L} \tag{4.21}$$

90

By using (4.19), (4.20) and (4.21), we can find an expression for the change in V_X as a result of a small length dz at position z:

$$\frac{dV_X}{dz}(z) = \frac{dV_S}{dt} \cdot C_{MT} \cdot R_L \cdot \frac{(R_S + z \cdot R_T) \cdot ((1-z) \cdot R_T + R_L)}{(R_S + R_T + R_L)^2} \quad (4.22)$$

The crosstalk voltage V_X can be found by integrating over z from 0 to 1.

$$V_X = \frac{dV_S}{dt} \cdot R_L \cdot \frac{R_T \cdot C_{MT}}{2} \cdot \frac{R_S + \frac{R_T}{3} + R_L \cdot (1 + \frac{2 \cdot R_S}{R_T})}{(R_S + R_T + R_L)^2} \quad (4.23)$$

This is the same results as in (4.18).

If crosstalk is suppressed by the use of twists in the interconnects, the integration of (4.22) should be carried out for every twist section separately. By adding the results with the right sign, the total crosstalk can be found. If the twist positions are well chosen, the result of the summation will be zero and the crosstalk will be canceled.

With the help of (4.22), we are able to explain the optimal position of z_2 as found in the previous section. In figure 4.10, the result of (4.22) is plotted for both termination schemes of figure 4.8. Because of the difference in load impedance, also the shape of the two lines in the figure is different. The area below the lines is a measure for the amount of crosstalk. For minimal crosstalk, the twist should divide the area below the line into two equal halves. For a conventional scheme with $Z_S = 100\ \Omega$ and $Z_L = 100$ fF, the optimal position for the twist is at 0.7, while for the scheme with $Z_S = Z_L = 100\ \Omega$, the optimal position is at 0.5.
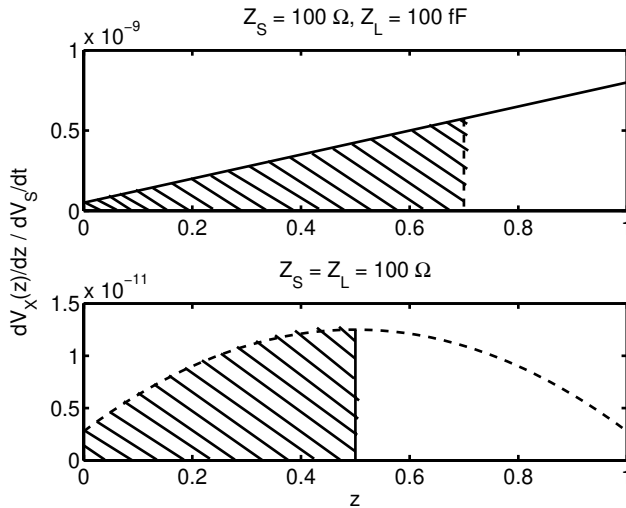
Figure 4.10: Optimum twist position for different termination schemes.

For the conventional scheme, the last part of the interconnect contributes more to the crosstalk at V_X than the first part of the interconnect. Therefore, the optimum twist position is not in the middle, but shifted to the end of the interconnect. However, if the interconnect is terminated in a symmetrical way, the curve that describes the contribution to V_X is also symmetrical and the optimum twist position is in the middle.

The model of this section is only valid for low frequencies. For higher frequencies, the attenuation of the interconnects makes that also for the conventional scheme the crosstalk contribution of the end of the interconnect becomes less important. Therefore, for high frequencies, the optimum position shifts to 0.5 again, also for the conventional scheme. This means, that while with a symmetrical termination scheme the twist position can be chosen optimal for all frequencies, this is not possible with the asymmetrical conventional termination scheme.

4.3.5 Optimal position of the double twist

With the twist at z_2, crosstalk is only canceled differentially. This means, that there still can be crosstalk at the end of the interconnect, but it is equal on both single-ended halves. With the twists at z_1 and z_3, also common-mode crosstalk is canceled.

With figure 4.10, we are also able to predict the optimal positions for the twists at the positions z_1 and z_3. We then have to divide the area below the curves into four equal parts. For the conventional scheme, this is accomplished by choosing $z_1 = 0.5$ and $z_3 = 0.9$, while for the scheme with $Z_S = Z_L$ we should choose $z_1 = 0.3$ and $z_3 = 0.7$.

Figures 4.11 and 4.12 show for different values of z_1 and z_3, both the relative differential eye-height and the maximum value of the common-mode crosstalk symbol response, as calculated with the distributed model given in section 4.3.2. The common-mode crosstalk is normalized relative to the value when $z_1 = 0$ and $z_3 = 1$. This is done for both the conventional termination scheme with $z_2 = 0.7$ in figure 4.11 and the scheme with $Z_S = Z_L$ and $z_2 = 0.5$ in figure 4.12. The optimal positions for z_1 and z_3, for which the differential eye-height is maximized and common-mode crosstalk is minimized, are indicated in the figures with a cross. The figures show that the optimal positions of the double twist also depend on the termination impedances, as already predicted with the help of figure 4.10. By choosing optimal values for z_1, z_2 and z_3 both the differential and the common-mode crosstalk are minimized.

4.3.6 3D EM-field simulation

In order to check the analytical results on the optimal twist positions, a configuration with two differential interconnects has been simulated in the 3D EM-field simulator [43]. The length l_T of the interconnects is only 1 mm to limit the simulation time. Note that for $l_T = 1$ mm, the crosstalk voltage is much lower than for $l_T = 10$ mm.

One of the differential interconnects has one twist and the other has two twists. Figures 4.13 and 4.14 show the simulated crosstalk voltage (step response) for different positions of the twists. The source resistance $R_S = 50\ \Omega$ and the load resistance is either $50\ \Omega$ or $20\ k\Omega$.
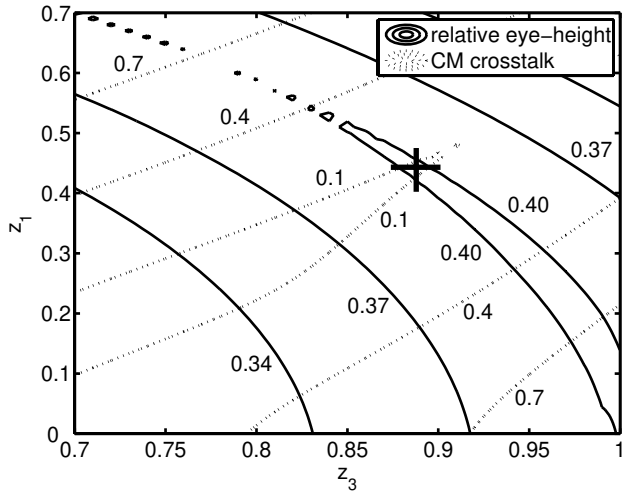
Figure 4.11: Contour plot of relative differential eye-height and normalized maximum common-mode crosstalk as a function of twist positions z_1 and z_3 for the conventional termination scheme with $Z_S = 100\ \Omega$, $Z_L = 100$ fF and $z_2 = 0.7$.
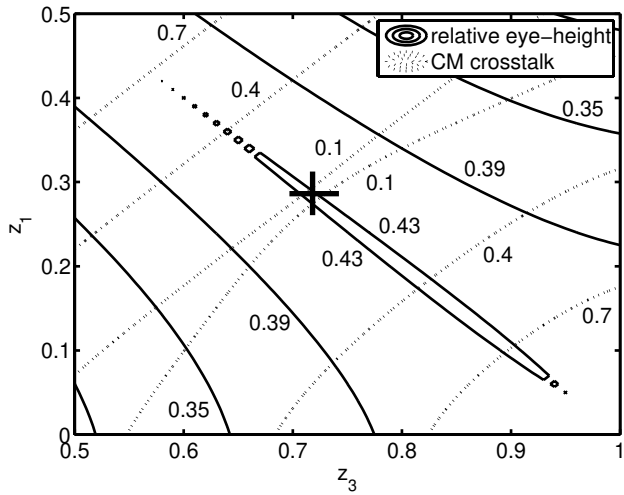
Figure 4.12: Contour plot of relative differential eye-height and normalized maximum common-mode crosstalk as a function of twist positions z_1 and z_3 for the scheme with $Z_S = Z_L = 100\ \Omega$ and $z_2 = 0.5$.
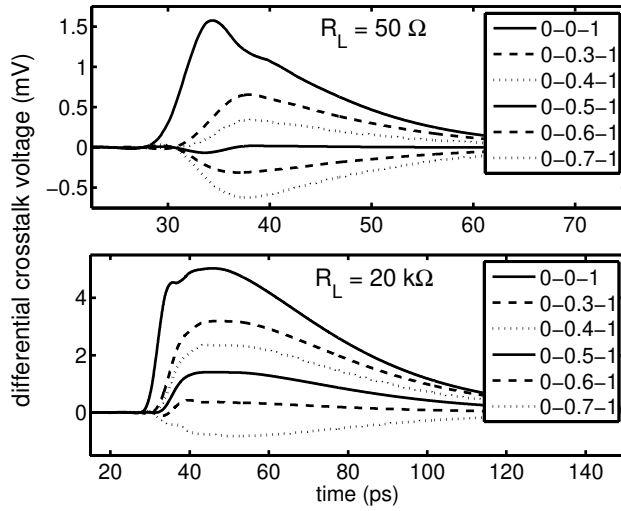
Figure 4.13: 3D EM-field simulation of differential mode crosstalk step response for different positions of the twist (z_1-z_2-z_3) and for two different load resistances. The length of the interconnects l_T = 1 mm.
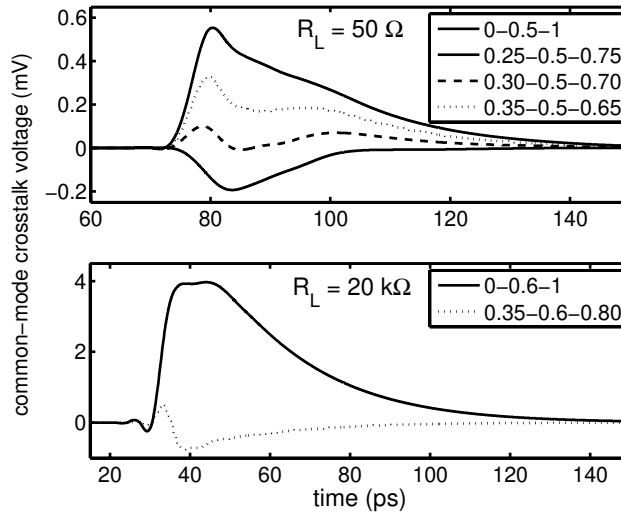
Figure 4.14: 3D EM-field simulation of common-mode crosstalk step response for different positions of the twist (z_1-z_2-z_3) and for two different load resistances. The length of the interconnects l_T = 1 mm.

For differential mode crosstalk, the optimal position of the twist (z_2) is at 0.5 for an R_L of 50 Ω and between 0.6 and 0.7 for an R_L of 20 kΩ. This coincides with the theory: the first-order model of section 4.3.4 predicts 0.5 and 0.64 respectively.

For common-mode crosstalk, the optimal positions of the twists (z_1 and z_3) are at 0.3 and 0.7 for an R_L of 50 Ω and at 0.35 and 0.8 for an R_L of 20 kΩ. Again, this agrees well with the theory that predicts $z_1 = 0.27$ and $z_3 = 0.73$ for an R_L of 50 Ω and $z_1 = 0.37$ and $z_3 = 0.82$ for an R_L of 20 kΩ.

4.3.7 Lumped circuit simulation

Next to the 3D EM-field simulations, we also simulated the optimal twist positions with a lumped circuit model. For these simulations, a lumped RC model was used with 200 lumps per interconnect. For 10 mm long interconnects, we chose $R_{lump} = 7.5$ Ω, $C_{G,lump} = 3.25$ fF and $C_{M,lump} = 2.5$ fF. The twist positions were varied and the crosstalk step responses are shown in figures 4.15 and 4.16. The figures show that $z_2 = 0.5$ gives minimal differential mode crosstalk for $R_L = R_S$ and $z_2 = 0.7$ for R_L high-ohmic. The figure also shows that common-mode crosstalk is reduced by twists at $z_1 = 0.3$ and $z_3 = 0.7$ for $R_L = R_S$ and by twists at $z_1 = 0.5$ and $z_3 = 0.87$ for R_L high-ohmic. Again, these twist positions agree with the analytical results.

4.3.8 Parameter spread

As we have seen in the previous sections, the eye-height at the output of the interconnect is largest for certain optimal positions of the twist. Furthermore, we saw that these optimal positions depend on the termination impedances. The question arises: how sensitive is the eye-height to small variations in either the twist positions or the termination impedances? As an example, figure 4.17 plots the relative eye-height against z_2 (upper figure) and R_L (lower figure). For this figure, the example interconnect from section 1.5.2 is used with a data rate of 1.2 Gb/s, $z_1 = 0$, $z_3 = 1$, $R_S = 100$ Ω and the nominal value of $R_L = 100$ Ω and of $z_2 = 0.5$.

The upper part of the figure shows that the relative eye-height is not much reduced by small changes in the optimal position of z_2. If the position of z_2 is varied with 1% (100 μm), the eye-height is only reduced from 0.428 to 0.427. So, the exact placement of the twists is not critical. The lower part of the figure shows that it is also not critical to exactly match R_S and R_L. The relative eye-height even increases for $R_L < R_S$: although there is slightly more crosstalk, the bandwidth of the interconnect is increased by the smaller load resistance.

4.4 Crosstalk from other metal layers

4.4.1 Introduction

So far, we have looked at crosstalk from neighboring interconnects that are running in parallel and are in the same metal layer. Furthermore, the termination impedances of the neighboring interconnects are equal. In this way, the voltage swing on these interconnects is also equal. However, interconnects that are in another metal layer than the victim metal layer may have full-swing signals. The crosstalk from these full-swing aggressors can be severe. In this section, we look at two cases. First we look at aggressor interconnects that
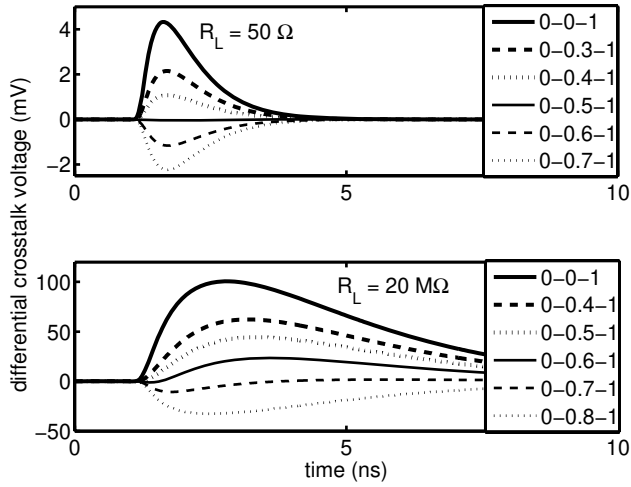
Figure 4.15: Lumped model simulations of differential mode crosstalk step response for different positions of the twist $(z_1\text{-}z_2\text{-}z_3)$ and for two different load resistances. The length of the interconnects $l_T = 10$ mm.
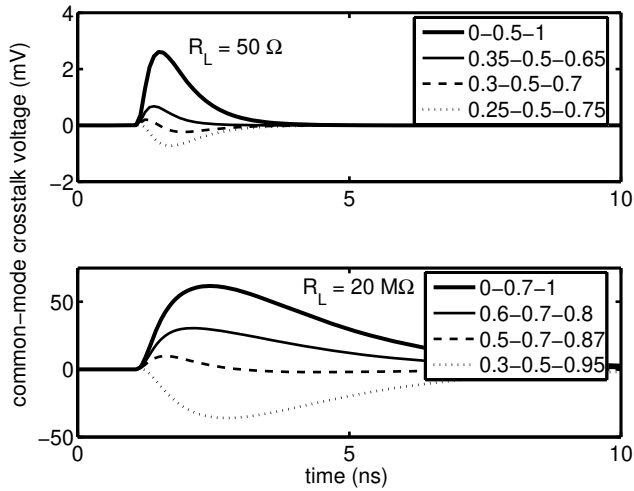
Figure 4.16: Lumped model simulations of common-mode crosstalk step response for different positions of the twist $(z_1\text{-}z_2\text{-}z_3)$ and for two different load resistances. The length of the interconnects $l_T = 10$ mm.
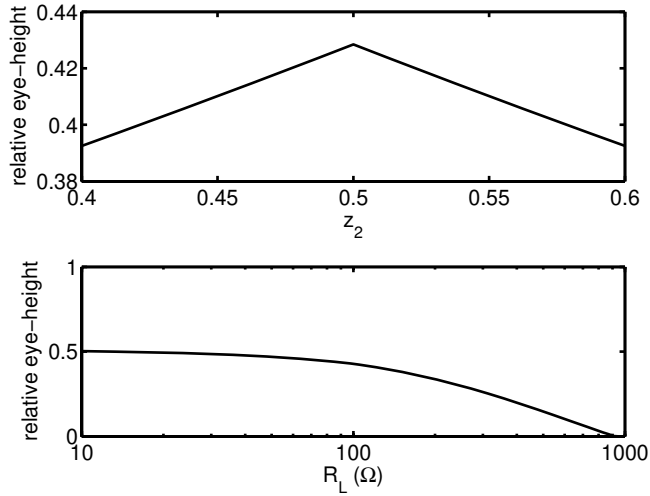
Figure 4.17: Relative eye height as a function of z_2 and R_L.

are routed in a perpendicular direction compared to the victim interconnect. After that, we look at aggressor interconnects that run in parallel with the victim interconnect for a certain length.

4.4.2 Perpendicular interconnects

Often interconnects that are one metal layer below or above an interconnect in a certain metal layer, are routed in a perpendicular direction (Manhattan routing style). This is shown in figure 4.18.
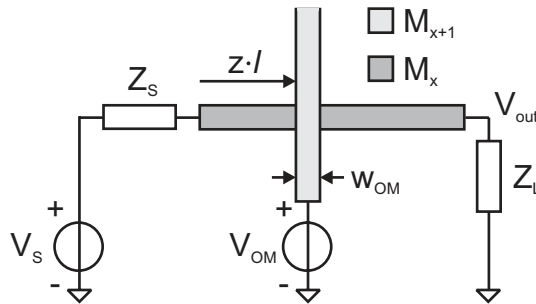
Figure 4.18: Crosstalk from an interconnect in another metal layer that runs in a perpendicular direction.

The interconnect in the other metal layer crosses at a position z and has a width w_{OM}. In this section, we calculate the transfer function from the voltage source V_{OM} to V_{out}. For simplicity, we assume that the aggressor interconnect (in metal layer M_{x+1}) has the same potential at every position, equal to the voltage V_{OM} (worst-case assumption).

We will do again an even and odd mode analysis. However, we cannot simply use $V_S = V_{OM}$ for the even case and $V_S = -V_{OM}$ for the odd case. Then, the voltage of the victim interconnect at position z (at the crossing) would not be equal to the voltage of the aggressor interconnect at position z and the Miller multiplication factor M is not simply 2, 3 or 4 as in the previous section.

Instead, we define the even and odd cases as follows. In the even case, the voltage of the victim interconnect at position z is equal to V_{OM}. In order to have this voltage at position z, V_S should have a higher voltage $a_{even}(z) \cdot V_{OM}$. Likewise, in the odd case, the voltage of the interconnect at position z is equal to $-V_{OM}$ and V_S is equal to $-a_{odd}(z) \cdot V_{OM}$.

The transfer functions are calculated as

$$H = \frac{V_{out}}{V_S} = \frac{a_{even}(z) \cdot H_{even} + a_{odd}(z) \cdot H_{odd}}{a_{even}(z) + a_{odd}(z)} \tag{4.24}$$

$$H_{XOM} = \frac{V_{out}}{V_{OM}} = \frac{a_{even}(z) \cdot H_{even} - a_{odd}(z) \cdot H_{odd}}{a_{even}(z) + a_{odd}(z)} \tag{4.25}$$

with

$$H_{even} = \left. \frac{V_{out}}{V_S} \right|_{V_S = a_{even}(z) \cdot V_{OM}} \tag{4.26}$$

$$H_{odd} = \left. \frac{V_{out}}{V_S} \right|_{V_S = -a_{odd}(z) \cdot V_{OM}} \tag{4.27}$$

The even and odd mode transfer functions can again be calculated with an s-parameter model of the interconnect. The s-parameter model now has three sections: the first section is from the source impedance to z, the second section is from z to z+w_{OM} and the last section is from z+w_{OM} to the load impedance.

Figure 4.19: S-parameter model for crosstalk from a perpendicular interconnect in another metal layer.

Every section k has a different characteristic impedance and propagation constant.

$$Z_{Ck} = \sqrt{\frac{R'}{j\omega M_k C_G'}} \qquad (4.28)$$

$$\gamma_k = \sqrt{j\omega R' M_k C_G'} \,. \qquad (4.29)$$

Note that because of the orthogonal routing, the crosstalk on the neighbors of the victim is equal. Therefore, in calculating the transfer functions, we should leave out the capacitances to these neighboring interconnects.

For the first and third section $M_k=2$, if we assume that there is metal in all metal layers and that these are connected to ground. The value of M_k in the second section depends on the even or odd mode. In even mode, we assume the voltage of the interconnect at position z to be equal to V_{OM} and thus $M_k=1$. In odd mode, we assume the voltage of the interconnect at position z to be equal to $-V_{OM}$ and thus $M_k=3$. Now the value of M_k is found for every section, the s-parameters of all sections are known and the transfer functions H_{even} and H_{odd} can be calculated with the help of Mason's rule.

However, in order to find H and H_{OM}, we also need to know $a_{even}(z)$ and $a_{odd}(z)$. These can be approximated by calculating the transfer function from V_S to the nodes b_{OM} and a_{OM} in the s-parameter model and taking the inverse of this transfer function. This can again be done with the help of Mason's rule.

As an example, figure 4.20 shows H_{XOM} for a 1 μm wide aggressor interconnect crossing the example interconnect from section 1.5.2. This is done for different positions z.
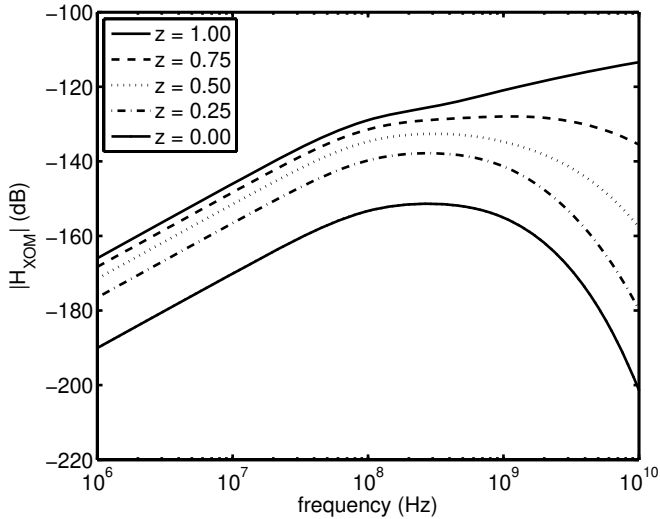
Figure 4.20: Crosstalk from a 1 μm wide crossing interconnect in another metal layer.

The figure shows that the closer the crossing interconnect is to the end ($z = 1$), the higher the crosstalk.

A worst-case scenario would have aggressor interconnects along the entire length of the interconnect. For figure 4.21, we have assumed 1 μm-wide full-swing aggressors with a spacing of also 1 μm. These aggressors are placed in both metal layers M_{x+1} and M_{x-1} along the entire length of the interconnect. The signals of these aggressors, with a symbol time of 1 ns and a rise time of 10 ps, are random and uncorrelated. The victim interconnect has the capacitive termination scheme, as discussed in section 3.3.
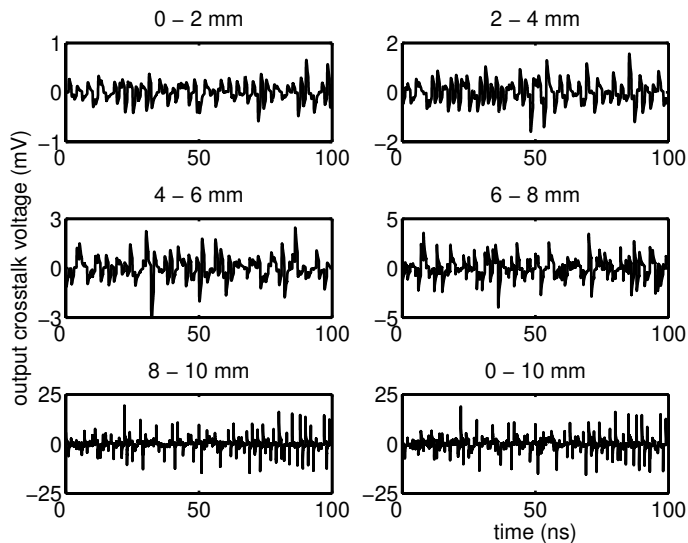
Figure 4.21: Crosstalk on a 10 mm long victim interconnect from 1 μm-wide, 1 μm-spaced orthogonal interconnects in both one metal layer above and one metal layer below the victim interconnect.

The figure shows that the amount of crosstalk that is introduced by aggressors crossing at the beginning of the victim interconnect is small. The crosstalk from aggressors at the end of the victim interconnect is much larger.

An effective way of reducing this crosstalk is the use of differential interconnects. The crosstalk from the orthogonal aggressor on both single-ended halves will be equal and the crosstalk will appear as common-mode. The receiver then should have good common-mode suppression. The differential interconnect has the further advantage of being robust to substrate noise and supply noise. Furthermore, as we have seen in the previous section, with a differential interconnect we can make twists to also cancel crosstalk from neighboring interconnects that run in parallel.

4.4.3 Full-swing interconnect running in parallel

Before we have assumed that full-swing aggressors in another metal layer are running orthogonal to the victim interconnect. We now look at a case where full-swing aggressors are running in parallel with a victim interconnect. We will look at the eye-opening of the victim interconnect and see how much it is degraded by crosstalk from a full-swing aggressor that runs in parallel only for a certain length.

Figure 4.22 shows a differential interconnect (victim), driven by V_{SP} and V_{SN} ($V_{SN} = -V_{SP}$), with a differential output voltage $V_{out} = V_{outP} - V_{outN}$. This differential interconnect is placed in metal M_x. In metal layer M_{x+1}, an aggressor interconnect is driven by voltage source V_{OM}. This aggressor interconnect runs in parallel with both single-ended halves of the victim for a certain length: from z_{start} to the end of the interconnect.
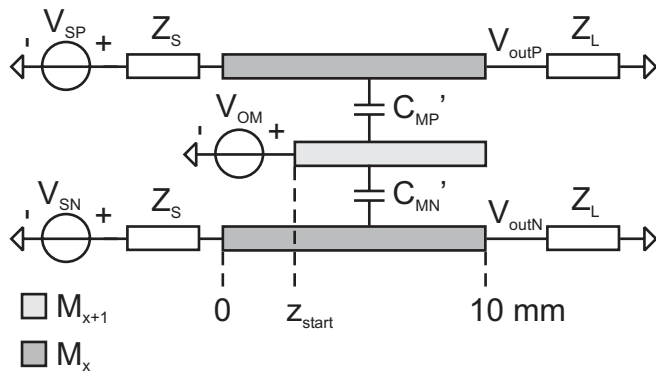
Figure 4.22: Crosstalk from a full-swing aggressor that runs in parallel with a differential interconnect from z_{start} to 10 mm.

There are distributed capacitances from the aggressor to both single-ended halves of the victim interconnect, C_{MP}' and C_{MN}'. These capacitances generate crosstalk from voltage source V_{OM} to V_{outP} and V_{outN}. The total mutual capacitances to the single-ended halves are $C_{MPT} = (10 \text{ mm} - z_{start}) \cdot C_{MP}'$ and $C_{MNT} = (10 \text{ mm} - z_{start}) \cdot C_{MN}'$. The total mutual capacitance to the differential interconnect is defined as $C_{MT} = C_{MPT} - C_{MNT}$.

C_{MPT} and C_{MNT} do not have to be equal and C_{MT} can have different values. The value of C_{MT} depends on the position of the aggressor interconnect relative to both single-ended halves, as shown in figure 4.23. In this figure, C_{MP}' would be larger than C_{MN}'.
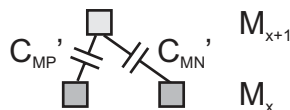
Figure 4.23: The capacitances C_{MP}' and C_{MN}' can have different values, depending on the relative position of the aggressor interconnect to both single-ended halves of the victim interconnect.

We assume that the bandwidth of the aggressor interconnect is large and the aggressor has full-swing signals along the entire interconnect (worst-case assumption). For the victim interconnect, we use the parameters of the example interconnect from section 1.5.2 with Z_S = 275 fF (capacive transmitter, see section 3.3).

Figure 4.24 shows the resulting absolute eye-height at the differential output of the victim interconnect for a data rate of 1 Gb/s. Without crosstalk (z_{start} = 10 mm), the absolute eye-height is 0.1 V. For other values of z_{start}, the absolute eye-height decreases due to the crosstalk from the aggressor interconnect. The figure gives the result for different values of C_{MT}. C_T is the total capacitance of a single-ended half of the victim interconnect
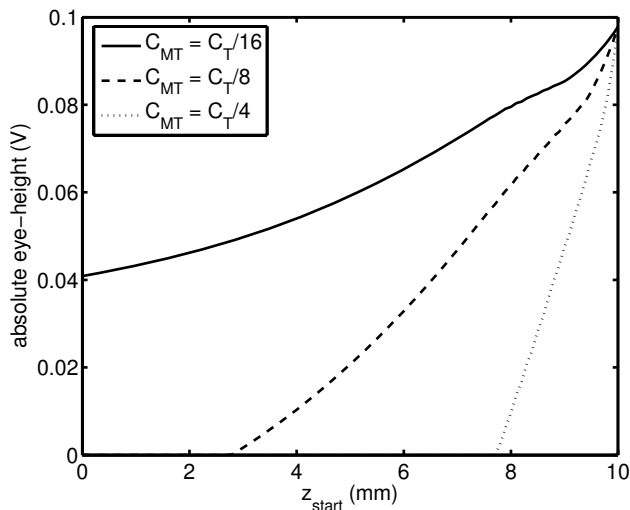
Figure 4.24: Eye-closure due to an aggressor interconnect that runs from z_{start} to 10 mm in parallel with a 10 mm long victim interconnect.

If only the positive single-ended half receives crosstalk ($C_{MPT} = C_T/4$ and $C_{MNT} = 0$), then the length of the aggressor interconnect can be a little more than 2 mm before the eye is closed. However, if the other single-ended half also receives crosstalk, for instance $C_{MNT} = C_T/8$, the total mutual capacitance seen by the differential interconnect is $C_{MT} = C_T/4 - C_T/8 = C_T/8$. Then, the full-swing aggressor interconnect can already run 7 mm in parallel. If the capacitance from the aggressor to two both single-ended halves is about equal, for instance with only a difference of $C_T/16$, the aggressor can run in parallel for the total 10 mm and the eye still stays open (40%). In summary, depending on the capacitance of the aggressor interconnect to both single-ended halves of the victim interconnect, the aggressor can run in parallel for 2 to the full 10 mm, without completely closing the eye.

4.5 Summary

− With a low-ohmic load resistance, the voltage swing at the receiver side can be very low. By using a transimpedance amplifier, the transceiver becomes less susceptible to offset.
− Crosstalk from neighboring interconnects in a bus can be canceled by using differential interconnects with twists. A single twist in every even differential interconnect cancels differential crosstalk and a double twist in every odd differential interconnect cancels common-mode crosstalk.
− The optimal positions of the twists depend on the termination impedances. For both a low source and a low load impedance, the optimum position of the single twist is at 50% of the interconnect. However, for a low source impedance, but a high load impedance, the optimum shifts to 70%.
− The optimum positions of the twist can be explained with a first-order model.
− Crosstalk from orthogonal aggressors in another metal layer than the victim interconnect can be large, but is suppressed with differential interconnects.
− A full-swing aggressor in another metal than a differential victim interconnect can run in parallel with the victim for 2 to 10 mm, depending on the capacitance to both single-ended halves of the victim.

Chapter 5

Circuit implementations

5.1 Introduction

In chapters 3 and 4 we discussed the design of interconnects, introduced termination and equalization concepts and discussed methods to counteract offset and crosstalk. In this chapter, we translate these concepts into CMOS circuits.

In sections 1.5.5 and 3.2 we mentioned that we design our interconnects for optimal bandwidth per cross-sectional area and in section 5.2 this is applied to interconnects in two different CMOS technologies. This results in a certain width and spacing for the interconnect and a 3D EM-field simulator [43] is used to extract the distributed parameters of the interconnect. Once these parameters are known, we are able to design transceivers that achieve a high data rate over these interconnects, preferable with low power consumption.

In section 5.3, a transceiver is described that is designed to achieve a high data rate. The transceiver uses the pulse-width pre-emphasis equalization from section 3.7.2 and a low-ohmic load resistance as described in section 3.4. Both transmitter and receiver circuits are discussed and measurement results are given. In order to reach the high data rate, crosstalk is canceled with the help of twists and crosstalk measurements that show the effectiveness of these twists are given in section 5.4.

In section 5.5 another transceiver is described that is not only designed to achieve a high data rate, but also to have low power consumption. Now, the capacitive pre-emphasis transmitter of section 3.3.6 is used together with the decision feedback equalization technique of section 3.7.3. Again, both transmitter and receiver circuits are discussed and measurement results are given. Finally, in section 5.6, the measured transceivers are compared to other solutions with respect to speed and power consumption.

5.2 Interconnect design

5.2.1 Introduction

A designer can choose the width and spacing of interconnects. As described in section 3.2, we choose both the width and spacing in such a way, that the bandwidth per cross-sectional area is maximized. In order to calculate the bandwidth as a function of width and spacing, we first describe the technologies, in which the interconnects are designed. Then, the optimal width and spacing are calculated. With the dimensions of the interconnect known, we find the distributed parameters of the interconnect with a 3D EM-field simulator [43].

5.2.2 Technology

In this thesis, two CMOS technologies are used: CMOS 0.13 μm and CMOS 90 nm. The CMOS 0.13 μm process is used for a transceiver that achieves a high data rate (see sections 5.3 and 5.4) and the CMOS 90 nm process is used for a transceiver that also achieves a high data rate, but now with much lower power consumption (see section 5.5). The first process has a low-V_T option. Figure 5.1 shows part of the layers that are present in the two processes.
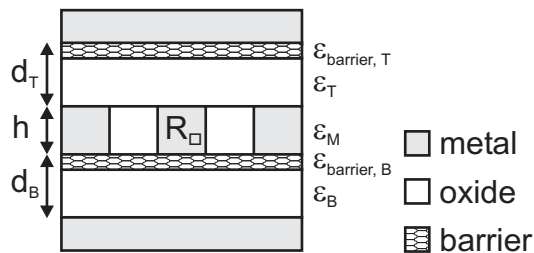
Figure 5.1: The interconnect with a certain height h has a sheet resistance R□ and is surrounded by oxides and barrier layers with a certain permittivity.

As already mentioned in section 2.2, we do not use the thick top metal layers for our interconnects. Firstly, these thick top metal layers are often reserved for the power and clock grid. Secondly, our research is focused on bandwidth-limited interconnects as interconnects in future CMOS technologies are expected to have a very low bandwidth (see chapter 1). Therefore, it makes sense to use one of the thinner metal layers, which have a higher distributed resistance.

The oxide above and below the interconnect (top and bottom oxide) have a barrier layer with a higher permittivity than the oxide itself, making the effective permittivity of the top and bottom oxide larger than the permittivity of the oxide between the interconnects.

5.2.3 Optimal bandwidth per cross-sectional area

For (3.6), which gives the optimal bandwidth per cross-sectional area, we have assumed only one permittivity for all oxides. However, as shown in figure 5.1, there are barrier layers that have another dielectric constant than the oxide layers. We can still use the equation by calculating an effective thickness d_{eff}, while taking a single permittivity ε_{eff} and assuming that all permittivities are equal to this ε_{eff}. Two oxide layers on top of each other and in between metal can be thought of as a series connection of two capacitors. If a first capacitor has a value of $\varepsilon_1 \cdot A/d_1$ (with A the area of the metal) and a second capacitor has a value of $\varepsilon_2 \cdot A/d_2$, then the effective capacitance of a series connection of the two capacitors is

$$C_{eff} = \frac{\dfrac{\varepsilon_1}{d_1} \cdot \dfrac{\varepsilon_2}{d_2}}{\dfrac{\varepsilon_1}{d_1} + \dfrac{\varepsilon_2}{d_2}} \cdot A \tag{5.1}$$

Using $C_{eff} = \varepsilon_{eff} \cdot A/d_{eff}$ and solving for d_{eff} gives

$$d_{eff} = \varepsilon_{eff} \cdot \frac{\dfrac{\varepsilon_1}{d_1} + \dfrac{\varepsilon_2}{d_2}}{\dfrac{\varepsilon_1}{d_1} \cdot \dfrac{\varepsilon_2}{d_2}} \tag{5.2}$$

For the CMOS 0.13 µm process, the effective thickness of the top oxide is:

$$d_{T,eff}\Big|_{0.13\ \mu m} \approx 0.55 \mu m \tag{5.3}$$

and of the bottom oxide:

$$d_{B,eff}\Big|_{0.13\ \mu m} \approx 0.38 \mu m \tag{5.4}$$

For (3.6), we further assume that the top and bottom oxides have the same thickness. Therefore, we take the average of the $d_{T,eff}$ and $d_{B,eff}$.

$$d_{eff}\Big|_{0.13\ \mu m} = \frac{0.55\mu + 0.38\mu}{2} \approx 0.46 \mu m \tag{5.5}$$

Now, we can optimize the bandwidth per area with the help of (3.6), which is repeated here for convenience.

$$\frac{BW}{area} \propto \frac{1}{\dfrac{w+s}{h} + \dfrac{w+s}{d} + \dfrac{h+d}{w} + \dfrac{h+d}{s}} \tag{5.6}$$

For the CMOS 0.13 µm process, we use this equation to find the optimal bandwidth per cross-sectional area to be proportional with the following expression.

$$\frac{BW}{area} \propto \frac{1}{\frac{w+s}{0.35\mu} + \frac{w+s}{0.46\mu} + \frac{0.81\mu}{w} + \frac{0.81\mu}{s}} \qquad (5.7)$$

The optimum of this expression is at w = s = 0.4 µm. For the circuits in the CMOS 0.13 µm process we use this value for both width and spacing of the interconnect.

However, by calculating the optimal bandwidth per cross-sectional area in this way, some second order effects are not included. For instance, the fact that we use differential interconnects makes that the side capacitance to the other single-ended half is seen double (Miller multiplication). Furthermore, in (5.6) we have neglected fringe capacitances. In order to include these effects for the CMOS 90 nm process, we do not use (5.6), but use the following approach.

In the 3D EM-field simulator [43] we have simulated interconnects with various widths and spacings. From these simulations, we approximate the total resistance R_T, the total capacitance C_T and the total mutual capacitance C_{MT} for a 10 mm long interconnect. R_T only depends on the width w and is for this process approximated by

$$R_T = \frac{72 \cdot 10^{-6}}{w} \quad \Omega \qquad (5.8)$$

The total capacitance as used in (5.6) has a term proportional to w and another term proportional to 1/s. In order to include the fringe capacitance, we now also use a term proportional to s. This approximation is only valid for small values of s. Then, if s increases, the fringe capacitance to the top and bottom metal layers will increase.

$$C_T = 1.90 \cdot 10^{-6} \cdot w + \frac{3.10 \cdot 10^{-19}}{s} + 1.25 \cdot 10^{-6} \cdot s \quad F \qquad (5.9)$$

The mutual capacitance to a neighboring interconnect only depends on the spacing s

$$C_{MT} = \frac{1.25 \cdot 10^{-19}}{s} \quad F \qquad (5.10)$$

The formulas are plotted in figures 5.2 and 5.3 together with the simulated results from the 3D EM field simulator [43], showing the usefullness of the formulas. Note that the formulas are only accurate for the CMOS 90 nm process and only for small values of w and s.
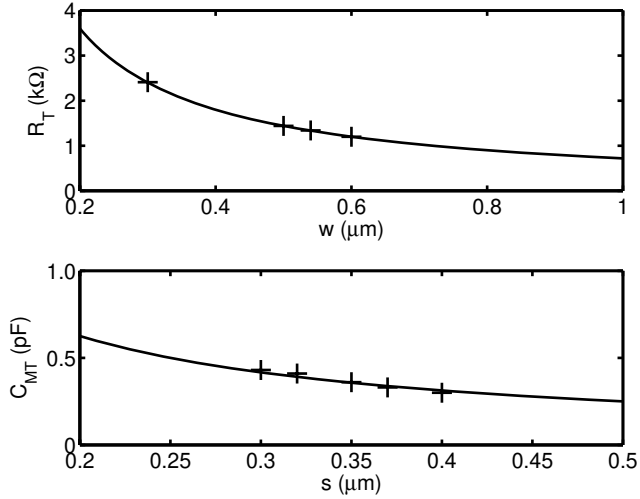
Figure 5.2: Formulas for R_T and C_{MT} for a 10 mm interconnect in CMOS 90 nm as a function of interconnect spacing.
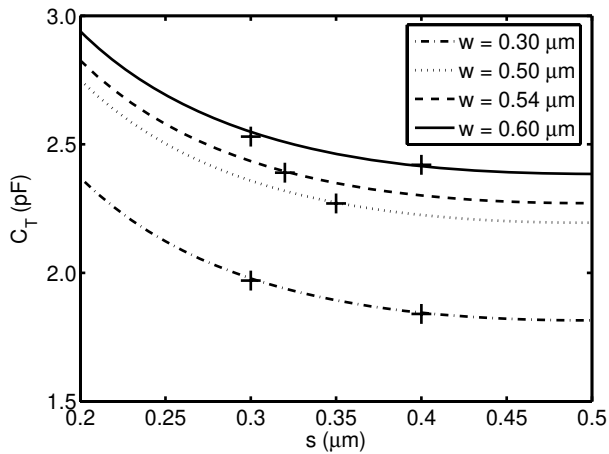
Figure 5.3: Formulas for C_T for a 10 mm interconnect in CMOS 90 nm as a function of interconnect spacing for different interconnect widths.

With these formulas for R_T, C_T and C_{MT}, we are able to plot the bandwidth per cross-sectional area as a function of w and s. This is done in figure 5.4, which shows the bandwidth of a 10 mm long differential interconnect,

$$BW = \frac{1}{\pi \cdot R_T \cdot (C_T + C_{MT})} \quad Hz \tag{5.11}$$

divided by a cross-sectional area of $(w + s) \cdot (0.62)$ μm^2.
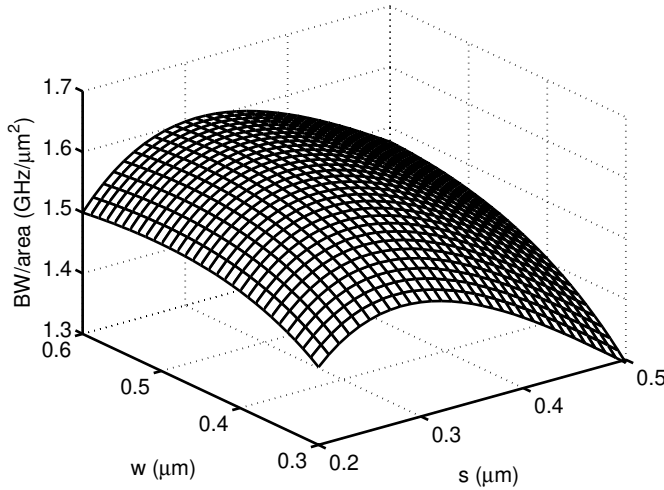
Figure 5.4: Bandwidth per cross-sectional area as a function of width (w) and spacing (s) for a 10 mm interconnect in CMOS 90 nm.

For the circuits in the CMOS 90 nm process, we use $w = 0.54$ μm and $s = 0.32$ μm, which is near the optimum bandwidth per cross-sectional area. Note that the inclusion of the second order effects to calculate the optimum bandwidth per cross-sectional area makes the optimal width somewhat larger, as we would expect an optimal width about equal to the optimal spacing. Further note, that the optimum is flat and for a small deviation in the optimal width and spacing not much bandwidth per cross-sectional area is lost.

5.2.4 Interconnect parameters

In the 3D EM-field simulator [43], interconnects with the dimensions of the previous section were simulated. As described in section 2.7, it is possible to map the results of this simulator to the distributed model of figure 2.4 or 2.15. In this way, the distributed parameters of the interconnect are found, which are used during the design of the transceivers in the following sections.

The structure of figure 5.5 is used in the 3D EM-field simulator [43]. The total width of the structure is 10 μm. In order to limit simulation time, the total length of the interconnects and metal plates is chosen at 1000 μm. For a 10 mm long interconnect, the total resistance, inductance, conductance and capacitance are simply ten times larger.

M6

M5

M4

M3

interconnects driven in even and odd mode
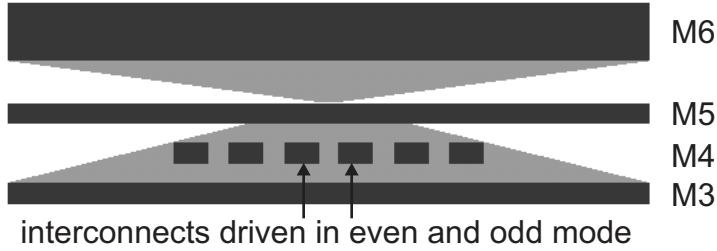
Figure 5.5: Structure used in 3D EM-field simulator [43]. The width is 10 μm and the length is 1000 μm.

For the CMOS 0.13 μm process, the total resistance R_T, total inductance L_T, total conductance G_T and total capacitance C_T of a 10 mm long interconnect are plotted in figure 5.6.
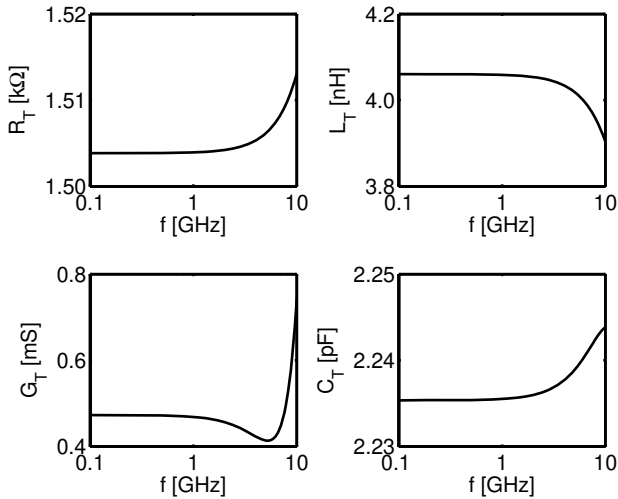
Figure 5.6: R_T, L_T, G_T and C_T for a 10 mm interconnect in CMOS 0.13 μm.

From the figure, we can calculate the distributed resistance R' = 150 kΩ/m, the distributed inductance L' = 406 nH/m, the distributed conductance G' = 47 mS/m and the distributed capacitance C' = 223 pF/m. The figure shows some frequency dependence of the parameters due to for instance the skin effect. However, for the frequencies of interest, this effect is small and we use constant numbers for the parameters.

The total ground resistance R_{GT}, total mutual inductance L_{MT}, total mutual conductance G_{MT} and total mutual capacitance are C_{MT} are plotted in figure 5.7. The resistance of the ground plane in the structure as drawn in the 3D EM field simulator is 15 Ω and is much smaller than the resistance of the interconnect. The mutual inductance of 1.4 nH/m and the
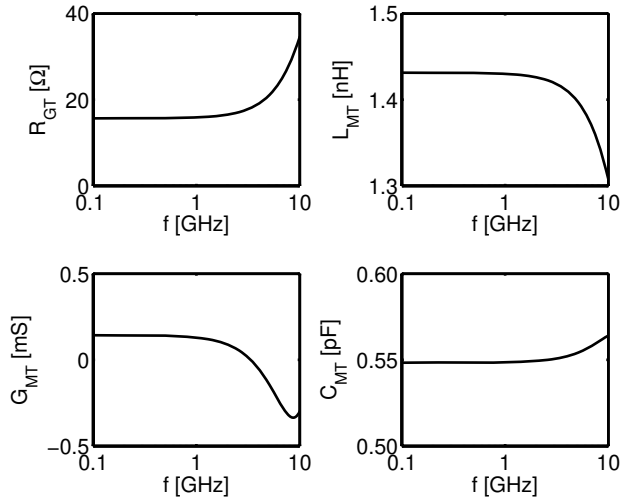
Figure 5.7: R_{GT}, L_{MT}, G_{MT} and C_{MT} for a 10 mm interconnect in CMOS 0.13 μm.

mutual capacitance of 55 pF/m are about one fourth of the total inductance L_T and total capacitance C_T respectively.

For the CMOS 90 nm process, simulations with the 3D EM-field simulator gave the results of figures 5.8 and 5.9 for a 10 mm long interconnect.
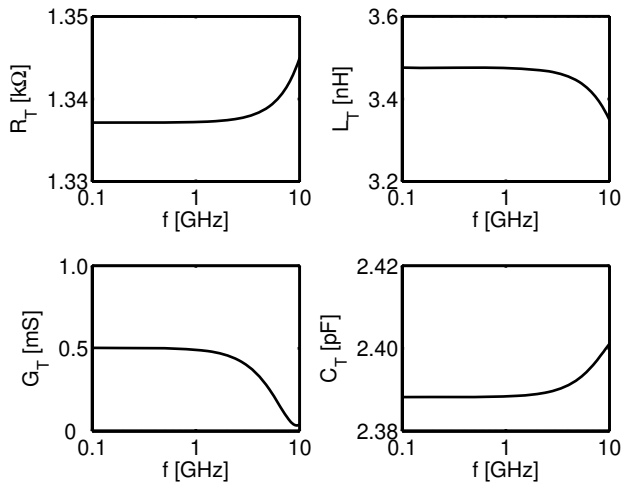
Figure 5.8: R_T, L_T, G_T and C_T for a 10 mm interconnect in CMOS 90 nm.
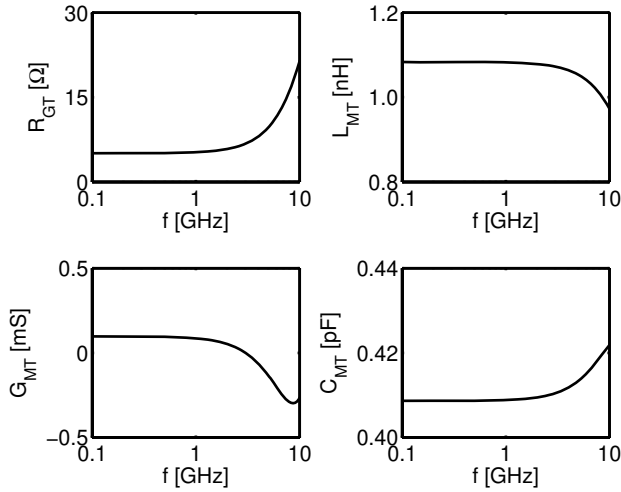
Figure 5.9: R_{GT}, L_{MT}, G_{MT} and C_{MT} for a 10 mm interconnect in CMOS 90 nm.

From the figures, the following parameter values are found: R' = 134 kΩ/m, L' = 350 nH/m, G' = 50 mS/m and C' = 239 pF/m. The ground resistance in this simulation is only 5 Ω. The mutual inductance L_M' is 110 nH/m and the mutual capacitance C_M' = 41 pF/m.

For the design of the transceivers of the following sections, the most important distributed parameters are the resistance and capacitance, which determine the bandwidth, and the mutual capacitance, which determines the amount of crosstalk. Table 5.1 summarizes these parameters for both technologies.

	CMOS 0.13 μm	CMOS 90 nm
R' (Ω/m)	150k	134k
C' (F/m)	223p	239p
C_M' (F/m)	55p	41p

Table 5.1: Distributed resistance, capacitance and mutual capacitance of the two used technologies.

5.3 High speed transceiver

5.3.1 Introduction

In the previous section, we have found the distributed parameters of interconnects in two CMOS technologies. In both technologies, the distributed resistance and the distributed capacitance are large, giving a low bandwidth, limiting the achievable data rate.

In order to increase the achievable data rate, we have designed a transceiver [6, 52] in the CMOS 0.13 μm process that uses two of the techniques of chapter 3: pulse-width pre-

emphasis at the transmitter side (see section 3.7.2) and a low-ohmic termination at the receiver side (see sections 3.4 and 4.2). In order to be able to compare the techniques with a conventional transceiver, both the pulse-width pre-emphasis equalization and low-ohmic termination can be turned on and off.

5.3.2 Transmitter circuits

In the conventional case, the interconnect is driven by a large inverter. This drive inverter should be strong enough, as (3.18) predicts that the source impedance should be well below the resistance of the interconnect for lowest delay and highest bandwidth. However, the input capacitance of this large drive inverter will also be large. Therefore, if we drove this inverter with a minimum sized inverter, the delay would be large. A delay-optimal solution is to create a chain of inverters that have an increasing width, as shown in figure 5.10.
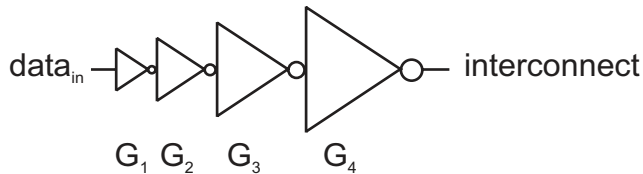
Figure 5.10: Inverter chain with increasing width for minimal delay.

For the CMOS 0.13 μm process, the inverter chain was designed with the following parameters.

Inverter	PMOS	NMOS
	W/L [μm/μm]	W/L [μm/μm]
G_1	0.7/0.13	0.3/0.13
G_2	2.1/0.13	0.9/0.13
G_3	6.3/0.13	2.7/0.13
G_4	23.2/0.13	7.0/0.13

Table 5.2: Dimensions for the chain of inverters.

The ratio between PMOS width and NMOS width is 2.3 for the first three (small) inverters and 3.3 for the last (large) inverter. Simulations showed this to be optimal. For minimal delay, the inverter sizes from one inverter to the next are increased with a factor 3. Again, simulation showed that in this way, the delay of the total inverter chain is minimized.

In order to increase the achievable data rate, pulse-width pre-emphasis equalization is used (see section 3.7.2). The implementation is shown in figure 5.11. Details about the circuit implementation can be found in [6, 42].

The amount of equalization is controlled by a current, I_{bias}. In this way, the pulse-width of the input signals can be set between 50% and 100%. For 100% pulse-width, the pulse-width
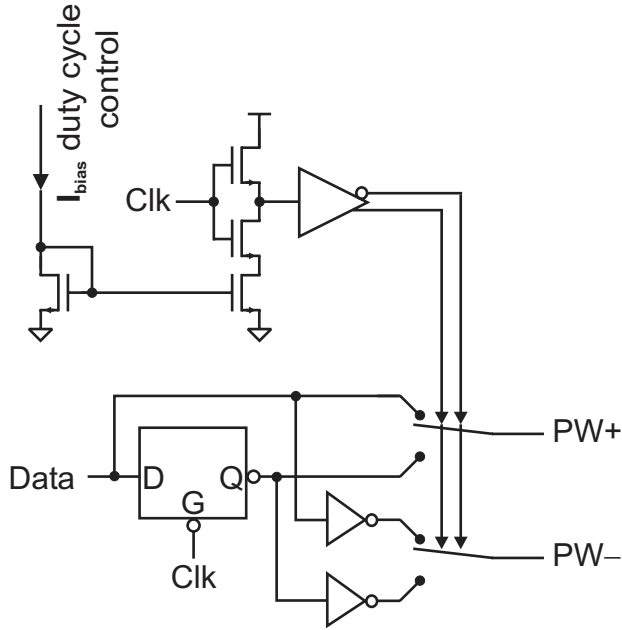
Figure 5.11: Circuit implementation of the pulse-width pre-emphasis equalization transmitter.

pre-emphasis equalization is effectively turned off. The differential output (PW+ and PW−) are both input to an inverter chain, as described above. The total area of the transmitter is about 300 μm^2. Half of the area is occupied by the inverter chains and half of the area by the pulse-width pre-emphasis equalization.

5.3.3 Receiver circuits

At the receiver side, a low-ohmic load resistance is used to further increase the achievable data rate. The low-ohmic termination is implemented with a transimpedance amplifier (see section 4.2). The schematic of the transimpedance amplifier is shown in figure 5.12.
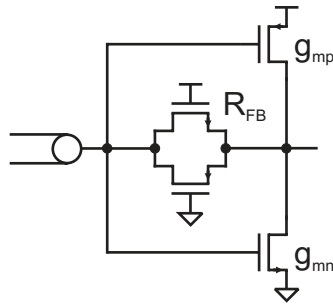
Figure 5.12: Receiver circuit with low input impedance.

This circuit is an implementation of figure 4.2. The G_M of the inverter is $g_{mp} + g_{mn} = 5.2 + 5.0 = 10.2$ mS with a total output impedance of 2 kΩ. For the PMOS W/L = 27.5 μm / 0.13 μm and for the NMOS W/L = 9.5 μm / 0.13 μm, giving a bias current of 0.7 mA.

The feedback resistance R_{FB} has a value of 1.07 kΩ and is made with a passgate. On the test chip we are able to turn the passgate on or off. In this way, we can either have a low-ohmic load resistance or a high-ohmic load impedance.

With the equations of section 4.2, we are able to calculate the input impedance of the transimpedance amplifier, the swing at V_{out} (the output of the interconnect and input of the transimpedance amplifier) and the small-signal gain from input to output of the transimpedance amplifier. The results for a single-ended half are in table 5.3.

Passgate	R_{in}	Swing at V_{out}	Small-signal gain
	[Ω]	[V]	
off	∞	0.60	-20.4
on	143	0.10	-6.5

Table 5.3: Equivalent input resistance, swing at output of the interconnect and gain by the transimpedance amplifier for passgate on or off.

So, on our test chip we are able to either choose a high-ohmic load impedance or a low-ohmic load resistance of 143 Ω.

The total area of the transimpedance amplifier is about 300 μm^2 and the output-equivalent offset is about 7.5 mV (one-sigma).

After the transimpedance amplifier, a sense amplifier is used to restore the signals to full-swing. The schematic is in figure 5.13.
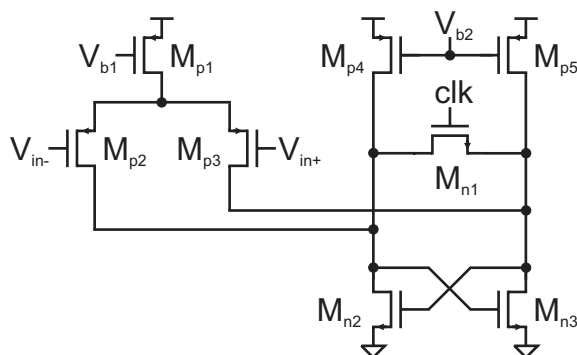
Figure 5.13: Sense amplifier to restore signals to full-swing.

The dimensions of the transistors are chosen for low-offset with minimal power consumption. The input-equivalent offset of the sense amplifier is comparable to the output-equivalent offset of the transimpedance amplifier. This gives a total one-sigma

offset of 10 mV at the input of the sense amplifier. The current of the input pair is 200 μA and the current source M_{p4} and M_{p5} have a current of 100 μA. So, the total static current is 400 μA.

The current sources are biased with V_{b1} = 0.85 V and V_{b2} = 0.6 V. The circuit that is used for this is in figure 5.14. Both V_{b1} and V_{b2} are generated with current mirrors from a single resistor current.
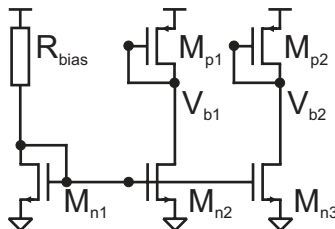
Figure 5.14: Bias circuit for sense amplifier.

The output of the sense amplifier is followed by a dynamic latch, as shown in figure 5.15.
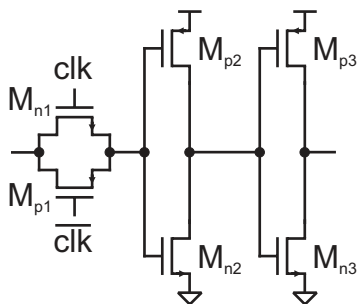
Figure 5.15: dynamic latch.

The total area of the sense amplifier and dynamic latch is about 500 μm^2.

5.3.4 Measurement setup

The micrograph of a demonstrator IC is shown in figure 5.16. The chip is used to see whether the pulse-width pre-emphasis equalization at the transmitter and the low-ohmic termination at the receiver indeed increase the achievable data rate.

A seven-channel differential bus with twisted wires (width and spacing of 0.4 μm each, optimized as explained in section 5.2.3) is placed in metal 5 (6 metal layer process) and is completely surrounded by GND/V_{DD}-connected metal stripes. These metal stripes emulate high-density metal use in all surrounding metal layers. An additional seven-channel single-ended bus with perpendicular orientation is placed below the differential bus for additional characterization purposes (a variety of wire pitches is used in this bus) and to provide an indication of interlayer crosstalk. An external single-channel 3.2 Gb/s pattern
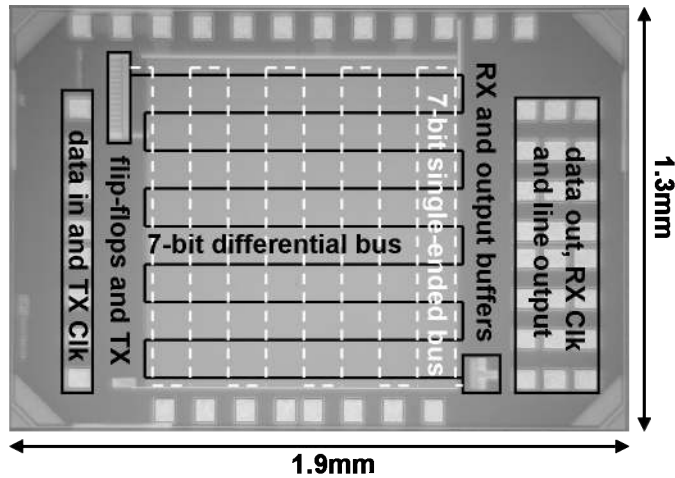
Figure 5.16: Chip micrograph, fabricated in CMOS 0.13 μm.

generator/analyzer is used for the data generation and BER measurement. Large on-chip delay lines (chains of flip-flops, ten per channel) provide all bus channels with pseudo-independent data. This setup allows for random-data BER testing in a realistic crosstalk environment while deterministic data patterns (for e.g. step response measurement) can also be applied. Different twisting patterns and receiver configurations are used for the different channels of the differential bus, as shown in figure 5.17.

Channels 1, 4, and 6 are equipped with 50 Ω output buffers and pads for measurements. The output buffers can accommodate a full-swing input range, with some large-signal compression. They attenuate the signal by about 6 dB (small-signal) to 9 dB (large-signal). Channel 4 is used for the BER measurements while the other two channels are used for e.g. crosstalk and eye-diagram measurements. The receiving ends of the single-ended bus interconnects are directly connected to pads to enable measurements directly on the interconnect. The chip has been measured in a probe station using 50 Ω GSSG probes for the high-speed signals. At the receiver side, dedicated GSSG pads are available for the various channels to enable wide-band measurements directly on the specific channel.

5.3.5 Measurement results

The measured interconnect parameters are 0.19 kΩ/mm and 0.25 pF/mm (for a differential interconnect). These values agree with the EM-field simulations, given the tolerance bounds of the process. Indirect measurement of the capacitance suggests that it is composed of roughly 0.05 pF/mm to each of the four sides; the part of the capacitance between the differential wires is doubled due to Miller multiplication.

The configurability of the transmitter and receiver enables measurements with or without PW pre-emphasis and with high-ohmic (conventional) or low-ohmic termination at the receiver. Eye-diagrams for each of the four settings are shown in figure 5.18, as measured at the output of channel 6.

118
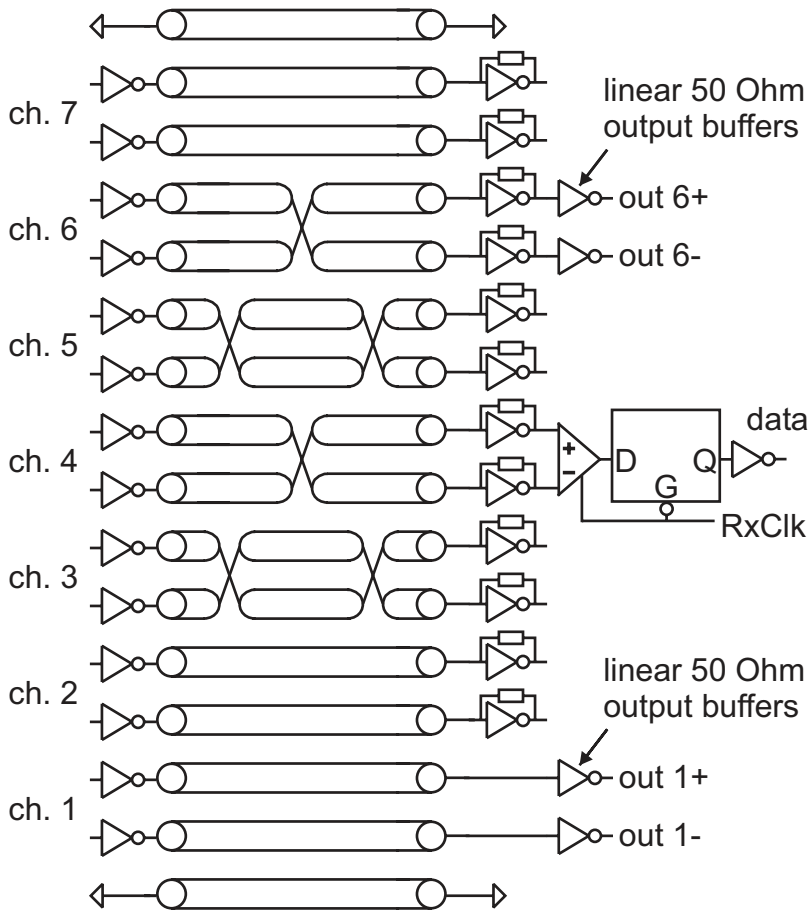
Figure 5.17: Configuration of the 7-bit differential bus.

BER measurements (with PRBS data patterns) for the four settings were carried out at channel 4 and table 5.4 shows the highest data rate at which bit-errors are not yet measurable (BER<10^{-12}). At the boundary of error-free operation the BER drops sharply, as the primary bit-error sources are deterministic (ISI) or static (offset) and a BER much lower than 10^{-12} is expected at the shown data rates.

The transceiver circuits of channel 4 have a dedicated supply to measure their energy consumption separately. The energy consumption of the channel is also shown in the table above (measured with PRBS data patterns with 50% data activity). Simulated values for the energy consumption of the various parts of the transceiver are also shown.

The results show good agreement with the analysis. The 550 Mb/s achieved in the conventional case is only slightly lower than the theoretical limit of 600 Mb/s. Resistive termination improves the achievable data rate by nearly a factor of three. The improvement of PW pre-emphasis together with conventional termination is a factor of four and another
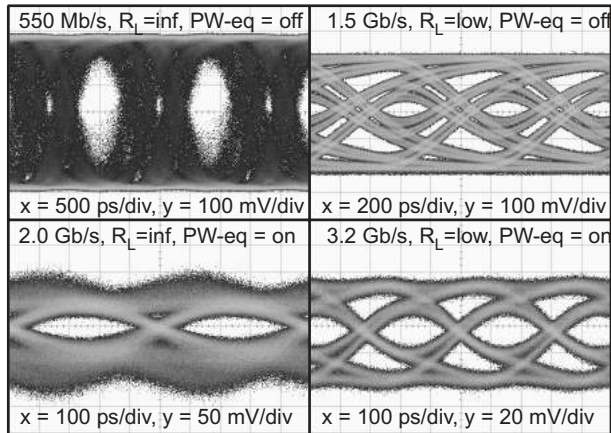
Figure 5.18: Eye-diagrams for four different schemes. The output buffers compress the vertical scale by 6 to 9 dB.

	High-ohmic termination		**Low-ohmic termination**	
Without PW pre-emphasis	Data rate:	550 Mb/s	Data rate:	1.5 Gb/s
	Energy consumption:	3.4 pJ/b (TX: 0.2, wire: 1.2, TIA: 0.7, SA: 1.4)	Energy consumption:	2.5 pJ/b (TX: 0.2, wire: 0.8, TIA: 0.9, SA: 0.5)
With PW pre-emphasis	Data rate:	2.0 Gb/s	Data rate:	3.0 Gb/s
	Energy consumption:	2.5 pJ/b (TX: 0.5, wire: 0.8, TIA: 0.7, SA: 0.5)	Energy consumption:	2.0 pJ/b (TX: 0.5, wire: 0.6, TIA: 0.5, SA: 0.3)

Table 5.4: Measurement results for different configurations.

factor of two if used in combination with resistive termination. So, the measurements prove that the achievable data rate is increased by using pulse-width pre-emphasis equalization and/or a low-ohmic load resistance.

We also did some measurements to test the robustness of the transceiver. The eye at the receiver is still open at 3.2 Gb/s as visible in the bottom right of figure 5.18, but the opening is so small (40 mV$_{pp}$) that effects such as hysteresis and offset in the clocked receiver prevent reliable detection at this data rate (BER 10^{-8}). At 3 Gb/s, error-free operation is possible for all ten measured samples (I_{bias} = 400 μA), but without much V_{DD} or I_{bias} tolerance. At 2.5 Gb/s, the design is robust and the BER remains immeasurable with large

external parameter deviations. Figure 5.19 illustrates this robustness by plotting the measured eye-width as a function of an external parameter, while keeping the other parameters at their nominal value (V_{DD} = 1.2 V, TX Clk duty cycle = 50% and I_{bias} = 200 μA).
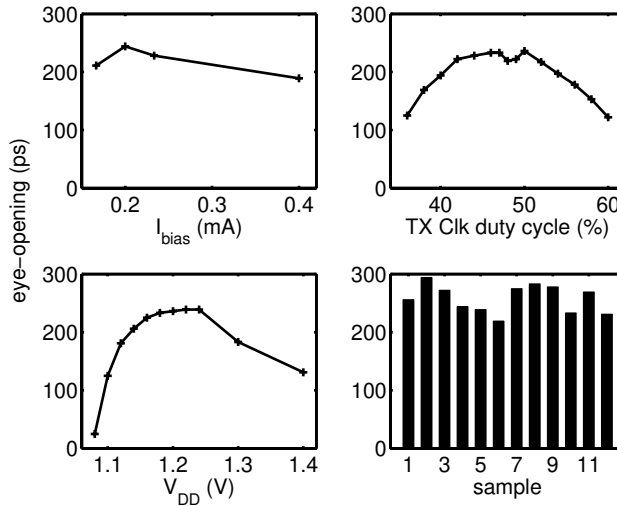
Figure 5.19: Measured eye-width versus parameters and over different samples at 2.5 Gb/s.

To measure the eye-width, a phase-shifter was used to vary the skew of the receiver clock and find the phase-shifts where the BER just becomes measurable. The optimal bias current of 200 μA gives a PW pre-emphasis duty cycle of about 58%. The measured relationship between the external TX Clk duty cycle and the eye-width behaves as expected, except for a small drop in eye-width around 50% TX Clk duty cycle which can probably be attributed to measurement tolerances. The highest measured eye-width of 250 ps is lower than the theoretical value of almost 400 ps due to the required setup and hold times of the sense amplifier.

5.4 Crosstalk reduction with twists

5.4.1 Introduction

The previous section showed a transceiver that achieves high data rates. However, in order not to be limited in data rate by crosstalk, we use twists in the interconnects (see section 4.3). In order to prove the effectiveness of these twists, some more measurements were done on the chip of figure 5.16 [3, 53].

5.4.2 Measurement setup

In order to measure crosstalk transfer functions, we use the same measurement setup as in the previous section (see figure 5.17). Since we only have one data generator available, the transmitters are all driven by the same data. In order to create pseudo-independent data on

the seven channels (needed for data rate measurements on channel 4), there is a delay of 10 clock periods between every transmitter, realized via on-chip shift registers. Because of the pulse-width equalization in the transmitters, the data is multiplied with a rectangular wave with controllable pulse-width. By setting the pulse-width to 50%, the transmitters transmit a square wave. For a 'zero', the square wave is first half a clock period low and after that half a clock period high; for a 'one' the square wave is inverted (Manchester code).

To understand how crosstalk information is extracted, assume that the data generator has been generating a 'zero' for longer than 70 clock delays. Then, all seven transmitters transmit the same square wave. If the data generator then starts transmitting a 'one', first the square wave of channel 1 is inverted. 10 clock delays later, also the square wave of channel 2 is inverted. Another 10 clock delays and the square wave of channel 3 is inverted, and so on. The square wave on the channels is filtered by the interconnect. Figure 5.20 shows the result as measured on channel 6 (out6+ – out6–).
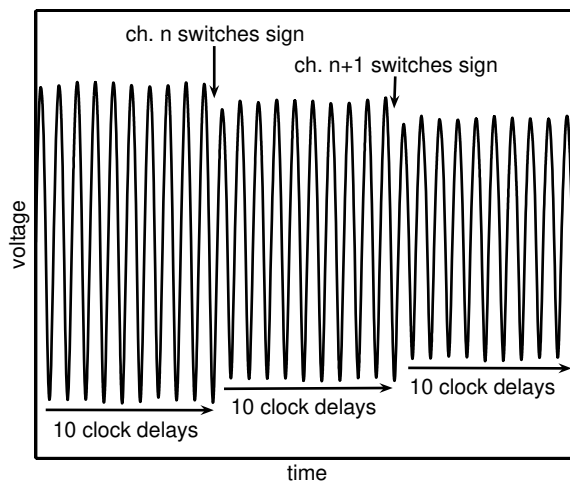
Figure 5.20: Output signal of channel 6 during crosstalk measurements.

The figure shows that the amplitude (and phase) of the sine wave is changing every 10 clock delays. This is because the amplitude and phase depend on the total resistance and capacitance of the interconnect. The capacitance of the interconnect also depends on the signals on the other channels (Miller Multiplication) and every 10 clock delays, another channel changes sign. By carefully correlating the output voltage with the clock frequency and filtering the results, the amplitude and phase steps are found. These steps are a measure for the crosstalk transfer functions at the clock frequency. The crosstalk transfer functions are found by repeating these measurements over a range of clock frequencies.

5.4.3 Measurement results

Figure 5.21 shows the measured transfer function from channel 6 and the crosstalk transfer functions from channel 5 and 7 to channel 6 (low-ohmic termination). As expected, the crosstalk from channel 5 is less than the crosstalk from channel 7: the double twist in channel 5 at $z_1 = 0.3$ and $z_3 = 0.7$ reduces CM crosstalk (see top of figure). Both the crosstalk from channel 5 and channel 7 is reduced for the differential output: the single twist in channel 6 at $z_2 = 0.5$ reduces DM crosstalk (see bottom of figure).
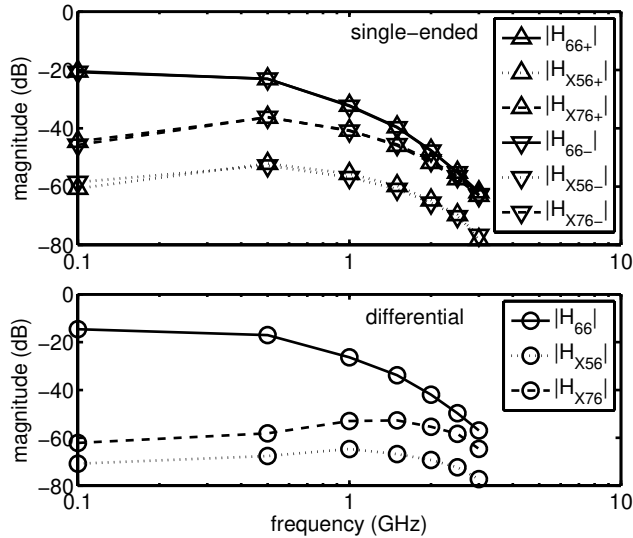
Figure 5.21: Measured transfer function at channel 6 and crosstalk transfer functions from channels 5 and 7 to channel 6.

We also measured the transfer function of channel 1 and the crosstalk transfer function from channel 2 to channel 1 (see figure 5.22). These transfer functions have a smaller bandwidth due to the high-ohmic termination of channel 1. There is more crosstalk from channel 2 on out1+ than on out1−, because out1− has no signal carrying neighbor. The bottom graph shows that the crosstalk is not reduced for the differential output as there is no twist in channel 1.

The influence of crosstalk on the eye-diagram is shown in figure 5.23. The figure shows both the output of the single-ended (SE) halves and the differential output of channel 6 at a rate of 2.5 Gb/s. Each SE half of channel 6 receives crosstalk mainly from the wire-piece that runs alongside channel 7 (but the other channels in the bus and the perpendicular bus also generate some common-mode crosstalk). The eye-closure due to the crosstalk in the single-ended output is clearly visible in the figure, while the crosstalk is mitigated in the differential output. If the twist in channel 6 would not be present, then the crosstalk on both SE halves would be even higher and it would not be canceled in the differential output.
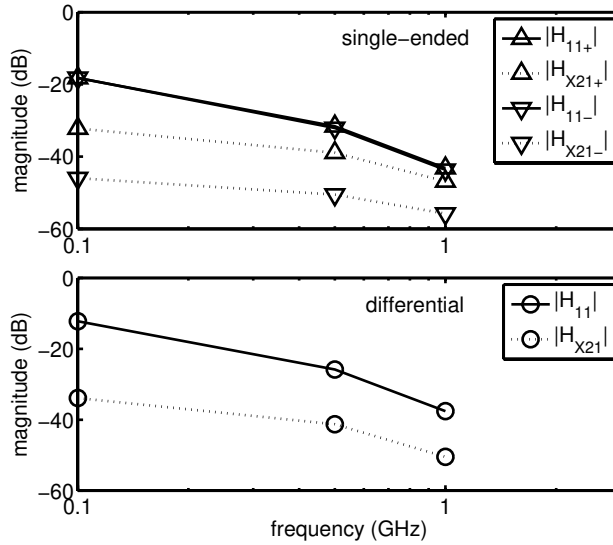
Figure 5.22: Measured transfer function at channel 1 and crosstalk transfer function from channel 2 to channel 1.
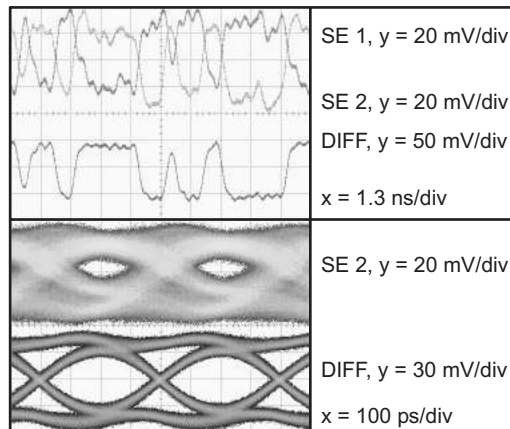
Figure 5.23: Effect of crosstalk on the single-ended and twisted differential output of channel 6 at 2.5 Gb/s. On-chip signals are about 6 dB larger.

5.5 High speed and low power transceiver

5.5.1 Introduction

The transceiver of section 5.3, in combination with the twisting scheme of section 5.4, achieves a high data rate. However, the power consumption is quite high, especially for low data activity. The reason for this is that pulse-width equalization switches every symbol period, thus consuming power, and the low-ohmic load resistance has a static current. In

order to achieve similar high data rates, but with both low power consumption and a power consumption that depends on data activity (low static power consumption), we use the capacitive pre-emphasis transmitter of section 3.3.6 [54, 55]. For the receiver, we have implemented a power efficient sense amplifier with the decision feedback equalization of section 3.7.3 [55, 56].

5.5.2 Transmitter circuits

The capacitive pre-emphasis transmitter of section 3.3.6 was implemented in the CMOS 90 nm process. The circuit implementation is given in figure 5.24.
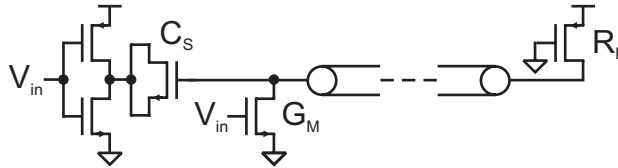
Figure 5.24: Circuit implementation of a capacitive pre-emphasis transmitter.

G_M (5.2 µS) and R_L (16 kΩ) are implemented with MOS transistors with W/L = 0.17 µm / 3.9 µm and W/L = 0.765 µm / 0.71 µm respectively. For C_S, the gate capacitance of an NMOS transistor is used. As the gate oxide thickness is much smaller than the oxide between interconnects, the area that is consumed by C_S (270 fF) is relatively small (W/L = 6.02 µm / 6.02 µm). The gate-source and gate-drain capacitance of NMOS transistors is largest for a large gate-source voltage. Therefore, we choose the low-swing signals close to V_{DD} (between 1.1 V and 1.2 V). With reference to figure 3.13, I_{DC} = 0 and V_{bias} = V_{DD} = 1.2 V. This makes the bias voltage of the interconnect V_{DC} = V_{DD} − ½·G_M·V_{DD}·R_L = 1.2 − 0.05 = 1.15 V. Of course, the capacitance C_S can also be made with a PMOS transistor. Then, the gate should be connected to the inverter and source and drain to the interconnect.

The inverter in front of C_S is implemented with a chain of three inverters that are dimensioned for minimum total delay (see section 5.3).

Inverter	PMOS	NMOS
	W/L [µm/µm]	W/L [µm/µm]
G₁	0.72/0.1	0.24/0.1
G₂	2.4/0.1	0.96/0.1
G₃	9.6/0.1	3.84/0.1

Table 5.5: Dimensions for the chain of inverters.

The total area of the differential transmitter is about 200 μm^2.

5.5.3 Receiver circuits

In the CMOS 90 nm process, we have implemented a sense amplifier with feedback equalization, as shown in figure 5.25.
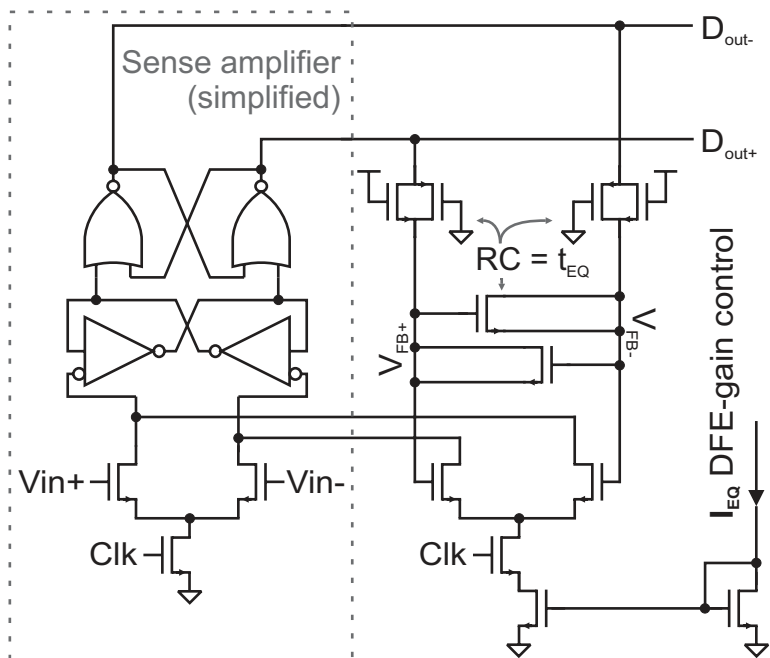
Figure 5.25: Simplified sense amplifier with analog DFE feedback filter.

Details about the sense amplifier itself (shown is a simplified schematic) and the dimensions of the transistors can be found in [42, 56]. The gain of the DFE is controlled with the help of a current I_{EQ}, which sets the gain of the feedback filter. The feedback filter is implemented in an analog way and consumes only 0.02 pJ/b and occupies 30 μm^2. The power consumption of the total sense amplifier is 0.12 pJ/b with an area of about 100 μm^2.

5.5.4 Measurement setup

The chip micrograph of the demonstrator IC is shown in figure 5.26. The chip is used to prove that with a capacitive pre-emphasis transmitter and decision feedback equalization at the receiver, similar high data rates can be achieved as with the transceiver of section 5.3, but with much lower power consumption.

The 10 mm long interconnects are placed in metal 4 (7 metal layer process). The other metal layers are filled with GND- and VDD-connected metal stripes to emulate high-density metal use. An external pattern generator/analyzer is used for data generation and BER measurement. The receiver clock is generated externally in order to adapt its phase to the eye position and be able to measure eye widths. In an application a simple skew circuit or a source-synchronous approach could be used to generate the proper clock phase. Eye-diagrams are measured via 50 Ω output buffers that are connected to the output of a differential interconnect.
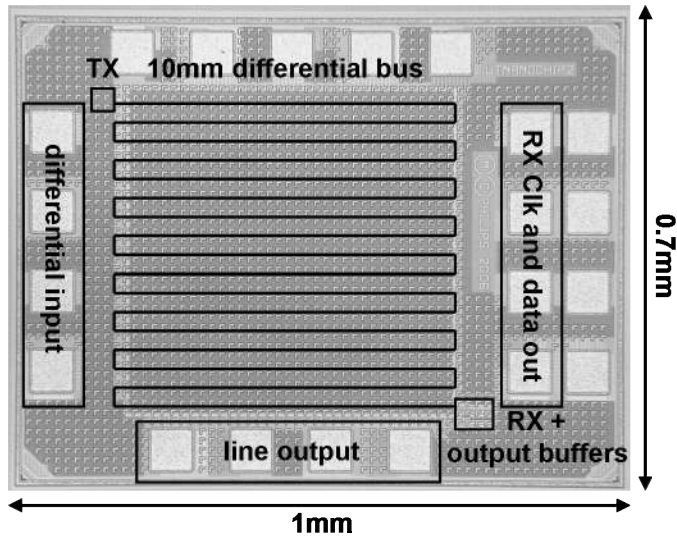
Figure 5.26: Chip micrograph, fabricated in CMOS 90 nm.

5.5.5 Measurement results

The measured interconnect parameters are 0.20 kΩ/mm and 0.28 pF/mm (for a differential interconnect). The capacitance is as expected, but the resistance is higher than expected from the design manual of the process. These parameters were not measured directly, but deduced from the measurement of step responses. Unfortunately, these step responses also show a mismatch between the time constants C_S/G_M and $C_T \cdot R_L$. By comparing simulation results and measurements, the following parameters are found: C_S = 290 fF (design: 270 fF), G_M = 4.3 μS (design: 5.2 μS), C_T = 2.8 pF (design: 2.8 pF) and R_L = 16 kΩ (design: 16 kΩ). These measured parameters do agree with a newer version of the design manual with updated transistor parameters, which was only available after the measurements.

So, due to an outdated version of the design manual, the circuits were not designed optimally. This is shown in figure 5.27 that shows the transfer function with the parameters of the design (outdated) and the measurement parameters (and new parameter set). The figure shows the mismatch in the time constants. This has a negative influence on the absolute eye-height, as shown in figure 5.28. The figure shows that the eye-height is reduced because of the mismatch in time constants. Therefore, lower data rates will be achievable. However, even with these mismatches the achievable data rate is higher than with conventional techniques, as will be shown next. This proves the robustness of the technique.

Figure 5.29 shows a measured eye-diagram at a data rate of 1 Gb/s. The measured BER at the edges of the eye is also shown. The BER drops rapidly below a clock skew of -150 ps and above 180 ps, giving an eye-opening of 670 ps. Data rates up to 1.35 Gb/s are achieved without decision feedback equalization (DFE) at the receiver side (I_{EQ}=0). Data rates up to 2 Gb/s are measured with DFE. Figure 5.30 shows that DFE improves the eye-opening for a wide range of I_{EQ}. In an application I_{EQ} can therefore be fixed at design time.

127
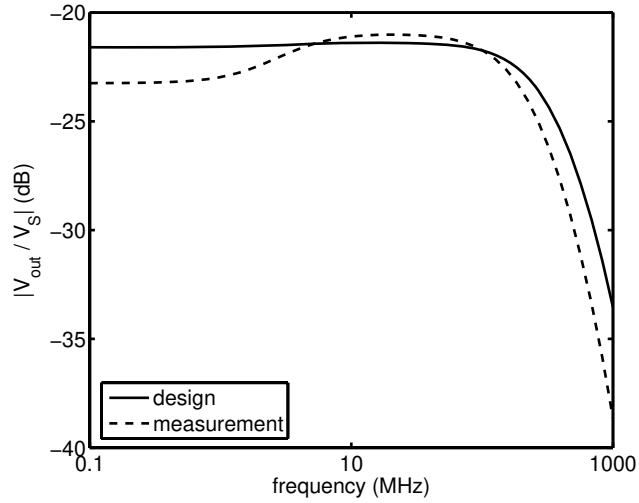
Figure 5.27: Transfer function of 10 mm interconnect on CMOS 90 nm chip with parameters during design process and actual parameters.
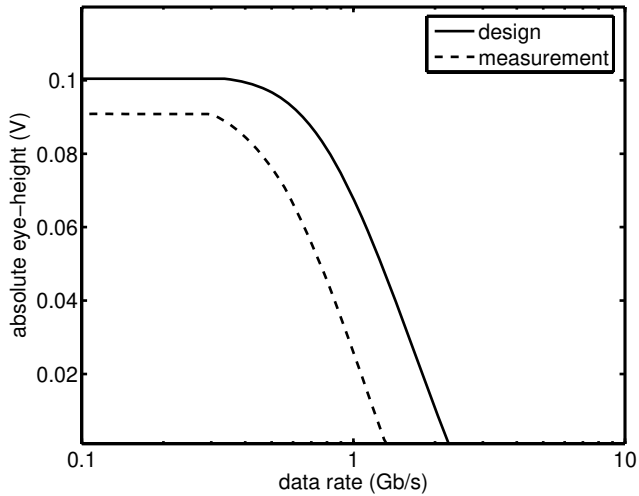
Figure 5.28: Absolute eye-height of 10 mm interconnect on CMOS 90 nm chip with parameters during design process and actual parameters.
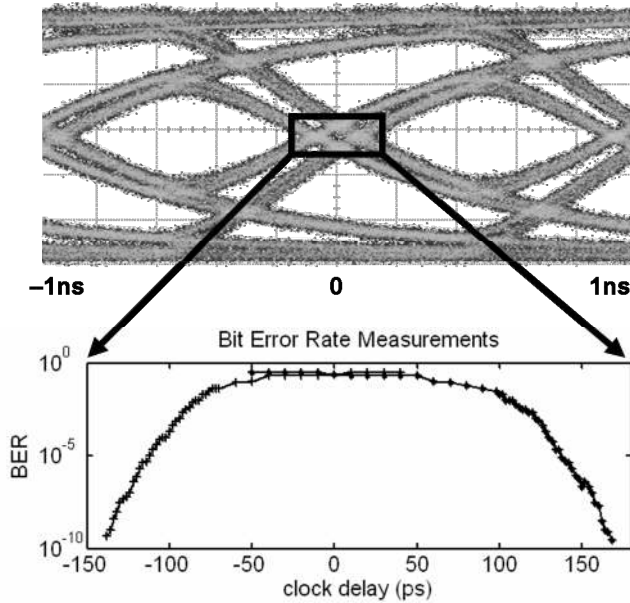
Figure 5.29: Eye-diagram at 1Gb/s and BER at the edges of the eye.
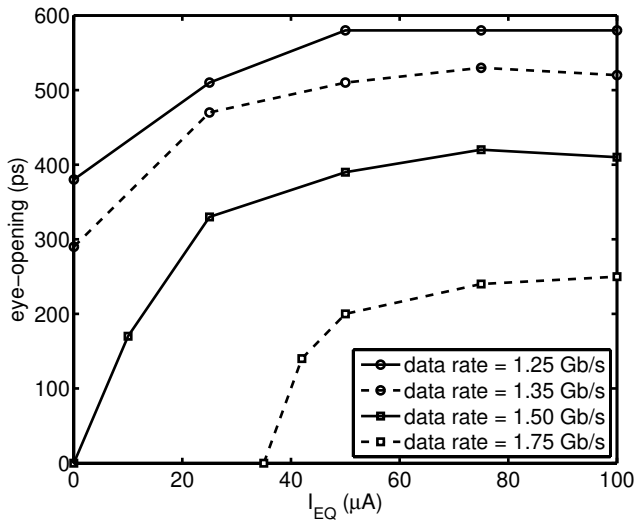
Figure 5.30: Eye-opening for different data rates as a function of I_{EQ}.

So, the measurements show that with the capacitive pre-emphasis transmitter and DFE at the receiver, the achievable data rate is increased. However, for this chip we do not only want a high data rate, but also low power consumption. In figure 5.31 the measured energy per bit is plotted as a function of transition probability at different data rates.
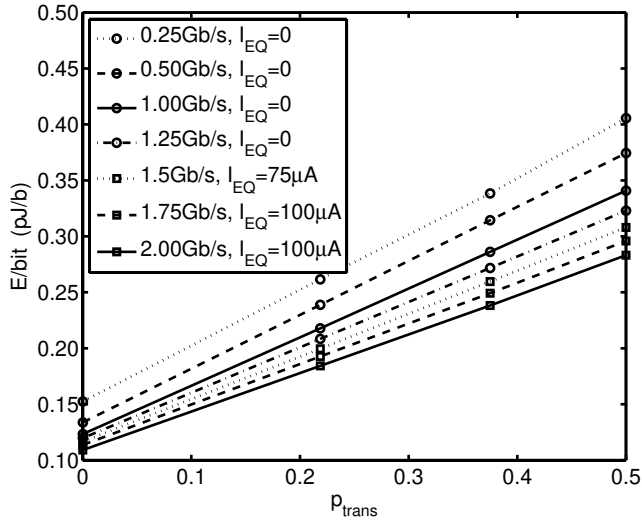
Figure 5.31: Energy per bit as a function of transition probability for different data rates.

With random data at 2 Gb/s, only 0.28 pJ/b is dissipated. This is seven times lower than the lowest power consumption of the transceiver of section 5.3. The power dissipation of 0.12 pJ/b at zero data activity is mainly due to the power dissipation in the sense amplifier, which has large transistors to get a low offset (σ_{os}=8 mV). Clock-gating can be used to eliminate power consumption during inactive periods. The DFE part of the circuit requires less than 7% of the total transceiver power, while it can increase the achievable data rate with a factor 1.5.

Also for this chip, measurements were done to prove the robustness of the transceiver. The one-sigma offset of the total transceiver is 11mV, measured over 20 samples. Due to this offset, not all samples achieve 1.35 Gb/s (without DFE), but a slightly lower data rate of 1 Gb/s is achieved by all samples. Simulations over process corners also indicate that the circuit is robust for PVT (process, voltage and temperature) variations at a rate slightly lower than the maximum achievable data rate.

Measurements for different supply voltages and different samples are given in figure 5.32. The eye stays open for about 20% variation in V_{DD} and the power scales linearly with V_{DD}. All samples have an open eye at 1 Gb/s with comparable power consumption. The eye-opening of some samples is smaller due to higher offsets in the circuits.

5.6 Comparison

In this section, we compare the measurement results of our two transceivers with other published global interconnect transceivers. Table 5.6 shows the used technology, interconnect length, interconnect dimensions, achievable data rate and energy consumption per transmitted bit.
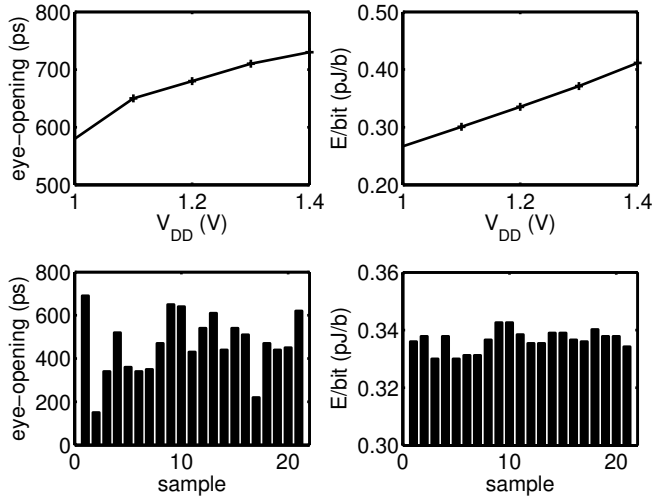
Figure 5.32: Measurements of eye-opening and power consumption as a function of V_{DD} and for different samples.

Ref.	[17]	[21]	[19]	[27]	[22]	[16]
CMOS Technology node (µm)	0.18	0.18	0.18	0.18	0.18	0.18
Length (mm)	2	20	7.2	14	3	5
Width (µm)	2·4	2·16	2·9	2·8	2·(4+4)	2+0.8
Spacing (µm)	2·4.2	2·1[1]	2·1[1]	2·8	2·4	2·1.2
Height (µm)	0.9	2	2.34	0.5[1]	0.53	1.04
Thickness (µm)	7.35	1.9	1.5	0.5[1]	0.5[1]	3.52
Cross-sectional area (µm²)	135	133	76.8	32.0	24.7	23.7
Achievable data rate (Gb/s)	4	1	14	3	8	3
Energy per bit (pJ/b)		16.1		2	0.29	

[1] Estimated values.

Table 5.6-I: Dimensions and performance of different on-chip transceivers.

Ref.	[13]	[15]	[9]	[14]	[6] (this work)	[55] (this work)
CMOS Technology node (µm)	0.18	0.35	0.13	0.13	0.13	0.09
Length (mm)	10	17.5	8	10	10	10
Width (µm)	4.5	2	0.6	0.60	2·0.4	2·0.54
Spacing (µm)	1[1]	1[1]	0.6	0.63	2·0.4	2·0.32
Height (µm)	0.5[1]	0.5[1]	0.6	0.35	0.35	0.33
Thickness (µm)	0.5[1]	0.5[1]	0.3	0.36	0.46	0.27
Cross-sectional area (µm²)	5.5	3	1.08	0.87	1.30	1.03
Achievable data rate (Gb/s)	2	1	1.5	0.2	3	2
Energy per bit (pJ/b)	2.3	5.8		1.7	2.0	0.28

[1] Estimated values.

Table 5.7-II: Dimensions and performance of different on-chip transceivers.

In the table, a width of for instance 2·4 µm means that a differential interconnect is used with both single-ended halves having a width of 4 µm. A width of for instance 2+0.8 µm means that there is also a ground line or shield of 0.8 µm in width next to every interconnect. References [6] and [55] give the results of the two chips as described in sections 5.3 and 5.5 respectively.

As the table shows, all implementations use different dimensions and therefore have different cross-sectional areas. Also, the length of the interconnect can differ. In order to have a fair comparison, we will specify the achievable data rate divided by cross-sectional area. This number is then multiplied by the interconnect length squared, as the bandwidth of the interconnect depends on the length squared. Furthermore, we also specify the energy consumption per transmitted bit divided by the interconnect length. The longer the length, the higher the capacitance of the interconnect and hence the higher the power consumption will be. Note, that the results are not scaled for technology. Furthermore, the power consumption is measured at 50% data activity. Some of the solutions have high static power consumption and will relatively perform worse at lower data activities.

Figure 5.33 plots the results of table 5.8 with power on the y-axis and speed on the x-axis. For high speed and low power consumption, a transceiver should be located in the bottom right corner of the figure.

132

Ref.	Speed $\mathbf{Gb \cdot mm^2 / s \cdot \mu m^2}$	Power $\mathbf{pJ / b \cdot mm}$
[17]	0.12	
[21]	3.0	0.81
[19]	9.4	
[27]	18	0.14
[22]	2.9	0.097
[16]	3.2	
[13]	36	0.23
[15]	$1.0 \cdot 10^2$	0.33
[9]	88	
[14]	23	0.17
[6] *(this work)*	$2.3 \cdot 10^2$	0.20
[55] *(this work)*	$1.9 \cdot 10^2$	0.028

Table 5.8: Performance comparison between different on-chip transceiver solutions.
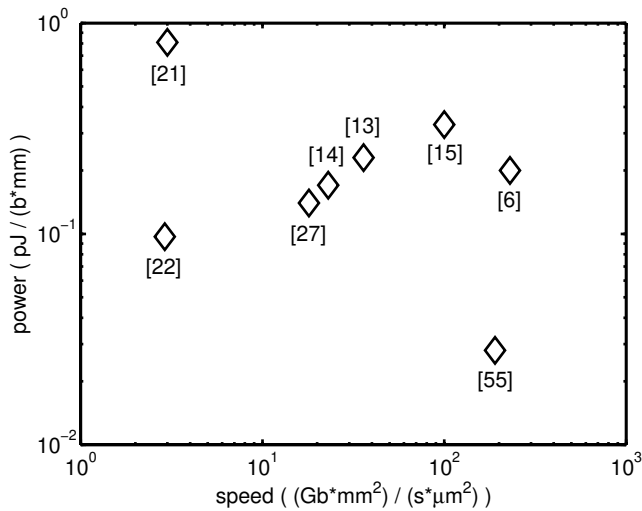
Figure 5.33: Comparison of the transceivers described in this chapter ([6] and [55]) with on-chip interconnect transceivers as found in literature.

The figure shows that both transceivers, as described in this chapter, have a considerably higher achievable data rate rate at a certain length per cross-sectional area compared to the

other solutions. These solutions have either a much higher cross-sectional area or have a lower data rate.

The transceiver with pulse-width equalization and low-ohmic load resistance [6] has high static power consumption. Nonetheless, at 50% data activity the power consumption is comparable to other solutions. The transceiver with the capacitive pre-emphasis transmitter and decision feedback equalization [55] has low static power consumption and has considerable less energy consumption per transmitted bit than all other solutions.

5.7 Summary

−   Two chips, one in CMOS 0.13 μm and another in CMOS 90 nm, were implemented to test the concepts of the previous chapters. With these concepts, a high data rate is achieved over RC bandwidth-limited interconnects.
−   The 10 mm long interconnects have a measured total resistance of 1.9 kΩ (CMOS 0.13 μm) or 2.0 kΩ (CMOS 90 nm) and a measured total capacitance (for differential interconnects) of 2.5 pF (CMOS 0.13 μm) and 2.8 pF (CMOS 90 nm). Due to these high resistances and capacitances, the RC bandwidth of the interconnects is severely limited (< 100 MHz).
−   A conventional transceiver (inverter as both transmitter and receiver) in 0.13 μm CMOS achieves 550 Mb/s/ch over a 10 mm long differential interconnect. Power consumption is 3.4 pJ/b at 50% data activity.
−   Another transceiver in 0.13 μm CMOS for 10 mm long differential on-chip interconnects achieves 3 Gb/s/ch. The transceiver uses pulse-width pre-emphasis at the transmitter and a low-ohmic load resistance at the receiver. Power consumption is 2 pJ/b at 50% data activity.
−   For this transceiver, a single twist in the even and a double twist in the uneven differential interconnects is effective against crosstalk.
−   A transceiver in 90m CMOS for 10 mm long differential on-chip interconnects achieves 2 Gb/s/ch. The transceiver uses a capacitive pre-emphasis transmitter and decision feedback equalization at the receiver. Power consumption is only 0.28 pJ/b at 50% data activity with low static power consumption.
−   Compared to other solutions, both measured chips have a higher achievable data rate per cross-sectional area times interconnect length squared ($2 - 10^3$ x). The second chip also has lower energy consumption per bit per interconnect length (3 – 30 x).

Chapter 6

Conclusions

6.1 Central question

In the introduction, global interconnects are defined as interconnects that span a large portion of a chip and do not scale in length with technology. The predicted total resistance of these global interconnects increases with technology, while the capacitance stays more or less equal. Therefore, the RC bandwidth will decrease and the achievable data rate will be limited.

This thesis describes methods to increase the achievable data rate. However, in circuits proposed in literature there often is a trade-off between achievable data rate on the one side and power, area and data integrity on the other side. Therefore, the central question of this thesis is:

How can we design global interconnects and transmitter and receiver electronics in future IC technologies to

- *maximize data capacity,*
- *minimize chip area consumption,*
- *minimize power consumption and*
- *maintaining data integrity?*

In order to answer this question, first a summary will be given of the conclusions of this thesis. After that, the solutions as presented in this thesis are discussed.

137

6.2 Summary

- In order to calculate transfer functions, eye-opening, latency and power consumption, we model the interconnects with distributed parameters. The value of these distributed parameters can be found by mapping the results of a 3D EM-field simulator on the model.
- Global interconnects are dominated by their distributed resistance and capacitance and have dominant first-order RC behavior. Due to a large distributed resistance and capacitance, the bandwidth is severely limited and is in the order of 100 MHz.
- For highest bandwidth per area, all interconnect dimensions (width, spacing, height and oxide thickness) should be chosen equal. If second order effects are included like fringe capacitances and miller multiplication due to differential signaling, the optimal width increases somewhat. However, the optimum is shallow.
- The bandwidth can be increased by about a factor of three by using a capacitive source impedance. In addition, the capacitive source impedance decreases power consumption in the interconnect. A small transconductance and large load resistance can be used to define the DC potential without increasing the static power consumption much.
- The bandwidth can also be increased by about a factor of three by a small resistive load impedance. Also now, the dynamic power consumption decreases. However, the static power consumption can be high. In order to be less susceptible to offset, a transimpedance amplifier can be used to create the low-ohmic load impedance.
- The achievable data rate can be further increased by using equalization techniques, including pulse-width pre-emphasis equalization at the transmitter and decision feedback equalization at the receiver.
- In order to maintain data integrity, a low offset sense amplifier is used to restore the low voltage swing at the output of the interconnect to full-swing. Furthermore, differential interconnects are used: noise sources like supply and substrate noise become common-mode and can be canceled by the receiver. Also, crosstalk from orthogonal aggressors in another metal layer than the victim interconnect is suppressed with differential interconnects.
- Crosstalk from neighboring interconnects in a bus is canceled with a single twist in every even interconnect for differential crosstalk and a double twist in every odd interconnect for common-mode crosstalk. The optimal positions of the twists depend on the termination impedances.
- A conventional transceiver (inverter as both transmitter and receiver) in 0.13 μm CMOS achieves 550 Mb/s/ch over a 10 mm long uninterrupted differential interconnect. Power consumption is 3.4 pJ/b at 50% data activity.
- Another transceiver in 0.13 μm CMOS for 10 mm on-chip interconnect is designed to increase the achievable data rate. The transceiver achieves 3 Gb/s/ch by using pulse-width pre-emphasis at the transmitter and a low-ohmic load resistance at the receiver. Power consumption is 2 pJ/b at 50% data activity.
- A transceiver in 90m CMOS for 10 mm on-chip interconnect is designed to both increase the achievable data rate and decrease the power consumption. The transceiver achieves 2 Gb/s/ch by using a capacitive pre-emphasis transmitter and decision feedback equalization at the receiver. Power consumption is only 0.28 pJ/b at 50% data activity with low static power consumption.

6.3 Discussion

6.3.1 Presented solutions

Four different techniques have been proposed to increase the data capacity or achievable data rate of global interconnects. Two techniques use one of the termination impedances of the interconnect to enlarge the bandwidth. Placing a capacitive source impedance at the transmitter or a low-ohmic load impedance at the receiver both increase the bandwidth by about a factor of three. The other two techniques perform equalization by altering the transmitted or received symbol shape. At the transmitter, pulse-width pre-emphasis equalization divides the symbol time into two periods. In the first period the interconnect is charged, while the second period is used to discharge the interconnect again to remove inter-symbol interference. At the receiver, decision feedback equalization cancels the long tail by subtracting a filtered version of the previous decision of a comparator from the output voltage of the interconnect. Combinations of these four techniques are also possible.

All techniques increase the achievable data rate. However, the central question of this thesis states that we do not only want a high achievable data rate, but also as low area and power consumption as possible, while maintaining data integrity.

The area consumption of all techniques is similar. All transmitters and receivers have an area in the order of hundred square micrometer. Compared to the interconnect area, which is in the order of ten thousand squared micrometer, this is small. Note that the area of the interconnect itself is kept as small as possible by optimizing the dimensions of the interconnect for highest bandwidth per cross-sectional area.

While the area consumption of the four techniques is similar, the power consumption is not. The power consumption of a scheme with a capacitive source impedance is low due to a low-voltage swing on the interconnect. The resistive load impedance also causes a small voltage swing, but only at the receiver side. Closer to the transmitter, the voltage swing is larger. Therefore, although this scheme has a smaller dynamic power consumption than a conventional scheme, it is larger compared to the scheme with the capacitive source impedance. Also, the static power consumption due to the resistive load impedance is high and especially for low data activities cancels the benefit in dynamic power consumption.

The power consumption of pulse-width pre-emphasis equalization is also relatively high for low data activities. Every symbol period, first a positive and then a negative pulse is transmitted. Thus, the transmitter consumes energy every symbol period, even if there is no data transition. The decision feedback equalization, on the other hand, does not cost much power. It only increases the total power consumption by a few percent. A benefit of this technique is that it does not require extra charging or discharging of the large interconnect capacitance.

In summary, low power transceivers can use a capacitive source impedance at the transmitter side and decision feedback equalization at the receiver side. Note that at the receiver a sense amplifier [56] is used to restore the low-swing signals back to full-swing. This sense amplifier will also cost power. However, this power can be much smaller than the decrease in power consumption due to the low voltage swing. This is shown in the

transceiver of section 5.5, where at a data rate of 2 Gb/s the power consumption is only 0.28 pJ/b (57% in transmitter and interconnect, 43% in sense amplifier). Note, that the full-swing transceiver of section 5.3 consumes 3.4 pJ/b.

For maintaining data integrity, both designed transceivers used the same techniques. First, low offset sense amplifiers are used to restore the low-swing output of the interconnects to full-swing. Second, differential interconnects are used. In this way, most noise sources become common-mode. The sense amplifiers that are used have good common-mode rejection and are designed to work for a large common-mode range. In order to cancel crosstalk from neighboring interconnects, that run in parallel in the same metal layer, the interconnects are twisted. Every even differential interconnect has one twist to cancel differential mode crosstalk and every odd differential interconnect has two twists to also cancel common-mode crosstalk. The optimal positions of the twists depend on the termination impedances of the interconnects. If the interconnects are terminated with low-ohmic impedances at both sides, the optimal position of the single twist is at fifty percent and the optimal positions of the double twist are at thirty and seventy percent. However, if the interconnects have a low-ohmic source impedance, but a high-ohmic load impedance, the optimal position of the single twist shifts to seventy percent and the optimal positions of the double twist to fifty and eighty-seven percent.

6.3.2 Future

How will the presented transceivers perform for future CMOS technologies? As explained in the introduction chapter of this thesis, the resistance of global on-chip interconnects is predicted to increase for future technologies. The capacitance will stay roughly equal. The result of this is that the RC bandwidth will decrease. If the architecture of current large CMOS chips is not changed, techniques as described in this thesis will be needed to increase the achievable data rate.

When the resistance of the interconnects increases, the low-ohmic load resistance will become a more attractive solution. The load resistance has to be small with respect to the resistance of the interconnect and if this last resistance increases, also the load resistance can increase. This will cost less power for creating a low impedance and also the static power consumption will becomes less.

The capacitance of the capacitive pre-emphasis transmitter is made with an NMOS or PMOS transistor in order to benefit from the thin gate oxide of these transistors. In this way, the area of this capacitance can be kept small. The capacitance of interconnects in future technologies stays roughly equal. Therefore, for the same voltage swing, the source capacitance should also stay equal over technology. In the worst case, if the gate oxide does not decrease much further and no high-k dielectric is used, the area consumption of the source capacitance does not scale. However, since the area consumption of the presented transceiver is still much smaller than the area of the interconnect, this should not be a problem.

The capacitive pre-emphasis transmitter can also be used for shorter interconnects. There is a tendency to move computer architecture in the direction of locally connected, reconfigurable hardware meshes that merge processing and memory. The processing

elements are connected by a communication network. The length of interconnects in these networks on chips (NoC) are often chosen to be a few millimeters. By using a capacitive pre-emphasis transmitter for these shorter interconnects, the power consumption is decreased. Furthermore, the transmitter will increase the bandwidth of the interconnects. This allows to use very small driver inverters, which further reduce area and power consumption.

The capacitive pre-emphasis transmitter also opens up other opportunities. By using an extra source capacitance that is driven by a delayed version of the input signal, a FIR (finite impulse response) filter is created at the transmitter [54], which enables additional equalization. Another application of the capacitive pre-emphasis transmitter can be the implementation of a multi-drop bus, where several transmitters and receivers share the same global interconnect.

6.3.3 Comparison

The transceivers as presented in this thesis are compared to other solutions as found in literature in section 5.6. Two metrics are used: the achievable data rate per cross-sectional area and the power per transmitted bit.

Both transceivers of this thesis outperform the solutions in literature with respect to the first metric: we achieve higher data rates per cross-sectional area. The main reason for this is that most other solutions try to create LC behavior by using very wide and thick interconnects. Therefore, large cross-sectional area is used. Our transceivers only have two or three times minimum width and do not use the thick top level metal layers. These thin and narrow interconnect do have a small bandwidth, but by using the proposed termination and equalization techniques, still high data rates are possible.

With respect to power, our solution with pulse-width pre-emphasis and low-ohmic load resistance has approximately the same power consumption as other solutions. Unfortunately, the solution has high static power consumption, thus power is not decreased for lower data activities. Our solution with a capacitive source impedance and decision feedback equalization has a much lower power consumption than all other solutions. The capacitive source impedance does not only increase bandwidth, but also decreases the voltage swing on the entire interconnect, thus reducing power consumption.

References

[1]   R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," *Proceedings of the IEEE*, vol. 89, pp. 490-504, April 2001.

[2]   E. A. M. Klumperink, R. Kreienkamp, T. Ellermeyer, and U. Langmann, "Transmission Lines in CMOS: An Explorative Study," *Circuits, Systems and Signal Processing (ProRISC), Annual Workshop on*, Nov. 2001.

[3]   E. Mensink, D. Schinkel, E. Klumperink, E. van Tuijl, and B. Nauta, "Optimally-placed twists in global on-chip differential interconnects," *European Solid-State Circuits Conf. (ESSCIRC), Proc. of the*, pp. 475-478, Sept. 2005.

[4]   W. J. Dally and J. W. Poulton, *Digital Systems Engineering*: Cambridge University Press, 1998.

[5]   H. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*: Reading, MA: Addison-Wesley, 1990.

[6]   D. Schinkel, E. Mensink, E. A. M. Klumperink, E. van Tuijl, and B. Nauta, "A 3-Gb/s/ch transceiver for 10-mm uninterrupted RC-limited global on-chip interconnects," *Solid-State Circuits, IEEE Journal of*, vol. 41, pp. 297-306, Jan. 2006.

[7]   P. Saxena, N. Menezes, P. Cocchini, and D. A. Kirkpatrick, "Repeater scaling and its impact on CAD," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 23, pp. 451-463, April 2004.

[8]   N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-Power Dissipation in a Microprocessor," *System-Level Interconnect Prediction (SLIP), Proc. of the Int. Workshop on* pp. 7-13, Feb. 2004.

[9]   H. Kaul and D. Sylvester, "Low-power on-chip communication based on transition-aware global signaling (TAGS)," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, pp. 464-476, May 2004.

[10]   H. Hong-Yi and C. Shih-Lun, "Interconnect accelerating techniques for sub-100-nm gigascale systems," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, pp. 1192-1200, Nov. 2004.

[11]   A. Nalamalpu, S. Srinivasan, and W. P. Burleson, "Boosters for driving long onchip interconnects - design issues, interconnect synthesis, and comparison with repeaters," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 21, pp. 50-62, Jan. 2002.

[12]   E. Seevinck, P. J. van Beers, and H. Ontrop, "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's," *Solid-State Circuits, IEEE Journal of*, vol. 26, pp. 525-536, April 1991.

[13]   L. Zhang, J. Wilson, R. Bashirullah, L. Lei, X. Jian, and P. Franzon, "Driver pre-emphasis techniques for on-chip global buses," *Int. Symp. on Low Power Electronics and Design (ISLPED), Proc. of the*, pp. 186-191, Aug. 2005.

[14]   A. Katoch, H. Veendrick, and E. Seevinck, "High speed current-mode signaling circuits for on-chip interconnects," *Circuits and Systems (ISCAS), proc. of the IEEE Intern. Symp. on*, pp. 4138-4141, May 2005.

[15]   R. Bashirullah, L. Wentai, R. Cavin, III, and D. Edwards, "A 16 Gb/s adaptive bandwidth on-chip bus based on hybrid current/voltage mode signaling," *Solid-State Circuits, IEEE Journal of*, vol. 41, pp. 461-473, Feb. 2006.

[16]   P. Caputa and C. Svensson, "A 3Gb/s/wire global on-chip bus with near velocity-of-light latency," *VLSI Design, Intern. Conf. on*, pp. 6 pp., Jan. 2006.

[17]   H. Ito, H. Sugita, K. Okada, and K. Masu, "4 Gbps on-chip interconnection using differential transmission line," *IEEE Asian Solid State Circuits Conference*, pp. 417-420, April 2005.

[18]   J. J. Kang, J. Y. Park, and M. P. Flynn, "Global High-Speed Signaling in Nanometer CMOS," *IEEE Asian Solid State Circuits Conference*, pp. 393-396, April 2005.

[19]   M. P. Flynn and J. J. Kang, "Global signaling over lossy transmission lines," *Computer-Aided Design (ICCAD), Digest of Techn. Papers IEEE/ACM Intern. Conf. on*, pp. 985-992, Nov. 2005.

[20]   P. Caputa and C. Svensson, "Low-power, low-latency global interconnect," *ASIC/SOC Conference, Proc. of the IEEE International*, pp. 394-498, Sept. 2002.

[21]   R. T. Chang, C. P. Yue, and S. S. Wong, "Near speed-of-light on-chip electrical interconnect," *VLSI Circuits, Digest of Tech. Papers, Symp. on*, pp. 18-21, June 2002.

[22]   A. P. Jose, G. Patounakis, and K. L. Shepard, "Pulsed current-mode signaling for nearly speed-of-light intrachip communication," *Solid-State Circuits, IEEE Journal of*, vol. 41, pp. 772-780, April 2006.

[23]   W. Pingshan, G. Pei, and E. C. C. Kan, "Pulsed wave interconnect," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, pp. 453-463, May 2004.

[24]   M. F. Chang, V. P. Roychowdhury, Z. Liyang, S. Hyunchol, and Q. Yongxi, "RF/wireless interconnect for inter- and intra-chip communications," *Proceedings of the IEEE*, vol. 89, pp. 456-466, April 2001.

[25]   R. Ho, K. Mai, and M. Horowitz, "Efficient on-chip global interconnects," *VLSI Circuits, Digest of Tech. Papers, Symp. on*, pp. 271-274, June, 2003.

[26]   K. Chang-Ki, R. Kwang-Myoung, and L. Kwyro, "High speed and low swing interface circuits using dynamic over-driving and adaptive sensing scheme," *VLSI and CAD, Intern. Conf. on*, pp. 388-391, Oct. 1999.

[27]   A. P. Jose and K. L. Shepard, "Distributed Loss Compensation for Low-latency On-chip Interconnects," *Int. Solid State Circuits Conf. (ISSCC), Dig. Tech. Papers*, pp. 516-517, Feb. 2006.

[28]   K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *Electron Devices, IEEE Transactions on*, vol. 49, pp. 2001-2007, Nov. 2002.

[29]   H. Zhang, V. George, and J. M. Rabaey, "Low-swing on-chip signaling techniques: effectiveness and robustness," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 8, pp. 264-272, June 2000.

144

[30]   C. Svensson, "Optimum voltage swing on on-chip and off-chip interconnect," *Solid-State Circuits, IEEE Journal of*, vol. 36, pp. 1108-1112, July 2001.

[31]   M. Ghoneima, Y. Ismail, M. Khellah, and V. De, "Reducing the Data Switching Activity of Serialized Datastreams," *Circuits and Systems (ISCAS), proc. of the IEEE Intern. Symp. on*, pp. 1015-1018, May 2006.

[32]   M. Aoki, Y. Nakagome, M. Horiguchi, H. Tanaka, S. Ikenaga, J. Etoh, Y. Kawamoto, S. Kimura, E. Takeda, H. Sunami, and K. Itoh, "A 60-ns 16-Mbit CMOS DRAM with a transposed data-line structure," *Solid-State Circuits, IEEE Journal of*, vol. 23, pp. 1113-1119, Oct. 1988.

[33]   M. Dong-Sun and D. W. Langer, "Multiple twisted dataline techniques for multigigabit DRAMs," *Solid-State Circuits, IEEE Journal of*, vol. 34, pp. 856-865, June 1999.

[34]   H. Hidaka, K. Fujishima, Y. Matsuda, M. Asakura, and T. Yoshihara, "Twisted bit-line architectures for multi-megabit DRAMs," *Solid-State Circuits, IEEE Journal of*, vol. 24-27, pp. 21, Feb. 1989.

[35]   K. Noda, K. Takeda, K. Matsui, S. Masuoka, H. Kawamoto, N. Ikezawa, Y. Aimoto, N. Nakamura, T. Iwasaki, H. Toyoshima, and T. Horiuchi, "An ultrahigh-density high-speed loadless four-transistor SRAM macro with twisted bitline architecture and triple-well shield," *Solid-State Circuits, IEEE Journal of*, vol. 36, pp. 510-515, March 2001.

[36]   K. Dong Gun, L. Heeseok, B. Seungyong, P. Bongcheol, and K. Joungho, "GHz twisted differential line structure on printed circuit board to minimize EMI and crosstalk noises," *Electronic Components and Technology Conference, Proceedings*, pp. 1058-1065, May 2002.

[37]   B. J. Wilkie, G. Parker, and J. E. Muyshondt, "Printed Circuit Board with an Integrated Twisted Pair Conductor," *U.S. Pat. No. 5,646,368*.

[38]   L. Deng and M. D. F. Wong, "Optimal algorithm for minimizing the number of twists in an on-chip bus," *Design, Automation and Test in Europe Conference and Exhibition, Proceedings*, vol. 2, pp. 1104-1109, Feb. 2004.

[39]   K. Dong Gun, A. Seungyoung, B. Seungyong, P. Bongcheol, S. Myunghee, and K. Joungho, "A novel twisted differential line for high-speed on-chip interconnections with reduced crosstalk," *Electronics Packaging Technology Conference*, pp. 180-183, Dec. 2002.

[40]   I. Hatirnaz and Y. Leblebici, "Modelling and implementation of twisted differential on-chip interconnects for crosstalk noise reduction," *Circuits and Systems (ISCAS), proc. of the IEEE Intern. Symp. on*, vol. 5, pp. V-185-V-188, May 2004.

[41]   G. Zhong, C. K. Koh, and K. Roy, "A twisted-bundle layout structure for minimizing inductive coupling noise," *Computer Aided Design (ICCAD), IEEE/ACM Intern. Conf. on*, pp. 406-411, Nov. 2000.

[42]   D. Schinkel, "Title not known yet," PhD Thesis, Faculty of Electrical Engineering, Mathematics & Computer Science, University of Twente, 2007/2008.

[43]   IMST GmbH, "Empire (3D electromagnetic time domain solver)," 2003.

[44]   D. K. Cheng, *Field and Wave Electromagnetics*, 2nd ed: Addison-Wesley Publishing Company, 1989.

[45]   P. A. Rizzi, *Microwave Engineering, Passive Circuits*: Prentice-Hall, 1988.

[46]   D. Divsalar and M. Simon, "Spectral Characteristics of Convolutionally Coded Digital Signals," *Communications, IEEE Transactions on*, vol. 28, pp. 173-186, Feb. 1980.

[47]   R. G. Gallager, *Information Theory and Reliable Communication*: John Wiley & Sons, 1969.

[48]   R. Bashirullah, L. Wentai, and R. Cavin, III, "Delay and power model for current-mode signaling in deep submicron global interconnects," *Custom Integrated Circuits Conference, Proc. of the IEEE*, pp. 513-516, May 2002.

[49]   K. Taeik, L. Xiaoyong, and D. J. Allstot, "Compact model generation for on-chip transmission lines," *Circuits and Systems I: Regular Papers, IEEE Transactions on [see also Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on]*, vol. 51, pp. 459-470, March, 2004.

[50]   S. Shekhar, J. S. Walling, and D. J. Allstot, "Bandwidth Extension Techniques for CMOS Amplifiers," *Solid-State Circuits, IEEE Journal of*, vol. 41, pp. 2424-2439, Nov. 2006.

[51]   V. Stojanovic, A. Ho, B. Garlepp, F. Chen, J. Wei, E. Alon, C. Werner, J. Zerbe, and M. A. Horowitz, "Adaptive equalization and data recovery in a dual-mode (PAM2/4) serial link transceiver," *VLSI Circuits, Digest of Tech. Papers, Symp. on*, pp. 348-351, June 2004.

[52]   D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl, and B. Nauta, "A 3Gb/s/ch transceiver for RC-limited on-chip interconnects," *Int. Solid State Circuits Conf. (ISSCC), Dig. Tech. Papers*, pp. 386-387,606, Feb. 2005.

[53]   E. Mensink, D. Schinkel, E. A. M. Klumperink, E. Van Tuijl, and B. Nauta, "Optimal Positions of Twists in Global On-Chip Differential Interconnects," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 15, pp. 438-446, April 2007.

[54]   R. Ho, T. Ono, F. Liu, R. Hopkins, A. Chow, J. Schauer, and R. Drost, "High-Speed and Low-Energy Capacitively-Driven On-Chip Wires," *Int. Solid State Circuits Conf. (ISSCC), Dig. Tech. Papers*, pp. 412-413,612, Feb. 2007.

[55]   E. Mensink, D. Schinkel, E. Klumperink, E. van Tuijl, and B. Nauta, "A 0.28pJ/b 2Gb/s/ch Transceiver in 90nm CMOS for 10mm On-Chip Interconnects," *Int. Solid State Circuits Conf. (ISSCC), Dig. Tech. Papers*, pp. 414-415,612, Feb. 2007.

[56]   D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl, and B. Nauta, "A Double-Tail Latch-Type Voltage Sense Amplifier with 18ps Setup+Hold Time," *Int. Solid State Circuits Conf. (ISSCC), Dig. Tech. Papers*, pp. 314-315,605, Feb. 2007.

146

Samenvatting

Een veel gebruikte technologie voor geïntegreerde schakelingen (chips) is CMOS. Door schaling van deze technologie, worden de transistors voor de schakelingen kleiner en daarmee sneller. De snelheid van verbindingsdraden, die zorgen voor de communicatie tussen verschillende onderdelen van de chip, is evenredig met hun lengte in het kwadraat. Vaak houden chips in een nieuwere technologie dezelfde afmetingen door meer functionaliteit op de chip. Verbindingsdraden die de hele chip overspannen houden dan dezelfde lengte. Daarmee worden ze niet sneller en kunnen ze de snellere schakelingen niet bijhouden. Er ontstaat een communicatieprobleem.

De snelheid van een schakeling komt naar voren in het aantal bits dat het per seconde kan versturen. De verbindingsdraad zal ditzelfde aantal bits per seconde moeten kunnen overbrengen naar een ontvanger. De snelheid van de verbindingsdraad wordt echter beperkt door de grote weerstand en grote capaciteit van de draad. In dit proefschrift wordt ingegaan op methoden om de snelheid van de verbindingsdraden te vergroten. De chip oppervlakte en vermogensconsumptie moeten hierbij zo laag mogelijk worden gehouden. Verder moet de integriteit van de verstuurde data worden gewaarborgd.

De verbindingsdraad kan beschreven worden met overdrachtsfuncties en oogopeningen. Deze kunnen gevonden worden met behulp van een gedistribueerd model dat beschreven kan worden met s-parameters. Het gedistribueerde model kent een gedistribueerde weerstand, inductantie, conductantie en capaciteit toe aan de draad. Als er meerdere draden zijn, die parallel naast elkaar lopen, zal er overspraak zijn tussen de draden. Het belangrijkste mechanisme is via een gedistribueerde mutuele capaciteit tussen twee draden. De waardes van de parameters van het gedistribueerde model kunnen gevonden worden met behulp van een 3D elektromagnetische veld simulator. Met het gedistribueerde model kan ook de vermogensconsumptie worden berekend.

Het gedrag van lange verbindingsdraden wordt gedomineerd door de gedistribueerde weerstand en de gedistribueerde capaciteit van de draden. Samen beperken deze parameters de bandbreedte van de draden. Voor de draden, zoals beschreven in dit proefschrift, is deze bandbreedte in de orde van 100 MHz. Deze bandbreedte is afhankelijk van de afmetingen van de draden. Een ontwerper kan de breedte en de afstand tot een buurdraad bepalen. De bandbreedte per chip oppervlakte is het grootst als deze breedte en afstand gelijk worden gekozen aan de hoogte van de draad en de verticale afstand tot andere metaallagen. Door de

bandbreedte per chip oppervlakte te maximaliseren, wordt de hoogste totale datacapaciteit bereikt voor een beschikbare hoeveelheid chip oppervlakte.

De bandbreedte van de draden kan vergroot worden door de juiste afsluitimpedanties te kiezen. Een conventionele zender bestaat uit een inverter. Door na deze inverter een serie-capaciteit te plaatsen wordt de bandbreedte tot een factor drie verhoogd. Een bijkomend voordeel is dat de spanningslag op de draad lager wordt, waardoor de vermogensconsumptie gereduceerd wordt. Een conventionele ontvanger bestaat ook uit een inverter. Door in plaats van deze inverter een laagohmige afsluiting te maken, bijvoorbeeld met een transimpedantieversterker, kan opnieuw de bandbreedte tot een factor drie worden verhoogd. De spanningslag wordt alleen aan het einde van de draad gereduceerd, waardoor de winst in vermogensconsumptie kleiner is dan bij de capacitieve afsluiting aan de zenderkant. Tevens zorgt de laagohmige afsluiting aan de ontvangerkant voor een ongewenste statische stroomconsumptie. De bandbreedte van de draden kan ook verhoogd worden door gebruik te maken van egalisatietechnieken als 'pulse-width pre-emphasis' en 'decision feedback equalization'.

Om de data integriteit te waarborgen kunnen verschillende technieken worden toegepast. Allereerst zorgt het gebruik van differentiële draden dat de storing van de meeste ruisbronnen op beide draden even groot is en daardoor kan worden opgeheven (de storing is 'common-mode'). Door in de differentiële draden een twist te leggen kan overspraak van buurdraden worden verminderd. Een schema dat zowel differentiële overspraak als common-mode overspraak reduceert heeft om en om telkens een enkele twist in het ene differentiële paar en een dubbele twist in het differentiële buurpaar. De positie van de twists bepaalt hoe goed de overspraak wordt onderdrukt. De optimale positie hangt af van de gebruikte afsluitimpedanties.

Een testchip toont aan dat een 10 mm lange draad met een conventionele zender en ontvanger een datacapaciteit haalt van 550 Mb/s. De vermogensconsumptie is 3.4 pJ/b bij vijftig procent data activiteit. Door het gebruik van 'pulse-width pre-emphasis' en een laagohmige afsluiting bij de ontvanger is een datacapaciteit van 3 Gb/s mogelijk met een vermogensconsumptie van 2 pJ/b. De serie-capaciteit aan de zenderkant in combinatie met 'decision feedback equalization' haalt 2 Gb/s met een vermogensconsumptie van slechts 0.28 pJ/b. Beide oplossingen halen een hogere datacapaciteit per chip oppervlakte dan oplossingen uit de literatuur ($2 - 10^3$ x). De laatste oplossing heeft ook nog een significant lagere vermogensconsumptie ($3 - 30$ x).

Publicaties

1. E. Mensink, D. Schinkel, E. A. M. Klumperink, E. Van Tuijl, and B. Nauta, "Optimal Positions of Twists in Global On-Chip Differential Interconnects," Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, pp. 438-446, vol. 15, April 2007.

2. E. Mensink, D. Schinkel, E. Klumperink, E. van Tuijl, and B. Nauta, "A 0.28pJ/b 2Gb/s/ch Transceiver in 90nm CMOS for 10mm On-Chip Interconnects," Int. Solid State Circuits Conf. (ISSCC), Dig. Tech. Papers, pp. 414-415, 612, Feb. 2007.

3. E. Mensink, D. Schinkel, E. A. M. Klumperink, E. Van Tuijl, and B. Nauta, "Global On-Chip Differential Interconnects with Optimally-Placed Twists," Circuits, Systems and Signal Processing (ProRISC), Annual Workshop on, Nov. 2005.

4. E. Mensink, D. Schinkel, E. Klumperink, E. van Tuijl, and B. Nauta, "Optimally-placed twists in global on-chip differential interconnects," European Solid-State Circuits Conf. (ESSCIRC), Proc. of the, pp. 475-478, Sept. 2005.

5. E. Mensink, D. Schinkel, E. A. M. Klumperink, and E. Van Tuijl, "Interconnects and On-Chip Data Communication Techniques," Circuits, Systems and Signal Processing (ProRISC), Annual Workshop on, Nov. 2004.

6. D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl, and B. Nauta, "A Double-Tail Latch-Type Voltage Sense Amplifier with 18ps Setup+Hold Time," Int. Solid State Circuits Conf. (ISSCC), Dig. Tech. Papers, pp. 314-315,605, Feb. 2007.

7. D. Schinkel, E. Mensink, E. A. M. Klumperink, E. van Tuijl, and B. Nauta, "A 3-Gb/s/ch transceiver for 10-mm uninterrupted RC-limited global on-chip interconnects," Solid-State Circuits, IEEE Journal of, vol. 41, pp. 297-306, Jan. 2006.

8. D. Schinkel, E. Mensink, E. A. M. Klumperink, E. Van Tuijl, and B. Nauta, "A Transceiver for High-Speed Global On-Chip Data Communication," Circuits, Systems and Signal Processing (ProRISC), Annual Workshop on, Nov. 2005.

9. D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl, and B. Nauta, "A 3Gb/s/ch transceiver for RC-limited on-chip interconnects," Int. Solid State Circuits Conf. (ISSCC), Dig. Tech. Papers, pp. 386-387, 606, Feb. 2005.

Dankwoord

"Het begin van alle kennis is ontzag voor de HEER." (Spreuken 1:7).

Tijdens zijn promotieonderzoek probeert een promovendus kennis op te doen over een specifiek onderwerp. In mijn geval was het onderwerp verbindingsdraden op een CMOS chip en het onderzoek heeft geresulteerd in dit proefschrift. In dit dankwoord wil ik een aantal personen bedanken die bij hebben gedragen aan de totstandkoming van dit proefschrift.

Allereerst denk ik aan mijn ouders, broer, zussen en hun gezinnen*: "Hoe goed is het, hoe heerlijk als broeders bijeen te wonen!"* (Psalm 133:1). Halverwege mijn promotie-onderzoek ben ik getrouwd met Alien: *"Een sterke vrouw, wie zal haar vinden? Zij is meer waard dan edelstenen."* (Spreuken 31:10).

Motiveren is een van de sterke punten van de leerstoelhouder van de IC-Design groep. Bram, jij bent degene die me overgehaald heeft om promovendus te worden en ik heb er zeker geen spijt van. Eric was mijn dagelijks begeleider en ook Ed (mijn promotor) was wekelijks bij het project betrokken. Jullie hebben beiden een wezenlijke bijdrage geleverd aan de behaalde resultaten.

Wie ik zeker niet kan vergeten te bedanken is Daniël. We hebben het onderzoek samen gedaan en ik vind dat er een goede samenwerking was. Ook was je een leuke kamergenoot, samen met Mustafa. Verder heb ik een aantal studenten mogen begeleiden, wat voor een leuke afwisseling zorgde.

Andere belangrijke mensen voor het project waren Gerard, die een gedeelte van de laatste chip heeft gelayout, Henk voor de testopstelling, Frederik en Cor voor computerproblemen en de secretaresses Gerdien en Annemiek. En verder alle leden van de vaste staf van IC-Design en de overige promovendi.

De leden van de gebruikerscommissie, André.Nieuwland, Atul Katoch, Gerrit den Besten, Jan Geralt bij de Vaate, Jan Visschers, Joop van Lammeren, Marc van Heijningen en Frank Karelse, wil ik danken voor hun waardevolle input. Verder wil ik graag de (voormalige) mixed-signal groep van Philips Research bedanken voor het beschikbaar stellen van silicium en de hulp bij het 'chip-finishen'.

About the author

Eisse Mensink was born on January 10, 1979, in Almelo, The Netherlands. He received the M.Sc. degree in electrical engineering (cum laude) from the University of Twente, The Netherlands, in 2003. The past four years, he has been working toward the Ph.D. degree at the same university on the subject of high-speed on-chip communication.