Interconnects and On-Chip Data Communication Techniques

Eisse Mensink, Daniel Schinkel, Eric Klumperink and Ed van Tuijl
IC-Design Group, University of Twente
P.O.Box 217, 7500 AE Enschede, The Netherlands
E-mail: e.mensink@utwente.nl

Abstract—**Global on-chip communication is rapidly becoming a speed and power bottleneck in CMOS circuits. In this paper, a 'mixed-signal' approach is taken to analyze on-chip interconnects and it is investigated how data-rates can be improved. It is shown that complex signaling schemes such as OFDM and CDMA are not efficient to improve the data-rate, while equalization can significantly improve the achievable data-rate. A combination of an equalizing transmitter with a low-ohmic receiver can improve the achievable data-rate by about a factor of 7.**

Keywords— **on-chip; communication; interconnect; CMOS; mixed signal communication techniques**

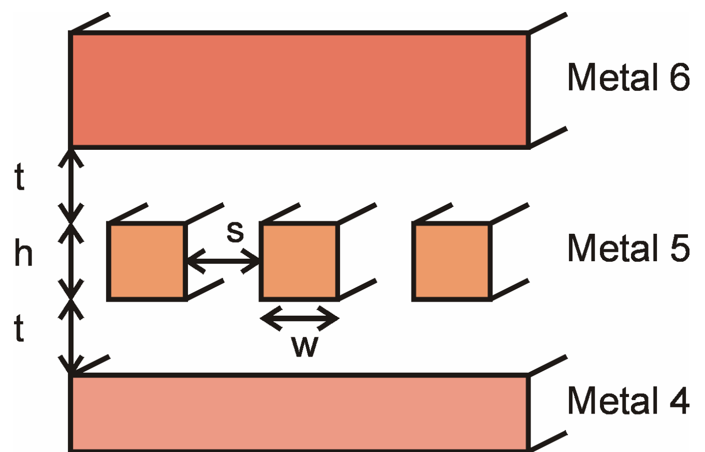

Figure 1: interconnect structure

I. Introduction

The ever-increasing disparity between on-chip interconnects and gate delays makes these interconnects a speed bottleneck for digital systems. While transistors get faster as technology scales, the shrinking dimensions of on-chip interconnects limit their speed.

[1] analyses the problem. Historically, the delays of gates (FO4) have scaled linearly with technology. As this trend may continue for future generations of transistors, the delays of gates are expected to decrease.

The delays of interconnects depend mainly on their distributed resistance and capacitance. Resistance per length grows under scaling, since the width and height of interconnects both scale down, while capacitance per length should be roughly constant with scaling. For constant interconnect length, therefore, the delay is increasing.

However, as [1] points out, in discussions about interconnect delays under technology scaling, an important distinction should be made between interconnects that scale in length and interconnects that do not scale in length.

The first kind of interconnects are between gates within blocks, and when devices (and blocks) get smaller, these interconnects get shorter. As the RC product is proportional to the length squared, the delay of these interconnects scales with technology. Therefore, these local interconnects have no performance problem.

The second kind of interconnects span the whole chip area. These global interconnects do not scale in length and therefore, their delay is increasing under scaling. They cannot keep up with the increasing speed of transistors and solutions need to be found.

Traditional communication techniques over on-chip interconnects take a more or less digital approach, using inverters as transmitters, repeaters and receivers and with a focus on delay time estimation and optimal repeater insertion [2].

In this paper, a more 'mixed-signal' approach is taken to analyze on-chip interconnects. Not only the latency of global on-chip interconnects increases with scaling, also the data capacity decreases as the bandwidth is reduced. This paper investigates how the data capacity of global on-chip interconnects can be maximized for minimal chip area and power consumption. It looks at point-to-point communication in a global bus.

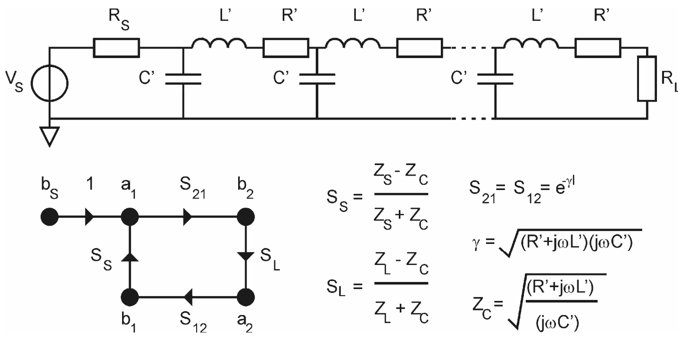

Figure 2: distributed RLC model of the interconnect and an equivalent model with s-paramters

The next section, interconnect design, describes the modeling, dimensions and termination of the interconnects. Section III, modem-techniques, describes which communication techniques efficiently improve the data capacity. Section IV gives the conclusions.

II. Interconnect Design

A. Interconnect model

In Fig. 1, a general structure for the interconnects is drawn. We assume that the top metal layer is reserved for power and clock routing and therefore place the global interconnects one level below. The higher and lower metal layers are filled with metal, to approximate the effect of other high-density interconnects.

With a 3D EM-field simulator, the s-parameters for such an interconnect structure can be found. These s-parameters can be mapped on a distributed RLC model (see Fig. 2) and the values of R', L' and C' can be calculated.

For a 0.4µm wide interconnect with a spacing of 0.4µm to the neighboring interconnects (see section II-B), R'≈0.15kΩ/mm, L'≈0.35nH/mm and C'≈0.22pF/mm.

In Fig. 3, the transfer function (distributed model) of this interconnect is drawn, with R_S=10Ω and R_L=100Ω. It is calculated from the s-parameters as follows.

$$H = \frac{Z_c}{Z_c + R_S} \cdot \frac{S_{21}(1+S_L)}{1 - S_{21}{}^2 S_S S_L} \tag{1}$$

The transfer function consists of three areas. The first part has a first order RC roll-off. The -3dB-point depends on R', C', R_S, R_L and the length of the interconnect and can be approximated by [3]:

$$BW_{-3dB} = \frac{1}{0.5 R_T C_T \cdot 2\pi} \cdot \frac{R_S + R_T + R_L}{R_S + \frac{R_T}{3} + R_L + \frac{2 R_S R_L}{R_T}} \tag{2}$$

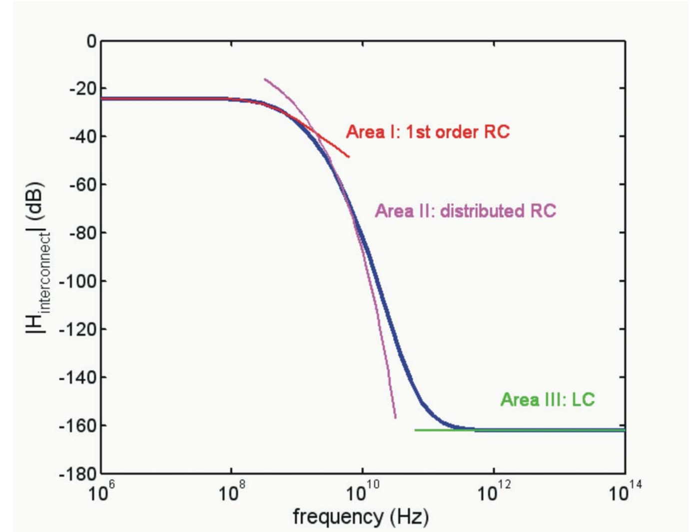

Figure 3: interconnect transfer function showing three areas

where R_T=R'*length and C_T=C'*length are the total resistance and capacitance of the interconnect.

In the second and third area, the influence of R_S and R_L (or S_S and S_L in Fig. 2) are not seen anymore and the transfer function follows the S21 of the interconnects. In area II, the distributed RC gives large attenuation. The third area starts when R<<ωL (>>70GHz). In this area, the interconnect behaves as a transmission line with LC behavior.

Using the transmission line characteristics of area III would be interesting. However, this area starts only at very high frequencies and the attenuation is extremely large. [4] tries to make the interconnects more inductive by using very wide interconnects (16µm). However, that solution uses a lot of area and power.

In this paper, the first area is used and the interconnects can be modeled by first order RC behavior.

B. Dimensions

In the previous section, the transfer function of a 0.4µm wide interconnect with a spacing of 0.4µm to its neighbors is shown. But what is the best width and spacing?

Since a high data capacity is desired on the RC-limited interconnect, the bandwidth is used as a performance measure. However, as chip area is to be minimized, the bandwidth is divided by the cross-section of the interconnect (w+s)*(h+t) (see Fig. 1). In this way, an optimal bandwidth per cross-sectional area can be found.

Figure 4 shows the simulated bandwidth per cross-sectional area as a function of the width (w) and spacing (s). There is a clear optimum at w=s=h=t. In this case, the maximum cross-sectional area (for low resistance) at minimum perimeter (for low capacitance) is reached.

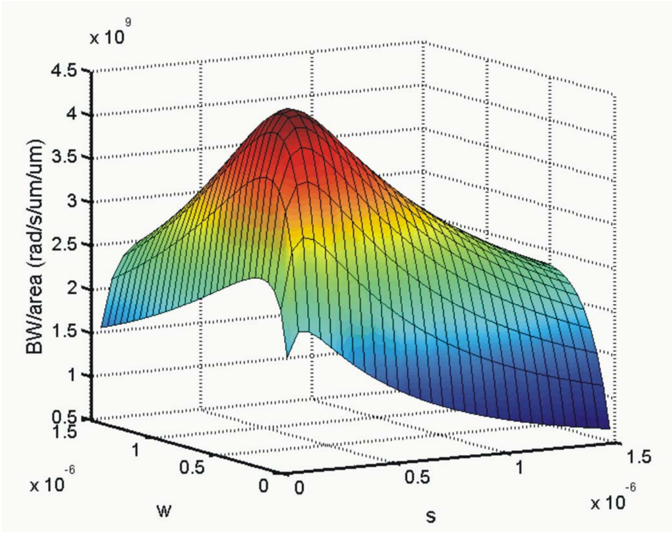

Figure 4: bandwidth per cross-sectional area as a function of width and spacing

The height (h) and vertical spacing (t) are determined by the process, width and spacing are chosen by the designer. The width and spacing of 0.4µm in the previous section were chosen for optimal bandwidth per area.

C. Termination

In traditional designs, global interconnects are terminated with the gate capacitance of a transistor, in Fig. 2 and Eq. 2 this means $R_L=\infty$. Terminating the interconnect with a low resistance can improve the bandwidth. [3] looks at the delay of bit lines in a CMOS SRAM and shows a large factor (20) of improvement by using a small R_L instead of a capacitive load. In that case, R_S is large compared to R_T. According to Eq. 2, the bandwidth with $R_L=\infty$ is

$$BW_{-3dB} = \frac{1}{0.5R_TC_T\cdot 2\pi}\cdot\frac{R_T}{R_T+2R_S} \xrightarrow{R_S>>R_T} \frac{1}{0.5R_TC_T\cdot 2\pi}\cdot\frac{R_T}{2R_S} \quad (3)$$

and with R_L = low-ohmic

$$BW_{-3dB} = \frac{1}{0.5R_TC_T\cdot 2\pi}\cdot\frac{R_S+R_T}{R_S+\frac{R_T}{3}} \xrightarrow{R_S>>R_T} \frac{1}{0.5R_TC_T\cdot 2\pi} \quad (4)$$

For $R_S>>R_T$, the bandwidth is improved by a factor $2R_S/R_T$.

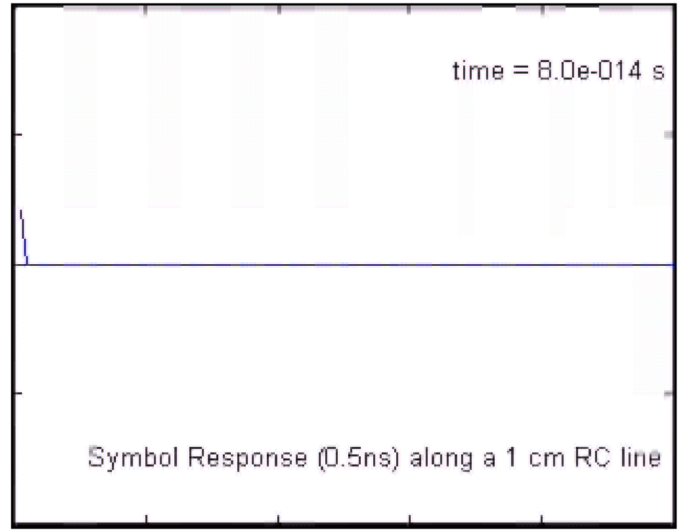

Figure 5: animated voltage across an interconnect with high-ohmic load

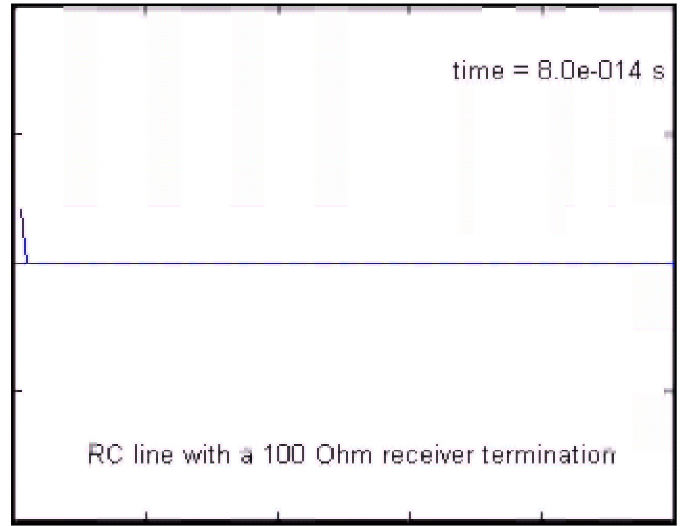

Figure 6: animated voltage across an interconnect with low-ohmic load

However, as Eq. 2 is symmetrical for R_S and R_L, it is for highest bandwidth better to have also a low R_S. Then the bandwidth with $R_L=\infty$ is

$$BW_{-3dB} = \frac{1}{0.5R_TC_T\cdot 2\pi}\cdot\frac{R_T}{R_T+2R_S} \xrightarrow{R_S<<R_T} \frac{1}{0.5R_TC_T\cdot 2\pi} \quad (5)$$

and with R_L = low-ohmic

$$BW_{-3dB} = \frac{1}{0.5R_TC_T\cdot 2\pi}\cdot\frac{R_S+R_T}{R_S+\frac{R_T}{3}} \xrightarrow{R_S<<R_T} \frac{1}{0.5R_TC_T\cdot 2\pi}\cdot\frac{1}{3} \quad (6)$$

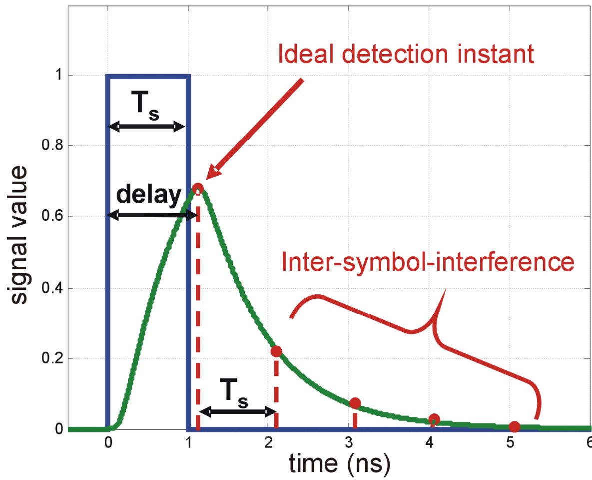

Figure 7: received symbol waveshape and ISI points

The results of Eq. 4 and 5 show the symmetry for R_S and R_L. Comparing with Eq. 6 shows that both R_S and R_L should be chosen low-ohmic (as compared to R_T) for highest bandwidth (3 times higher).

This effect can also be shown in the time domain. In the time domain, the limited bandwidth of the interconnect manifests itself in intersymbol interference (ISI). After a symbol is transmitted on the interconnect, there remains charge on the line for a long period of time, interfering with consecutive symbols. If R_S is low-ohmic, but R_L high-ohmic, all the remaining charge has to diffuse (a distributed RC line is governed by diffusion equations) to the source (see Fig. 5, which animates the voltage across the interconnect). If not only R_S, but also R_L is low-ohmic, the charge can diffuse in both directions and it takes less time for the interconnect to remove all remaining charge from previous symbols (click on Fig. 6).

III. Modem-Techniques

A. Analysis method

To evaluate different modem techniques, eye-diagrams are studied. Often, the properties of an eye-diagram are analyzed on an ad-hoc basis, using a simulation trial with a more or less arbitrary sequence of symbols that are put through a communication channel. However, it is explained here that it is also easily possible to extract eye-diagram properties directly from the received symbol waveshape, using a method similar to the method described in [5].

First, the received symbol waveshape has to be determined. In case of conventional binary signaling this is just the convolution of a square pulse with the line

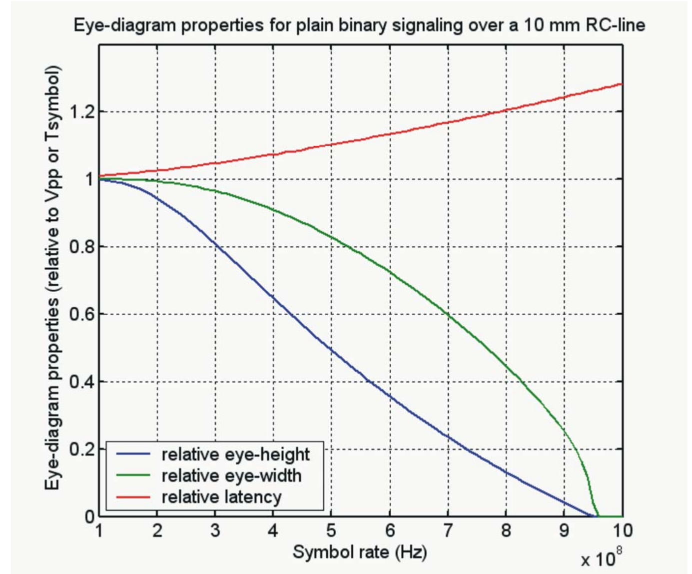

Figure 8: eye properties for conventional binary signaling

impulse response, as simulated with the distributed RLC model. Figure 7 shows an example of a received symbol waveshape. The figure shows that the interconnect, due to the limited bandwidth, suffers from severe intersymbol interference (ISI). The worst-case amount of ISI is determined by summing the absolute values of the symbol response at instants that are separated an integer number of symbol times (T_S) from the detection instant. The eye-opening of an eye-diagram is equal to the difference between the received value at the detection instant, minus the amount of ISI. So, by evaluating the eye-opening curve for different detection instants, both the ideal detection point (the point with maximum eye-opening) together with the width of the eye can be determined. If this is done for different data-rates, eye property diagrams as in Fig. 8 can be made. The figure shows the relative eye-width (relative to one symbol period), the relative eye-height (relative to the maximum received value) and the relative latency (again relative to one symbol period). Fig. 8 shows the result for conventional binary signaling. At a data-rate of 0.5Gb/s, the eye-opening is reduced to 50%.

B. Multi-carrier schemes

In this section it will be discussed if simple variants of multi-carrier schemes can be used to improve the data capacity of the interconnects. These schemes can be analyzed by transforming the signaling scheme to a functionally equal system that consists entirely of filters and samplers. OFDM and CDMA systems can be viewed as systems that use filter-banks for the transmission and modulation of different channels and matched filter-banks for the reception (see Fig. 9). With this analogy, it

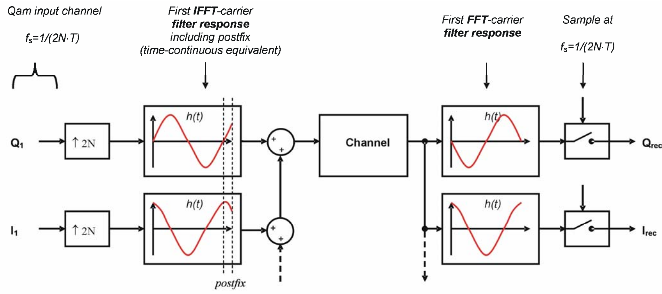

Figure 9: equivalent schematic for OFDM system (showing only the first carrier)

is again possible to derive the eye-diagram properties at the receiving end. One can analyze the detected symbol response of a certain channel by convolving the transmitter channel filter with the line-response and with the matched filter at the receiver.

The difference with plain (single-channel) binary signaling is that it is now also necessary to examine the inter-channel interference, because a non-ideal channel can result in a loss of channel-orthogonallity. The inter-channel interference (ICI) can be examined by convolving the received symbol response of one channel with the matched filter of another channel.

It was found in this case that the worst-case eye-opening analysis that is usefull in single-channel systems gives a pessimistic estimate of the achievable data-rate. With multi-channel systems, the chance that this worst-case situation occurs can become arbitrarily small if the numbers of channels increase. Therefore, the variance of the ISI and ICI is estimated. The variance is then used to calculate the BER, assuming that the total interference has a Gaussian distribution. This is a coarse approximation (validity increases with higher number of channels), but system simulations have verified that these a-priori BER estimations are accurate within an order of magnitude.

Figure 10 shows the BER for different detection instants at a data-rate of 1Gb/s for a 4-channel CDMA scheme. As the figure shows, the BER is far too large for reliable communication.

The problem with multi-carrier schemes is mainly the ICI: the interconnect destroys the orthogonallity between the different channels. Both CDMA and OFDM (even with long cyclic prefixes) schemes suffer from inter-channel interference due to this loss of orthogonallity and are therefore not useful for reliable signaling across on-chip interconnects, at least not without additional channel equalization. However, a combination of e.g. CDMA and channel equalization will become very complex and has little advantage over plain binary signaling with equalization, as described next.

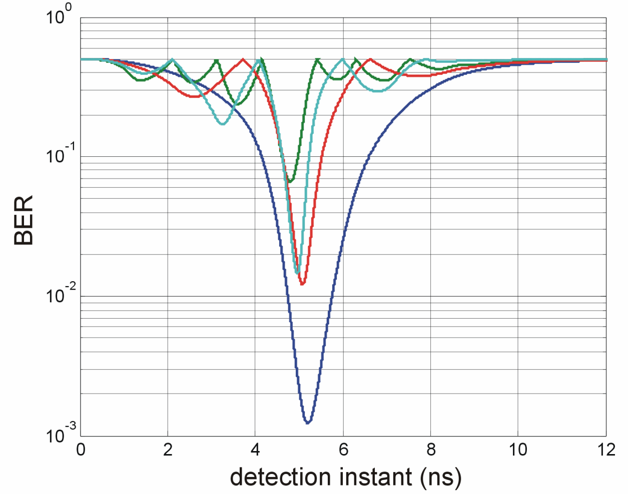

Figure 10: BER for a 4-channel CDMA scheme at 1Gb/s

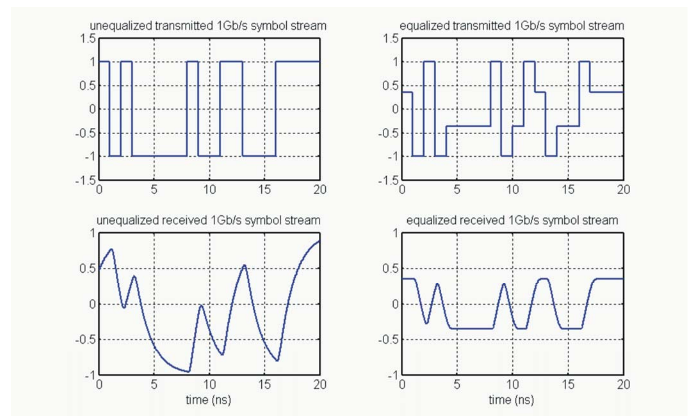

Figure 11: sent and received signals for conventional binary signaling and FIR equalization

C. Equalization

While complex signaling schemes such as OFDM and CDMA are not efficient to improve the data-rate, equalization can significantly improve the achievable data-rate. Due to the dominant first-order RC roll-off of the interconnect transfer function (see section II-A), a simple 2-taps equalization at the transmitter side can mitigate ISI. Figure 11 shows the effect of pre-emphasis/overdrive signaling (2-taps FIR-filter) and how it helps to overcome ISI effects.

In Fig. 12, the relative eye-diagram properties are shown for this FIR equalization ($R_L = \infty$). Since the maximum voltage swing at the receiver side is reduced by the equalization, the relative swing is also shown. This relative swing drops for higher data-rates, because more equalization is needed then. A minimum relative swing of 0.1 is used in the figure. The improvement in achievable data-rate, compared to Fig. 8, is apparent. Figure 13 shows that with a combination of low-ohmic

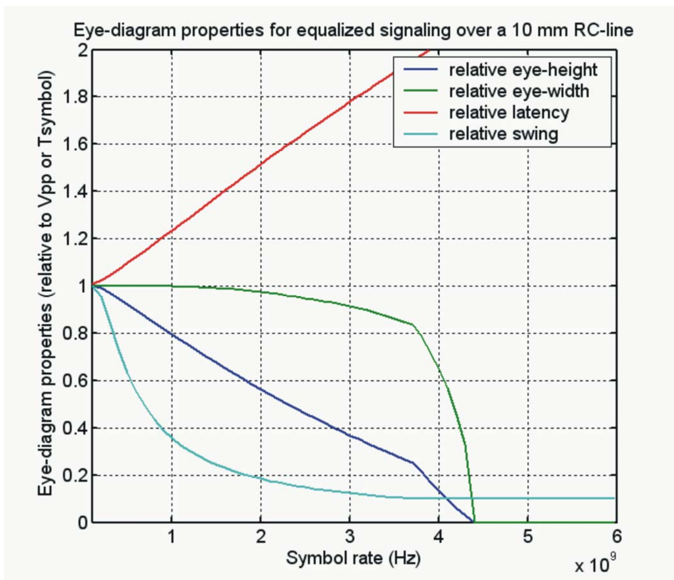

Figure 12: eye properties for FIR equalization and high-ohmic termination

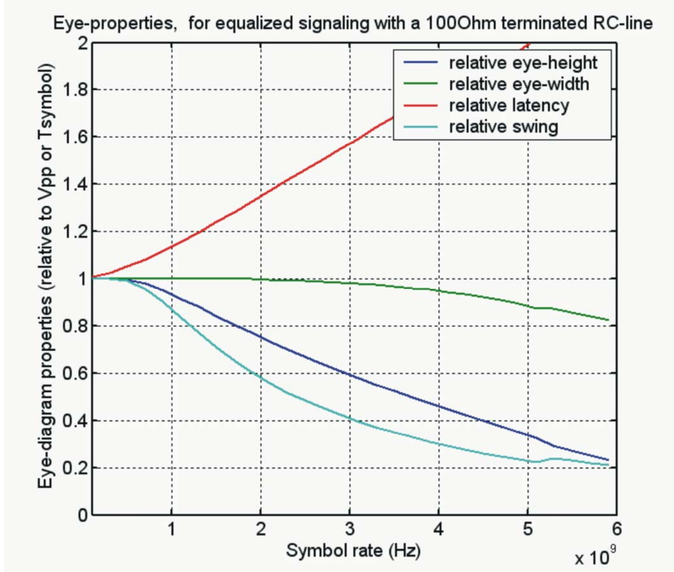

Figure 13: eye properties for FIR equalization and low-ohmic termination

termination (see section II-B) and equalization, data-rates of 3-4Gb/s are possible (50% eye-opening), which is a factor of 7 improvement compared to high-ohmic termination with no equalization (see Fig. 8).

The effect of FIR equalization is also shown in an animation, that shows the voltage across the interconnect (Fig. 14).

IV. Conclusions

In this paper, a ‘mixed-signal’ approach is taken to analyze on-chip interconnects. It is investigated how to improve the data-rate of global on-chip interconnects with minimal chip area and power consumption. A significant part of the transfer function can be

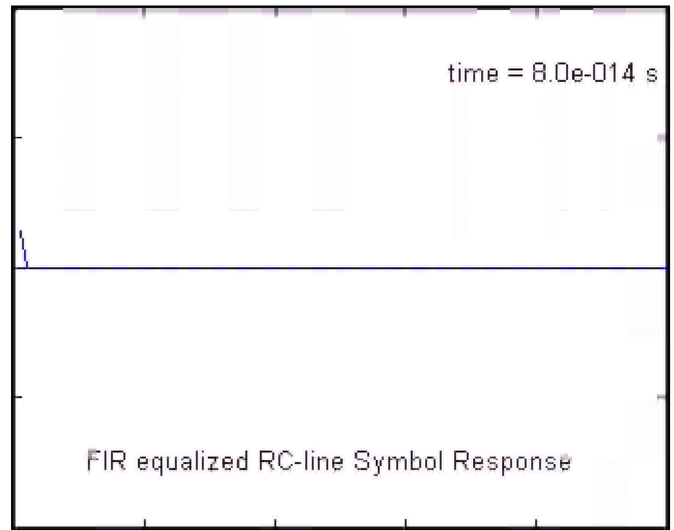

Figure 14: animated voltage across an interconnect with FIR equalization

approximated by a first-order RC model. The optimal bandwidth per cross-sectional area is found if all dimensions (width, height, horizontal and vertical spacing) are equal. The interconnects should be terminated with low impedance (both source and load impedances) for highest bandwidth. Multi-carrier schemes are not beneficial as the finite bandwidth of the interconnect destroys the orthogonallity between different channels. As a result, these schemes suffer from severe inter-channel interference. Equalization on the other hand, is beneficial for on-chip interconnects.

A combination of low-ohmic termination and equalization can improve the achievable data-rate by about a factor of 7.

Acknowledgement

Authors thank the Dutch Technology Foundation (STW, project TCS.5791) for funding.

References

[1] R. Ho, K. W. Mai, M. A. Horowitz, “The Future of Wires,” Proc. IEEE, pp. 490-504, April 2001.
[2] H. Bakoglu, “Circuits, Interconnections and Packaging for VLSI,” Addison-Wesley, 1990.
[3] E. Seevinck, P. van Beers, H. Ontrop, “Current-Mode Techniques for High-Speed VLSI Circuits with Application to Current Sense Amplifier for CMOS SRAM’s,” J. Solid State Circuits, vol. 26, pp. 525-536, April 1991.
[4] R. Chang, et. al., “Near Speed-of-Light Signaling Over On-Chip Electrical Interconnects”, J. Solid State Circuits, vol. 38, pp. 834-838, May 2003.
[5] P. Hanumolu et. al., “Analysis of PLL Clock Jitter in High-Speed Serial Links,” IEEE Trans. On Circuits and Systems II, vol. 50, no. 11, pp. 879-886, Nov. 2003.