Optimising Bandwidth Over Deep Sub-micron Interconnect[1]

Dinesh Pamunuwa, Li-Rong Zheng and Hannu Tenhunen

Royal Institute of Technology (KTH), IMIT, LECS
Electrum 229, SE-164 40 Kista, Sweden
dinesh/lrzheng/hannu@ele.kth.se

ABSTRACT

In deep sub-micron (DSM) circuits proper analysis of interconnect delay is very important. When relatively long wires are placed in parallel, it is essential to include the effects of cross-talk on delay. In a parallel wire structure, the exact spacing and size of the wires determine both the resistance and the distribution of the capacitance between the ground plane and the adjacent signal carrying conductors, and have a direct effect on the delay. Repeater insertion depending on whether it is optimal or constrained, affects the delay in different ways. Considering all these effects we show that there is a clear optimum configuration for the wires which maximises the *total* bandwidth. Our analysis is valid for lossy interconnects as are typical of wires in DSM technologies.

1. INTRODUCTION

Interconnects in deep sub-micron technologies are typically very lossy so that the RC delay dominates. In order to keep the resistance to a minimum the aspect ratio (width/height) of wires are kept low, which however gives rise to increased inter-wire capacitance. This inter-wire capacitance results in cross-talk which has an effect on the delay, depending on how the aggressor lines switch. Cross-talk is of especial significance in uniformly coupled parallel wires, causing unpredictable delays[1]. Recently there has been a profusion of research into block oriented architectures[2] which are being proposed as being suitable to overcome the interconnect bottleneck and to cope with complexity. The intra-block communication link in all of these will consist of a large number of parallel wires, with uniform coupling over most of the wire length in all probability. The question we pose and attempt to answer in this paper is, given a fixed area in which to distribute the interconnect, what is the best configuration of the wires to obtain the highest bandwidth? Is it to have a few fat wires and a high signaling frequency, or a large number of small wires with a lower signaling frequency, or anything in between? How does the wire spacing affect overall bandwidth? What effect does repeater insertion have?

In such an analysis, it is essential that the wire capacitance is distributed over a ground component and coupling component, and the effect of cross-talk *on delay* is taken into account. Our electrical model for investigating delay is shown in Fig. 1. Each line, except the two peripheral lines are coupled on both sides to aggressors. The exact effect of cross-talk on delay depends on the switching alignment of the aggressors with respect to the victim. In [1] and [3] we give an analysis of delay in such uniformly coupled lines with different switching patterns. Since in general, it is necessary to design for the worst-

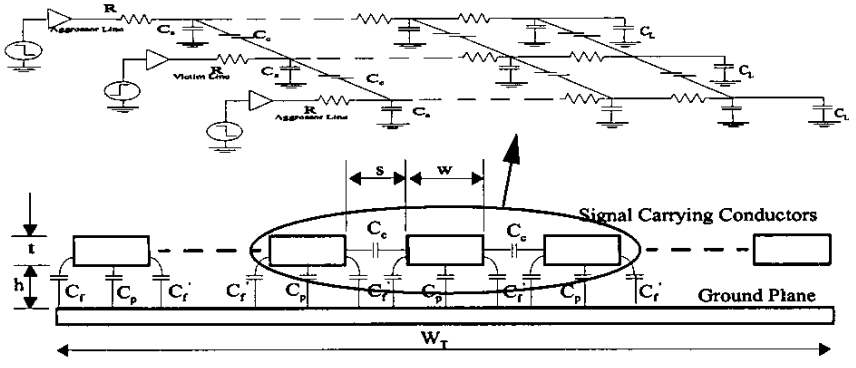

Figure 1. Configuration for investigating effect of cross talk

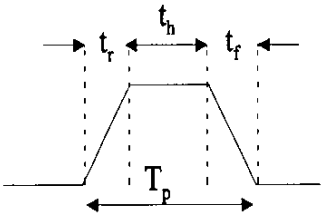

Figure 2. Minimum Bit Period

1. The funding support of Sida and that of Vinnova via the Socware and Exsite Programs are gratefully acknowledged

case, the appropriate expressions (which are reproduced in this article) are used for calculation of bandwidth. It is equally important to have accurate closed form equations for the capacitance terms shown in the figure. We use the expressions given in [4] which are empirical equations exhibiting good accuracy.

The rest of this document is structured in this manner. In section 2 we consider the intrinsic delay of the line, and derive expressions for the optimal bandwidth. In section 3 we include the effects of the non-ideal drivers and consider repeater insertion. Finally we give our conclusions.

2. LINE DELAY

All lines are *uniformly* coupled to two aggressor lines, except the two corner-most lines which are coupled to one aggressor. It has been shown that the dominant time constant approximation for the worst-case 50% delay for such lines is accurate to over 95% [3]. Eq. (1) gives the delay for the middle conductors while (2) gives the delay for the corner conductors[1] where the resistance and capacitances refer to the total values for the lines (see Fig. 1).

$$T_{0.5,mid} = 0.4RC_s + 1.51RC_c \quad (1)$$

$$T_{0.5,corn} = 0.4RC_s + 0.75RC_c \quad (2)$$

The delay is matched to the bit period by apportioning some percentage of the period to the rise and fall times so that a sufficient margin is allowed (Fig. 2). Usually 3 delays are considered to be sufficient for the response to pass the 90% threshold. We chose to be more conservative and use 4. Hence the bit period is given by (3) and (4)

$$T_{p,mid} = 4T_{0.5,mid} \quad (3)$$

$$T_{p,corn} = 4T_{0.5,corn} \quad (4)$$

For N conductors, (5) gives the total bandwidth.

$$BW = \frac{2}{T_{P,corn}} + \frac{N-2}{T_{P,mid}} \quad (5)$$

Now with reference to Fig. 1, R, C_s, and C_c are defined by equations (6) through (12). The expressions for the fringing and mutual capacitances (where ε_k is the permittivity and β an empirical constant for the technology) were originally presented in [4]. These are accurate to over 90% when the following inequalities are satisfied.

$$0.3 < (w/h) < 30, \quad 0.3 < (t/h) < 10, \quad 0.3 < (s/h) < 10$$

For DSM, typical geometries are well within this range. It must also be borne in mind that the delay expressions we give here are only valid for lossy lines, where the inductance has a minor effect and the delay is appreciably greater than the time of flight delay at the speed of light. Typically interconnects in DSM are such lines, where the line behaviour is diffusive in nature even for very fast rise times.

$$R = \rho\frac{l}{hw} \quad (6) \qquad C_f = \varepsilon_k\left[0.075\left(\frac{w}{h}\right) + 1.4\left(\frac{t}{h}\right)^{0.222}\right]l \quad (7)$$

$$C_f' = C_f\left[1 + \left(\frac{h}{s}\right)^{\beta}\right]^{-1} \quad (8) \qquad C_p = \varepsilon_k\frac{wl}{h} \quad (9)$$

$$C_{s,mid} = C_p + 2C_f' \quad (10) \qquad C_{s,corn} = C_p + C_f + C_f' \quad (11)$$

$$C_c = C_f - C_f' + \varepsilon_k\left[0.03\left(\frac{w}{h}\right) + 0.83\frac{t}{h} - 0.07\left(\frac{t}{h}\right)^{0.222}\right]\left(\frac{h}{s}\right)^{1.34}l \quad (12)$$

From the geometry of Fig. 1, we get the following relation.

$$W_T = NW + (N-1)S \quad (13)$$

Our problem definition is, for a constant width W_T, what are the N (number of conductors), S (spacing between conductors), and W (width of the conductors) values that give the optimum bandwidth. Of the three, only two are independent, as the third is defined by (13) for any values that the other two may take. We choose to vary N and S. The variables are discrete as S and W are dictated by the process as well, and there are geometrical limits which cannot be exceeded. As a first example, given in Fig. 3 is a plot of

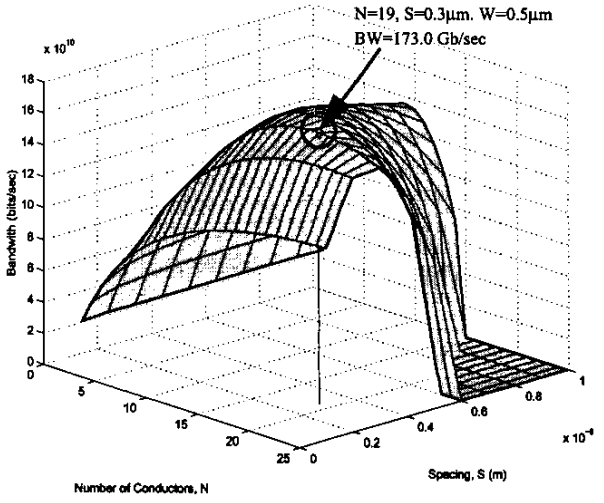

Figure 3. Variation of Bandwidth with N and S

1. Considering the different delays on the corner conductors may be an unnecessary refinement for certain applications where the same signaling frequency is used on all the lines.

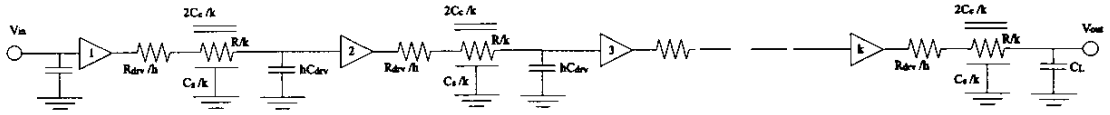

Figure 4: Repeater Insertion in a long interconnect

the total bandwidth varying with N and S for a width of 15μm in a representative 50nm technology. Copper wires are assumed with β=1.65, h=0.2μm, and t = 0.21μm. The minimum wire width and spacing are each assumed to be 0.1μm. It can be seen that there is a clear optimum which does *not* translate to the maximum parallelism possible under the technology constraints.

3. REPEATER INSERTION

The above analysis considered only the intrinsic delay of the lines. In practice the source will be a MOS driver (usually an inverter) with a considerable output impedance. Also to reduce delay the long lines in Fig. 1 are broken up into shorter sections, with a repeater driving each section. The analysis for repeater insertion is carried out by characterizing the non-linear buffers (inverters) by an output resistance R_{drv} and input capacitance C_{drv}. If the number of repeaters including the original driver is k, and the size of each repeater is h times a minimum sized inverter (all lines are buffered in a similar fashion), the output impedance of an h sized driver becomes $R_{drv,m}/h$, and the output capacitance $h*C_{drv,m}$ where $R_{drv,m}$ and $C_{drv,m}$ are the output resistance and input capacitance of a minimum sized inverter respectively in that particular technology. This configuration is sketched out in Fig. 4, where the symbol ≈ refers to a capacitively coupled interconnect as shown in Fig. 1. A complete analysis is given in [1] and the total delay is as defined by (14) where t_r refers to the rise time of the signal (it is assumed the rise times are typically much less than the delay of the line). Additionally the optimum k and h for minimising delay are defined by (15) and (16).

$$t_{0.5} = k\left[0.7\frac{R_{drv_m}}{h}\left(\frac{C_s}{k}+hC_{drv_m}+2.2\frac{2C_c}{k}\right)+\frac{R}{k}\left(0.4\frac{C_s}{k}+1.51\frac{C_c}{k}+0.7hC_{drv_m}\right)\right]+\frac{t_r}{2} \quad (14)$$

$$k_{opt} = \sqrt{\frac{0.4RC_s+1.5RC_c}{0.7R_{drv_m}C_{drv_m}}} \quad (15)$$

$$h_{opt} = \sqrt{\frac{0.7R_{drv_m}C_s+3.1R_{drv_m}C_c}{0.7RC_{drv_m}}} \quad (16)$$

Now using (14), the total delay which includes the delay of the repeaters can be calculated. It must be mentioned here that it is possible to pipeline the bits with a bit per section. In fact it is possible to obtain a gain in the bandwidth by introducing repeaters to pipeline bits, even when there is no significant improvement in the overall line delay. We choose to ignore pipelining in this particular analysis. If pipelining is carried out, the bandwidth is simply multiplied by the appropriate coefficient.

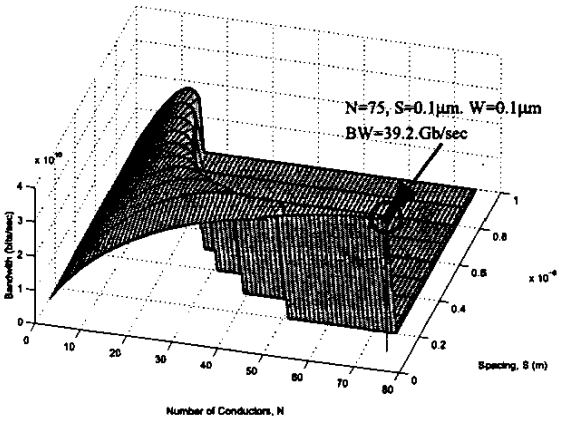

Figure 5: Variation of Bandwidth with Optimal Buffering

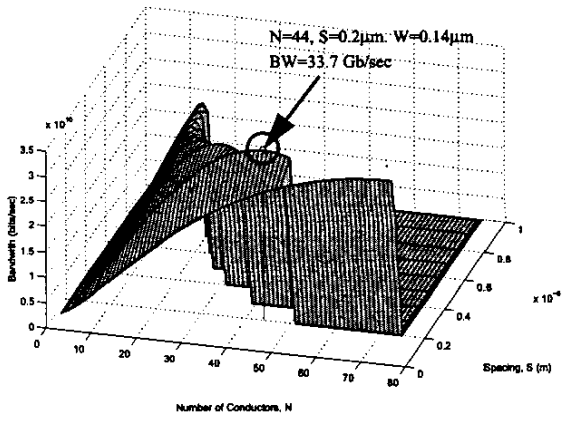

Figure 6: Variation of Bandwidth with a Fixed Number and Size of Buffers for Each Line

For the same boundary conditions as those corresponding to the plot in Fig. 3, the total bandwidth for changing N and S where the repeaters are optimally sized is plotted in Fig. 5. It can be seen that the maximum bandwidth is obtained when the parallelism is the maximum allowed by the physical constraints of the technology. This result is

logical because the buffers which are optimally sized for each configuration compensates for the increased resistance and cross-talk effect. However optimal repeater insertion results in a large number of huge buffers. Also it has been shown in [1] that the delay curve is quite flat, and the sizes can be reduced with little increase in delay. Instead of optimal repeater insertion, if a constraint is imposed on the number and size of buffers for each line, the optimal configuration does *not* equate to the maximum number of wires. Given in Fig. 6 is a plot of the bandwidth when a constraint of k=1 and h=20 is laid down for each line. The optimal configuration corresponds to N=44, so that the Nhk product is 880.

Typically the constraint would be on the total area occupied by the buffers, and hence k and h would be affected by N. If (18) describes the area constraint on the buffers, the optimum configuration is the solution to the constrained optimisation problem of maximising (5) subject to (18).

$$Nkh \leq A_{max} \tag{18}$$

This adds a third independent variable to the objective function (5), of either k or h since A_{max} is a constant. It is a simple matter to incorporate all the relevant equations presented here into an iterative algorithm that can be used to obtain a computer generated solution. As a final example, assume that A_{max} is set to 500 for the same boundary conditions. It turns out that the optimal configuration is when k=1, and shown in Fig. 7 is a plot of the bandwidth where k=1 and h changes according to N.

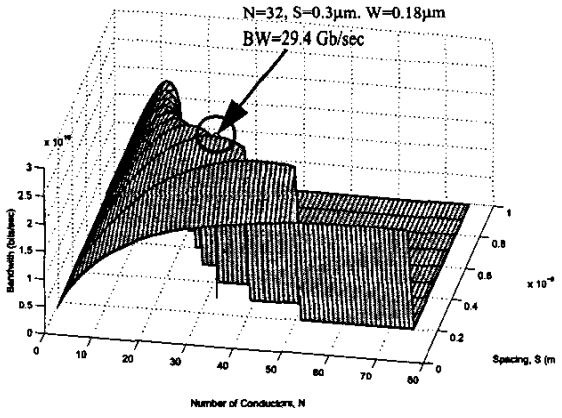

Figure 7: Variation of Bandwidth with Constrained Buffering

4. CONCLUSIONS

In this paper we have carried out an analysis of delay with worst-case cross-talk in capacitively coupled lossy interconnects. In the important configuration of a large number of parallel lines, several factors combine to affect the delay in various ways. Increased parallelism is desirable in general, but when the total area that is allowed for the wires is constrained, this results in smaller, more tightly coupled wires, causing greater line delay. Repeater insertion and especially area constrained repeater insertion further complicates the issue. However we have demonstrated models and a method of building an objective function that takes all these factors into account and predicts the optimum configuration. Additional considerations such as native word width constraints can easily be incorporated. We propose these equations and the method of analysis as being suitable for calculations early in the design flow, as the simplicity of the expressions allow for a large number of iterations.

5. ACKNOWLEDGEMENTS

Productive discussions with Johnny Öberg, Axel Jantsch and Mikael Millberg which helped identify the requirements from a systems perspective are gratefully acknowledged.

6. REFERENCES

[1] D. Pamunuwa and H. Tenhunen, "On dynamic delay and repeater insertion in distributed capacitively coupled interconnects" in *Proc. ISQED*, Mar. 2002.

[2] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron II: a global wiring paradigm", in *Proc. ISPD*, 1999, pp. 193-200.

[3] H. Tenhunen and D. Pamunuwa, "On dynamic delay and repeater insertion", in *Proc. ISCAS*, May 2002.

[4] L-R. Zheng, D. Pamunuwa and H. Tenhunen, "Accurate *a priori* signal integrity estimation using a dynamic interconnect model for deep submicron VLSI design", in *Proc. ESSCIRC*, Sept. 2000, pp- 324-327.

[5] J. Rubinstein, P. Penfield and M. Horowitz "Signal delay in RC tree networks", *IEEE Trans. Computer Aided Design*, vol CAD-2, no. 3, pp. 202-211, July 1983.

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

[7] Y. Ismail and E. Friedman, "Effects of inductance on the propagation delay and repeater insertion in VLSI circuits", *IEEE Trans. VLSI Systems*, April 2000, vol. 8, pp. 195-206

[8] S. Dar, M. Franklin, "Optimum buffer circuits for driving long uniform lines", *IEEE J. Solid State Circuits*, vol. 26, pp. 32-40, Jan. 1991.