INTERCONNECTION DELAY IN VERY HIGH-SPEED VLSI

D. Zhou, F. P. Preparata, and S. M. Kang

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

Abstract

Interconnection delay of VLSI has been a major bottleneck in the design of high-speed digital systems. This problem becomes even more pronounced as the minimum feature size is scaled down to the submicron level, and the transistor's intrinsic switching time is decreased well into the sub-nanosecond range [1-6]. This paper presents a rigorous analysis of the signal propagation delay in VLSI circuits. We show that wire inductance is an important factor in determining the delay when the delay is in the sub-nanosecond range and that this delay cannot be well estimated by the conventional modeling of the interconnection by *RC* circuits [7-10]. The effect of feature size scaling on circuit performance is also evaluated, and the inherent limitation on scaling due to interconnection delay is discussed. Finally, the relationship between the delay and the technology parameters, such as wire resistance, wire capacitance, wire inductance, wire width, wire length, and load capacitance is investigated to provide guidelines for the interconnection technology development.

Introduction

When the feature size of the interconnection wires is scaled down into the submicron range and the interconnection delay is in the sub-nanosecond range, the lumped *RC* model becomes inadequate and the transmission line property of interconnections must be taken into account [6]. Furthermore, since the *I–V* characteristics of the driving transistor play an important role in this issue, the conventional transmission line analysis with an ideal step input voltage fails to provide accurate information. In this paper we will analyze the interconnection delay by using a distributed *RLC* system with a transistor as the driver and a capacitor as the load.

Circuits Model and Its Solution

Figure 1 shows a basic circuit model for interconnection analysis. A similar model can be used when a p-channel transistor pulls up the drain node. Initially, the n-channel transistor T_1 is cutoff, and its drain load, the wire AB, and the gate capacitance of the driven transistor T_2 are charged to voltage V_0. The *I–V* characteristic curve of T_1 is modeled by a piecewise linear function (Figure 2), the gate load by a capacitor C_g, and the wire *AB* by a distributed RLC transmission line (Figure 3). The gate of T_1 is assumed to be driven by a step input, which is more realistic and quite different from the case with a step input at node *A* as considered by many previous researchers. In this paper the time interval in which the line voltage $V(x=l)$ drops from V_0 to $0.1V_0$ will be defined as the signal propagation delay along wire *AB*. According to the driver's *I–V* characteristic curve, the delay time can be divided into two segments. The first segment is the time interval τ_1 during which $V(x=0)$ drops from V_0 to $0.8V_0$, and the second segment is the time interval τ_2 which is defined as starting at the termination of τ_1 and ending at the point at which $V(x=l)$ drops to $0.1V_0$, as shown in Figure 4. The driving transistor operates in the saturation region during τ_1 and in linear region during τ_2.

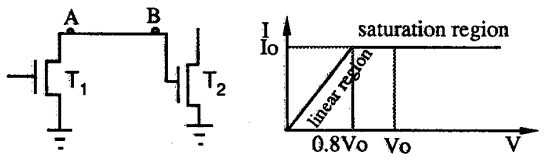

Figure 1

Figure 2

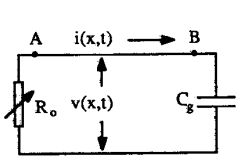

Figure 3

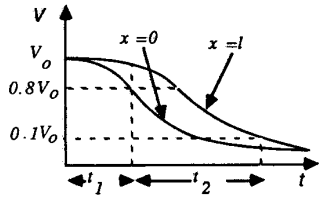

Figure 4

Let R be the resistance, L the inductance, and C the capacitance per unit length of the wire *AB*, and denote the voltage and the current at distance x from node *A* by $v(x,t)$ and $i(x,t)$ respectively. Then, from Kirchoff's laws, we obtain the following partial differential equations:

$$\frac{\partial v}{\partial x} + L\frac{\partial i}{\partial t} + Ri = 0 \tag{1}$$

$$\frac{\partial i}{\partial x} + C\frac{\partial v}{\partial t} = 0 \tag{2}$$

By substituting equation (1) into (2), we obtain the well-known telegraph equation

$$\frac{\partial^2 i}{\partial x^2} = LC\frac{\partial^2 i}{\partial t^2} + RC\frac{\partial i}{\partial t} \tag{3}$$

Two sets of boundary and initial conditions will be considered. The first case is when the transistor T_1 works in the saturation region, and the second one is when T_1 works in the linear region. In the first case, the initial conditions are

$$v(x,0) = V_0 \qquad 0 \le x \le l \tag{4}$$

$$i(x,0) = 0 \qquad 0 \le x \le l \tag{5}$$

$$\frac{\partial i(x,0)}{\partial t} = 0 \qquad 0 \le x \le l \tag{6}$$

and the boundary conditions are

$$i(0,t) = -I_0 \quad \text{and} \quad i(l,t) = C_g\frac{\partial v(l,t)}{\partial t} \tag{7}$$

The solutions for this initial and boundary condition are

† This research was supported by National Science Foundation Engineering Research Center for Compound Semiconductor Microelectronics contract NSF CDR 85-22666

$$i(x,t) = \sum_{n=1}^{k} M_n \sin\lambda_n x \left[\frac{e^{\zeta_{1n} t}}{1-\frac{\zeta_{1n}}{\zeta_{2n}}} + \frac{e^{\zeta_{2n} t}}{1-\frac{\zeta_{2n}}{\zeta_{1n}}} \right]$$

$$+ \sum_{n=k+1}^{\infty} M_n \sin\lambda_n x \; e^{\alpha t} \left[\cos\beta_n t - \frac{\alpha}{\beta_n} \sin\beta_n t \right] + I_0 \left[\frac{x}{l+\frac{C_g}{C}} - 1 \right] \quad (8)$$

and

$$v(x,t) = V_0 - \frac{1}{C} \left\{ \sum_{n=1}^{k} \lambda_n M_n \cos\lambda_n x \left[\frac{e^{\zeta_{1n} t} - 1}{\zeta_{1n} \left[1-\frac{\zeta_{1n}}{\zeta_{2n}} \right]} + \frac{e^{\zeta_{2n} t} - 1}{\zeta_{1n} \left[1-\frac{\zeta_{2n}}{\zeta_{1n}} \right]} \right] \right.$$

$$+ \sum_{n=k+1}^{\infty} \frac{\lambda_n M_n \cos\lambda_n x}{\alpha^2+\beta_n^2} \left[e^{\alpha t} \left(\alpha\cos\beta_n t + \beta_n \sin\beta_n t \right) - \alpha \right.$$

$$\left. \left. - \frac{\alpha}{\beta_n} \left[e^{\alpha t} (\alpha\sin\beta_n t - \beta_n \cos\beta_n t) + \beta_n \right] \right] + \frac{I_0 t}{l+\frac{C_g}{C}} \right\} \quad (9)$$

where α and β are constants, and λ_n, ζ_{1n}, ζ_{2n}, M_n, and N_n are coefficients of the above expansion series[1]. From equation (10), the first time segment τ_1 is found by setting $v(0,\tau_1)=0.8V_0$.

To find the solution in the linear region, we move the time origin to τ_1 and use the boundary conditions

$$i(0,t) = -\frac{1}{R_0} v(0,t) \quad \text{and} \quad i(l,t) = C_g \frac{\partial v(l,t)}{\partial t} \quad (10)$$

where R_0 is the output resistance of transistor T_1. The initial conditions are obtained by setting $t=\tau_1$ in solutions (9) and (10):

$$i(x,0) = \sum_{n=1}^{\infty} H_n \sin\lambda_n x + I_0 \left[\frac{x}{l+\frac{C_g}{C}} - 1 \right] \quad (11)$$

and

$$v(x,0) = \sum_{n=1}^{\infty} G_n \cos\lambda_n x + G_0 \quad (12)$$

where H_n and G_n are coefficients of expansion series. We use Laplace transform method to solve this problem. Let $V(x,s)$ and $I(x,s)$ be the Laplace Transform of $v(x,t)$ and $i(x,t)$ with respect to time t, respectively. Then we can obtain

$$V(x,s) = \cosh(h(s)x)V(0,s) - Z(s)\sinh(h(s)x)I(0,s) + A_1(x,s) \quad (13)$$

$$v(l,t) = \sum_{i=1}^{\infty} Res\left[V(l,s)e^{st}, s_i\right] \quad (14)$$

where $Res[V(l,s)e^{st}, s_i]$ is the residues of function $V(l,s)e^{st}$ at the pole s_i, and $A_1(x,s)$, $h(s)$, and $Z(s)$ are functions in s domain[1]. The second time segment τ_2 is then calculated from (19) by setting $v(l,\tau_2) = 0.1V_0$. The total delay τ is the sum of τ_1 and τ_2, to which the major contribution comes from τ_2.

Electrical Performance of Interconnection

We now investigate the relationship between the delay time and the technology parameters. The technology parameters can be grouped into two sets; (1) a set of electrical parameters such as wire resistance R, capacitance C, inductance L, and (2) a set of geometrical parameters such as wire width w, wire thickness d, and insulation layer thickness t_{ox}, as illustrated in Figure 5.

[1] The exact expressions of these coefficient are omitted for brevity.

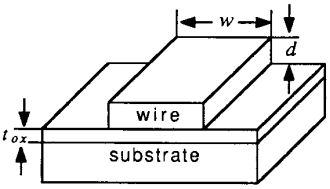

Figure 5

The following formulas, taken from [7], will be used to relate electrical parameters to geometrical parameters.

$$C = \varepsilon_{ox} \left\{ 1.15 \left[\frac{w}{t_{ox}} \right] + 2.80 \left[\frac{d}{t_{ox}} \right]^{0.222} \right\} \quad (15)$$

$$L = \frac{1}{C v_c^2} \quad (16)$$

and

$$R = \frac{\rho}{wd} \quad (17)$$

where v_c is the speed of light in the insulation material, and ρ is the resistivity of the interconnection material. We assume that uniform scaling [8] (sometimes called *ideal* scaling) is used when the wire width w is smaller than $1\mu m$. But, based on practical chip design consideration, we use $t_{ox}=1\mu m$ and $d = 1\mu m$ when w is larger than $1\mu m$. For a qualitative analysis of the interconnection problem, we divide the interconnection distance into three ranges: short (or local) distance for $0<l\leq 0.05cm$, medium distance for $0.05cm<l\leq 0.5cm$ and long distance for $0.5cm<l$. Typical parameter values used in the calculation are listed in Table 1. To study the sensitivity to a specific electrical parameter, we vary the parameter in question while all others are kept constant unless otherwise noted.

PARAMETERS	NOMINAL VALUES
gate capacitance: C_g	$7.2\times10^{-3}pf$
resistivity of Al: ρ	$.28\times10^{-7}\Omega m$
channel resistance: R_0	$4k\Omega$
saturation current: I_0	1ma
operation voltage: V_0	5v
length of wire: l	0.5cm
width of wire: w	$2\mu m$
thickness of oxide: t_{ox}	$1\mu m$
thickness of wire: d	$1\mu m$
thickness of substrate: d_s	$200\mu m$

Table 1

The numerical results based on our analytic solution are shown in Figures 6 through 11 (solid curves). For comparison, the dashed curves (some of them are coincident with the solid curves) of Figures 7 through 11 are calculated from the following empirical formula of [2].

$$\tau \approx (2.3R_0 + Rl)Cl = \left(\frac{2.3R_0}{Rl} + 1\right)RCl^2 \quad (18)$$

It can be seen that (18) is a good approximate formula when the delay is longer than one nanosecond where the transmission line effect is insignificant.

1. Delay vs. Wire Inductance (L)

The effect of L on the delay is plotted in Figure 6. When the interconnection delay is longer than $1ns$, it is almost unaffected by L. However, for delays in the sub-nanosecond range, L becomes an important factor. In the sub-nanosecond range, the problem of reducing reflections at both driving and receiving ends becomes important due to the significance of transmission line effects. Furthermore, the poor magnetic conductivity of the substrate material, which causes the "slow wave" phenomenon [6], deserves special attention in design.

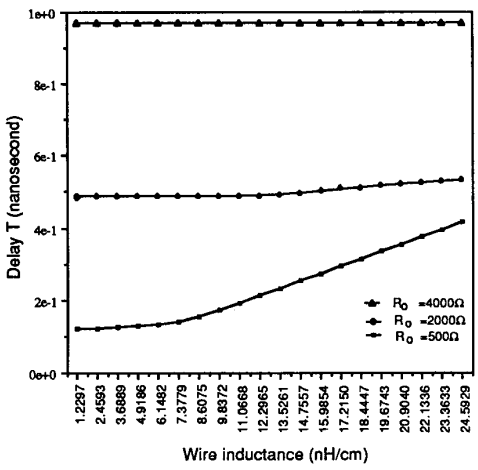

Figure 6

2. Delay vs. Wire Resistance (R)

In Figure 7, the delay time is plotted as a function of the wire resistance R ranging from $10^2\Omega/cm$ for gold wire to $2.4\times10^5\Omega/cm$ for polysilicon wire. It is clear that, in the medium and long distance range, larger resistance R will cause sharp increase in delay time. Even in local interconnection, R greater than $10^4\Omega/cm$ (polysilicon wire in $1\mu m$ technology) causes more than a one nanosecond delay, which requires special attention in submicron VLSI with the clock frequency higher than one gigahertz. Also, it can be observed that the delay time is not very sensitive to the metal wire resistance R in this short distance range.

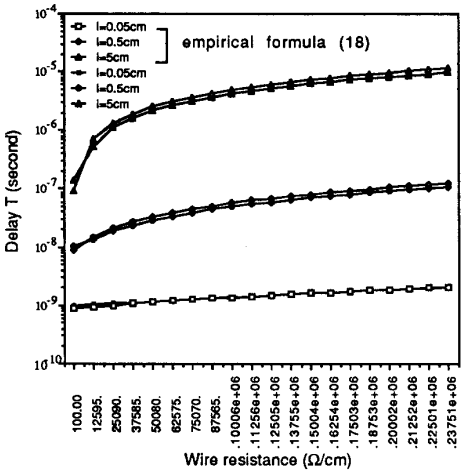

Figure 7

3. Delay vs. Wire Capacitance (C)

Figure 8 shows the strong dependency of propagation delay on the capacitance as reported by other researchers [1]. In our example, for incremental capacitance of $10fF/cm$, the delay time increases from $0.0054ns$ in the short distance range, to $0.054ns$ in the medium distance range and to $0.54ns$ in the long distance range. Clearly, the wire capacitance C plays a dominant role in controlling the delay, which suggests that for long interconnections narrowing the wire width can be an efficient speed up technique provided that the total wire resistance Rl is relatively smaller than R_0.

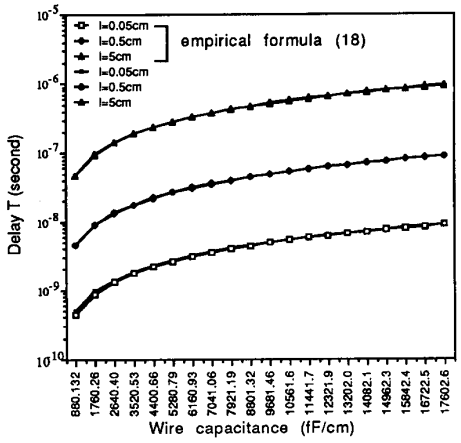

Figure 8

4. Delay vs. Driver Output Resistance (R_0)

Perhaps one of the most effective way to shorten the propagation delay is to increase the power of the driver, i.e., to reduce the output resistance R_0 of the driving transistor. The delay vs. R_0 curve is shown in Figure 9. The channel resistance should be less than 500Ω to achieve a sub-nanosecond delay in the medium distance range. This is exactly the case in nominal GaAs transistor [5,6]. In a typical 1 μm MOS technology, the ratio of channel width to length of the driving transistor should be about 20 to achieve such R_0. To achieve a nanosecond propagation delay in the long distance range, the driver's output resistance should be less than 200Ω , which corresponds to a channel width to length ratio bigger than 50.

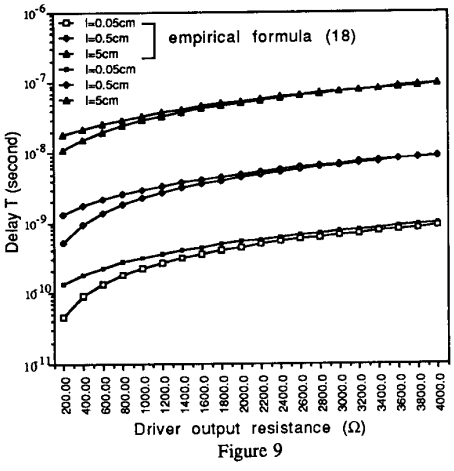

Figure 9

5. Delay vs. Wire Length (l)

Figure 10 shows the delay as a function of wire length l. From the short distance range to the long distance range, the relationship is almost linear. Physically, this means that the total resistance of the wire is much smaller than driver's output resistance. In this situation, the delay is determined mainly by the total wire capacitance Cl and driver's output resistance R_0. As a result, the delay of the wire with larger R will exhibit a more markedly nonlinear behavior. This fact has been shown clearly in Figure 10, where the delay of the wire $0.5\mu m$-wide increases faster than that of the wire $1\mu m$-wide. But the rate at which the delay increases should not be confused with the delay itself. In fact, the delay of the wire $5\mu m$-wide is still smaller than that of other wider wires ($w=1\mu m$, $w=2\mu m$, and $w=4\mu m$) in the wire length range $l\leq8.5cm$, which covers almost all practical on-chip interconnection distances.

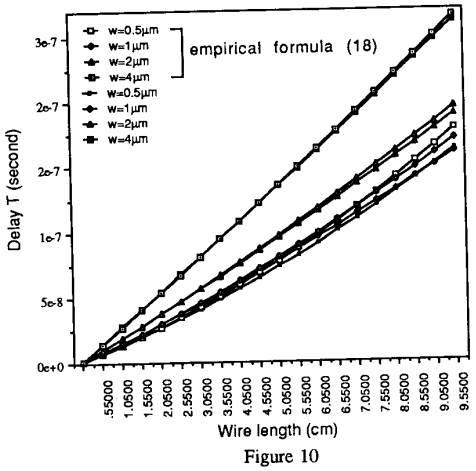

Figure 10

6. Delay vs. Wire Width (w)

Figure 11 shows the behavior of the delay as a function of the wire width. A critical wire width w_0 can be defined to express the influence of scaling on interconnection delay. The delay will increases drastically as w becomes smaller than w_0. With this in mind, we define w_0 as the wire width beyond which further reduction of the width increases the delay by at least $0.5dB$ for every $0.1\mu m$ reduction in w. Using this criterion, w_0 is about $0.3\mu m$ for Al, $1.1\mu m$ for silicide, and $1.6\mu m$ for polysilicon, respectively, under the hypothesis of uniform scaling. But, as mentioned earlier, the wire thickness d and the insulation layer thickness t_{ox} for practical chip design considerations are set to be $1\mu m$ when w is larger than $1\mu m$. Therefore for both silicide and polysilicon wires w_0 is conventionally equal to $1\mu m$ (Figure 11). This result shows that, for submicron processes, non-metal wire connections should be ruled out, except for very short distance communications. The behavior of the delay curve for $w \leq w_0$ can be explained by the following facts: the wire capacitance C and the wire resistance R make major contributions to the delay, and they change at different rates according to the wire width. Because of the fringe effect and the ideal scaling, the rate at which C decreases is smaller than the rate at which R increases. Consequently, the RC delay of the wire increases as the wire dimension is scaled down. On the other hand, the delay decreases slightly, or even increases as the wire width increases for $w \geq w_0$ due to the fact that the driver's output resistance R_0 and wire capacitance C control the delay. In this case, the wire resistance is not as important as the driver's output resistance, and hence the increment of C can result in an increase of the delay.

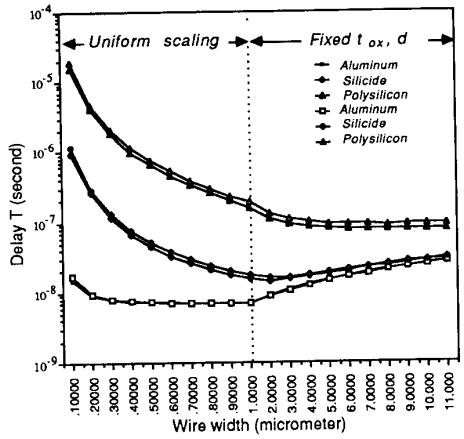

Figure 11

Summary

Some of our significant findings are now summarized as follows:

(1) The wire inductance L becomes prominent in the sub-nanosecond range, and therefore L should be taken into account in design of sub-nanosecond systems.

(2) The interconnection delay will be the bottleneck in the speed performance of high speed VLSI; it is more than hundreds of picoseconds even in the local distance range ($l \leq 500\mu m$) while high speed devices can provide a switching speed on the order of ten picoseconds.

(3) The interconnection delay places an inherent limitation on scaling, as had been qualitatively noted more than a decade ago [8]. From the analytic solution of RLC distributed model of this paper, it is demonstrated that for a given interconnection distance there exists a typical wire width below which further scaling drastically increases the propagation delay.

(4) The relationship between delay and interconnection wire length is nearly linear for metal wire in many practical cases ($l \leq 5cm$). The primary factors involved in determining the delay for these cases are the capacitance of the wire and the output resistance of the driver. For polysilicon wire, the resistance R becomes an important factor, and non-linear behavior is exhibited.

(5) The driving ability of the driver plays a principal role in determining the interconnection delay. Any incorrect assumptions on the behavior of the driver, such as the step input assumption, can lead to invalid conclusions.

ACKNOWLEDGMENTS

The authors wish to thank D. S. Gao and S. W. Hornick for their valuable comments.

REFERENCES

[1] G. Bilardi, M. Pracchi, and F. P. Preparata, "A critique of Network Speed in VLSI Models of Computation," IEEE J. Solid-State Circuits, Vol. SC-17, No.4, pp. 696-702, Aug., 1982.

[2] H. B. Bakoglu, and J. D. Meindl, "Optimal Interconnection Circuits for VLSI," IEEE Trans. Electron Devices, Vol. ED-32, No.5, pp. 903-909, May 1985.

[3] K. C. Saraswat and F. Mohammadi, "Effect of Scaling of Interconnections on the time delay of VLSI circuits," IEEE J. Solid-State Circuits, Vol. SC-17, pp. 442-448, June 1982.

[4] H. Hasegawa, M. Furukawa, and H. Yanai, "Properties of Microstrip lines on $Si-SiO_2$ System," IEEE Trans. Microwave Theory and Techniques, Vol. MIT-19, No. 11, pp. 869-881, Nov., 1971.

[5] V. Molutinovic, "GaAs Microprocessor Technology," Computer, Vol. 19, No. 10, Oct., 1986.

[6] B. K. Gilbert et. al., "Signal Processors based upon GaAs ICs: The Need For a Wholistic Design Approach," Computer, Vol. 19, No. 10, Oct., 1986.

[7] T. Sakurai and K. Tamaru, "Simple Formulas for Two- and Three-Dimensional Capacitances," IEEE Trans. Electron Devices, Vol. ED-30, No. 2, pp. 183-185, Feb., 1983.

[8] R.H. Dennard et. al., "Design of ion implanted MOSFET's with very small physical dimensions," IEEE J. Solid-State Circuits, SC-9, pp. 256-268, 1974.

[9] T Sakurai, "Approximation of Wiring Delay in MOSFET IC LSI," IEEE J. Solid-State Circuits, Vol. SC-18, pp. 418-426, Aug. 1983.

[10] H. B. Bakoglu, J. T. Walker and J. D. Meindl, "A Symmetric Clock-Distribution Tree and Optimized High-Speed Interconnections for Reduced Clock Skew in ULSI and WSI Circuits," Proceeding of International conference on Computer Design, pp.118-122, October 1986.