Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

Estimation of Delay Due to Overshoot Effect for CMOS Gates in Binary-Tree Timing Simulation

Molin Chang, Shuih-Jong Yih and Wu-Shiung Feng

Department of Electrical Engineering, R244
National Taiwan University
Taipei, Taiwan, R.O.C.

Abstract

A switch-level timing simulator has the advantage of fast speed and good adaptability for VLSI circuit, but it can not offer more accurate transient waveform information. A new approach for delay estimation is presented which is achieved by two equations : dominant delay equation and error delay equation. Both are derived by surface fitting to approximate the surface that is measured from the actual delay behavior of a CMOS gate.

1. Introduction

The switch-level simulation is a compromised method between circuit-level and logic-level simulations. It has the advantage of fast speed and good adaptability for VLSI circuit. Some switch-level simulators have been implemented, such as MOSSIM, RSIM [1], [2]. BTS (Binary tree Timing Simulator) is a switch-level timing simulator based on series-parallel circuit structure [3], which is two to three orders of magnitude faster than SPICE and has more accurate waveform approximation during the transient state.

Most switch-level algorithms emphasized how to calculate the time constant of charging/discharging the load capacitance more accurately. There are many researches on this topic [4], [5], [6], [7]. However, all of them can not offer us more accurate waveform information in the transient state; we want to know not only whether the logic gate changes state or not, but also when the output voltage begins to change and how fast it will change.

The high accuracy of BTS is attributed to a new waveform approximation technique, which includes delay estimation and slope estimation. The *delay* estimation tells us when the output begins to change and the *slope* estimation tells us how fast the output will change. In the delay estimation, the overshoot appearing on the output node is the thing we must consider. In the actual circuits, an uncertain amount of overshoot, chiefly due to parasitic capacitors, will almost always be produced at the output node while an event is happening at the input. If the width of overshoot can be predicted well, and then the delay will be estimated accurately. Second, the slope relates closely to the RC time constant of the discharging/charging path; the relationship between them is only a constant multiplication. Lin and Mead proposed an efficient method that can be implemented by recursive way [6].

In this paper, first, we describe the concept of series-parallel tree in section II. Next, in section III, we introduce the model used in BTS. The waveform approximation technique is presented in section IV. Then in sections V and VI the delay and slope estimation are stated in more detail, respectively. Finally, the simulation results are shown in section VII.

2. Series-parallel tree

If a gate circuit (also known as *cluster*) is connected by series-parallel way such as fully complementary CMOS, Pseudo-NMOS, Dynamic CMOS and so on, it can be represented as a merged series-parallel tree. A merged tree, also called a PN tree, consists of two series-parallel trees, which are the left subtree (P tree)

and the right subtree (N tree). When it is mapped to a gate circuit, the left subtree and the right subtree represent the *pull-up* and *pull-down* subcircuits, respectively. Fig. 1(b) illustrates the corresponding PN tree of the gate circuit as shown in Fig. 1(a) whose function is

$$Z=(a \cdot ((b \cdot c)+(d \cdot e))) \cdot f \quad (1)$$

In order to represent the non-complementary CMOS circuits, a new function expression is used. Its general form is **Z=(the function of P tree) #(the function of N tree)**, where # represents the root node of a PN tree. For fully complementary CMOS, the function of P tree is just a complement of that of N tree. Then Eq. 1 can be rewritten as follows:

$$Z=((a+((b+c)*(d+e)))+f)\#((a*((b*c)+(d*e)))*f) \quad (2)$$

where * and + represent an AND and an OR operations, respectively.

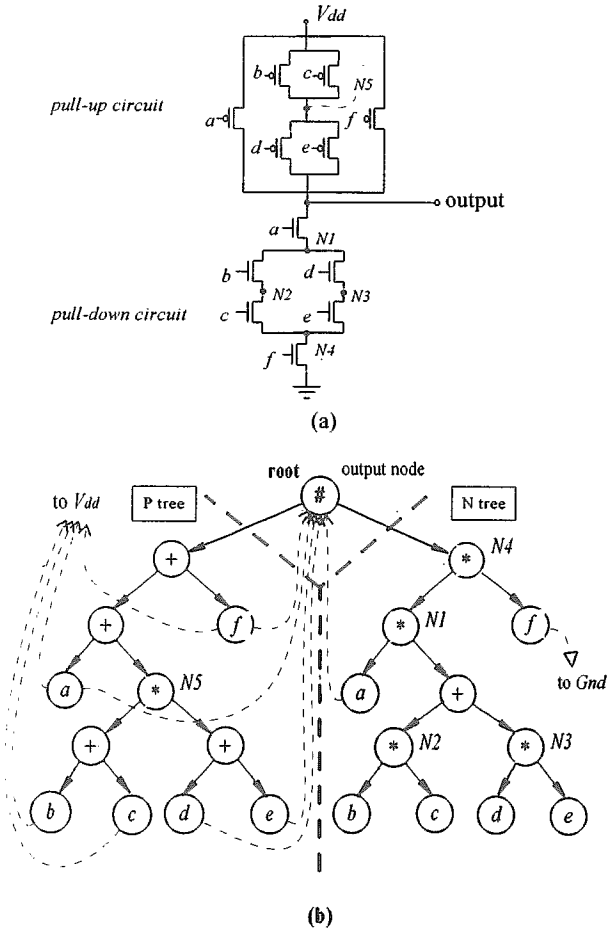

Figure 1: CMOS complex gate. (a) A gate circuit, (b) the equivalent merged series-parallel tree.

The equivalent PN tree of each cluster is established once at the beginning of simulation. The calculations of the equivalent resistances, the equivalent RC time constant and the charge sharing effect of a circuit can be solved efficiently because the methods are based on the structure of series-parallel tree.

3. MOS model

The MOS model in BTS is composed of voltage-controlled switch, effective resistance R_{eff} and equivalent grounded capacitances, as shown in Fig. 2. The transistor is *on* (the switch conducts) if and only if the gate voltage of the NMOS transistor is higher than its threshold voltage V_T. The turn-on effective resistor is distinguished by two cases: R_{on} (in the steady state) and R_t (in the transient state), because MOS transistor (denoted by MOST) has different response under different gate state. Therefore, the value of R_{eff} may be one of the three cases [3]:

$$R_{eff} = \begin{cases} \infty & \text{if } V_g < V_T \\ R_{on} & \text{if } V_g \text{ is high (steady state)} \\ R_t & \text{if } V_g \text{ changes from L to H (transient state)} \end{cases} \quad (3)$$

The values of R_{on} and R_t depend on the physical parameters and the load capacitance, and R_t depends also on the slope of the signal at the gate. In the actual implementation, we maintain two tables in our program, which are two-dimensional R_{on}-table and three-dimensional R_t-table,

$$R_{on} = f(W/L, C_{load}) \quad (4)$$

$$R_t = f(W/L, C_{load}, slope) \quad (5)$$

Because the load capacitance of a gate circuit is the summation of the input capacitances of the fanouts of this gate circuit, C_{load} will be discrete times of the input capacitance of a CMOS inverter with the minimum size (W/L). From the simulated results of SPICE, the relationship between the effective resistance and the slope of the gate signal is smoothly linear, so only a few slopes need be recorded, and the actual value can be obtained using interpolation.

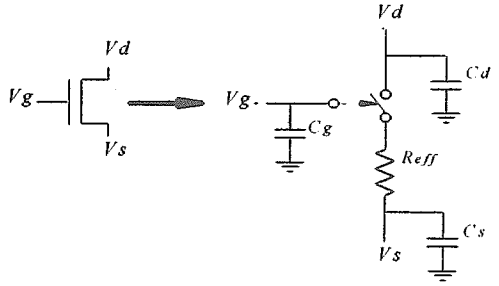

Figure 2: The MOS model used in BTS

4. Waveform approximation

The approximation work can be simplified if we cut the overshoot and use a linear segment followed by an exponential tail to approach the falling (or rising) signal. Fig. 3(a) illustrates this idea, and the solid line (the result of BTS) will fit the SPICE result well if we can calculate the time of point A and the slope of the linear segment between point B and point C. Next, we use two equations as follows to plot the transient waveform [8].

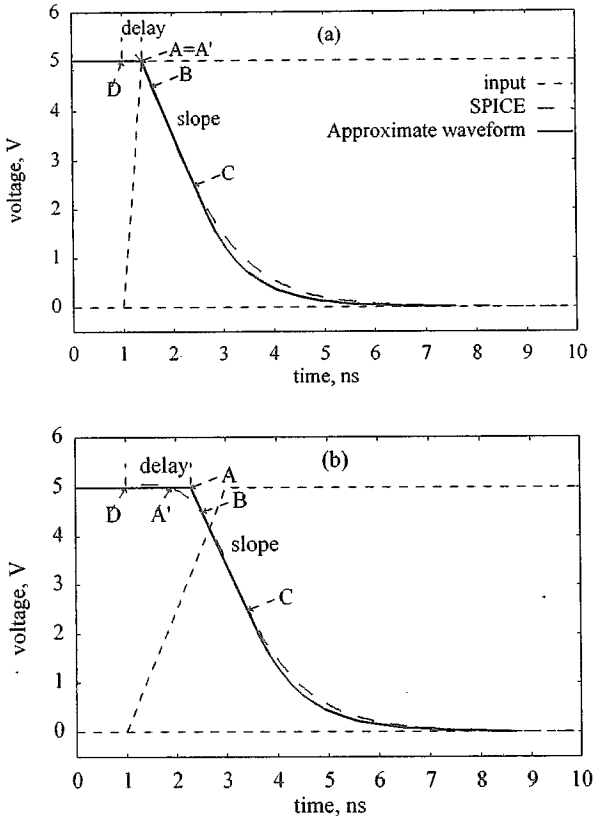

Figure 3 : Waveform approximation by delay and slope estimation method. (a) type A, (b) type B.

$$f=\begin{cases}\dfrac{0.2t}{T} & t<3T \\ 1-0.4\exp(-\dfrac{t-3T}{2T}) & t\geq 3T\end{cases} \quad \textit{for a rising signal} \qquad (6)$$

$$f=\begin{cases}1-\dfrac{0.2t}{T} & t<3T \\ 0.4\exp(-\dfrac{t-3T}{2T}) & t\geq 3T\end{cases} \quad \textit{for a falling signal} \qquad (7)$$

where T is half of the time spent by signal between 90% (for a falling signal) or 10% (for a rising signal) and 50% of the steady state. If the value of T can be obtained, the transient waveform will then be easily plotted.

In general, there are two types of waveforms; one is type A as shown in Fig. 3(a), which has a more near linear segment between point A' and point C, and the other is type B as shown in Fig. 3(b), which has a larger curvature of the curve between both points A' and C. However, for both types we use a linear segment to fit them. Thus, in contrast to the type A whose delay time is almost equal to the width of overshoot, the delay time ($t_A - t_D$) of type B is larger than the actual width ($t_{A'} - t_D$) of the overshoot. Because of this, the definition of delay and slope is not based on the actual waveform, but on the approximate waveform.

Definition1: The difference between the time when the output signal begins to change and the time when the input signal begins to change is defined as *switching delay* or *delay*, which is denoted by **D**.

Definition2: The changing rate after the output begins to change, but is restricted between 90% and 50% of the steady state for a falling signal, is defined as *slope*, which is denoted by **S**.

In Fig. 3(b) the time between point D and point A is the *delay*, and the changing rate between point B and point C is the *slope*. Choose some sample circuits and measure the values of delay time of them from the waveforms resulting from SPICE simulation. Next, model the behavior of delay by several dominant factors such as input slope, C_{load}, and so on. After the model being established, the delay prediction of an arbitrary circuit is possible.

5. Delay estimation

Basing on the definition of delay in BTS, we focus our attention on the overshoot of output. Owing to the electrical characteristics of MOS transistor, there are many parasitic capacitors existing inside a CMOS gate, e.g., C_{gs}, C_{gd} and so on. So the waveform of the drain of a MOST depends not only the turn-on mechanism of MOST but also the path formed by C_{gd}. The overshoot of output waveform, which can be treated as the excessive charge stored in the output node, is caused by the differential gate capacitor current. Observe that the amount of overshoot is determined by four factors as follows:

(1) S_I: the slope of input signal,

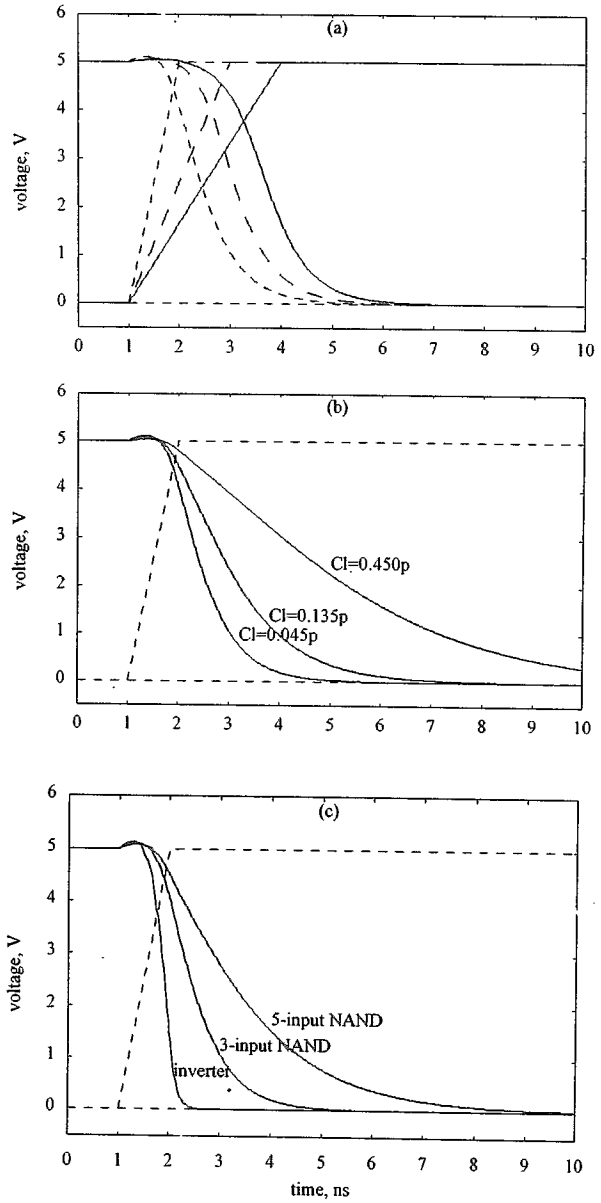

Figure 4 : Simulated waveforms to observe overshoot effect influenced by (a) Si, (b) Cl, and (c) Np

(2) Z_{gd}: the size of C_{gd},

(3) C_l: the load capacitance of output, and

(4) R_p: the resistance of discharging path in the N tree (or charging path in the P tree).

To simplify the problem, we assume that Z_{gd} is fixed, and then observe the overshoot influenced by the other three factors; we show them in Figs. 4(a), 4(b), and 4(c).

The structure and the processing method of N tree and P tree are identical, so only N tree is discussed hereafter. In Fig. 5 we can see that the overshoot is the major factor that produces the error between SPICE and

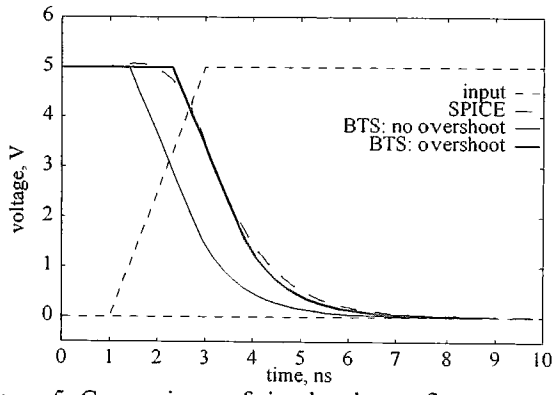

Figure 5: Comparisons of simulated waveforms.

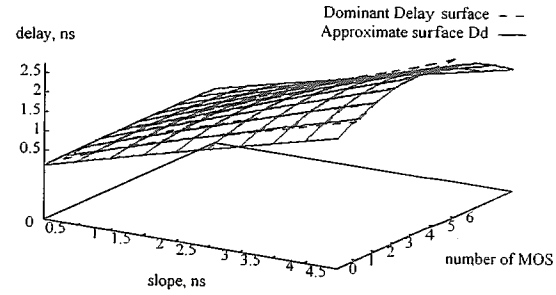

Figure 6: The dominant delay surface and its approximate surface D_d.

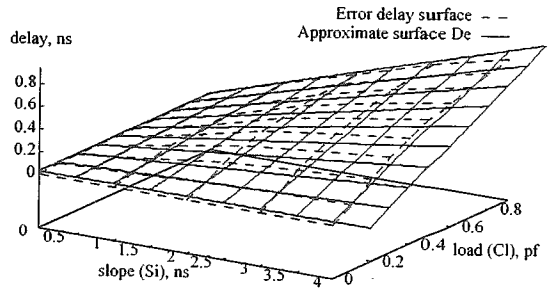

Figure 7: The error delay surface (N_p=1) and its approximate surface D_e.

BTS if we do not consider it into our simulator. In this case, BTS neglects the effect of C_{gd}, and uses the time that input voltage reaches V_T of a MOST as the turn-on time. Therefore, it is necessary to estimate the width of overshoot for obtaining more accurate transient waveform.

By analyzing some sample circuits using SPICE and varying the values of factors as mentioned above, we

measure the data of delay time (not the width of overshoot) and then we can model the delay behaviors of CMOS gates by two equations.

(1) ***Dominant delay equation***: C_l is fixed, so this equation describes the relationship among dominant delay, S_i, and R_p. Changing S_i is easy, but changing R_p is more difficult, because R_p is a discrete value due to the integer number of MOST. Therefore, an alternative method is used. We increase the number of MOST's in N tree circuit in order to change R_p discontinuously, and then R_p can be replaced with N_p. In other words, we use the circuits such as inverter, two-input NAND gate, three-input NAND gate, and so on, as the primitive cases; but only one MOST near the output accepts the input signal, and the others are kept in the turn-on state. The reason of placement of input signal is that the effect of internal charges can be avoided. The effect of internal charges we may meet in the actual circuits are extracted as an independent problem [3]. By analyzing data, we can plot a three-dimensional surface as shown in Fig. 6. To simplify the calculation, we can use a hyperbolic surface (Eq. 8a) to fit it. However, without obviously decreasing the speed, a more accurate approximate surface (Eq. 8b, also shown in Fig. 6) is preferred to obtain a more accurate delay time.

$$\mathbb{D}_d = (0.0292N_p+0.369)(S_i+0.3)+0.12 \qquad (8a)$$

$$\mathbb{D}_d = (-0.023N_p^2+0.19N_p+1.21)(-0.0047S_i^2+0.38S_i+0.91) \qquad (8b)$$

(2) ***Error delay equation***: N_p is fixed, so this equation describes the relationship among error delay (based on $\mathbb{D}_d$), S_i, and C_l. If N_p is adjusted, we obtain a set of surfaces. Similarly, we can also use a set of hyperbolic surfaces

$$\mathbb{D}_e = f(N_p)(0.293S_iC_l+0.023) \qquad (9)$$

to fit them, where $f(N_p)$ represents the coefficients that are the function of N_p. The surfaces when N_p=1 are shown in Fig. 7, which include the surface derived from experimental data and its approximate surface. The error delay is an offset value with respect to the dominant delay. Thus, the total delay is the sum of both.

In BTS the estimation of N_p is not easy in the simulation of real circuits because there may be several discharging paths. Therefore, the quickest and easiest method is calculated by total equivalent resistance R_{peq} of all discharging paths divided by the turn-on-resistance R_{on} of single MOST , i.e. $N_p \cong R_{peq}/R_{on}$. For example, in Fig. 1(a) let $V_b=V_c=0V$ and $V_a=V_d=V_e=V_f=5V$, then $R_{peq}=R_a+R_d+R_e+R_f$. Since there are four turn-on MOST's, N_p is four in this case if all MOST's are identical. In the most cases, N_p will not be exactly an integer because of two reasons : (1) parallel-connection almost always exists in the gate circuits, and (2) there may be more than one MOST that are situated in the transient (*off*-to-*on*) state in the CMOS gates. The latter reason means that the R_{eff} of the transient MOST is equal to R_t, which is always greater than R_{on}. But, the transient MOST that is most near the output node must be treated as a turn-on MOST because one transient MOST is always contained in the primitive cases (i.e., *i*-input NAND gate, *i*=integer number). Therefore, N_p will be greater than the actual number of MOST's in the discharging paths (only exists in the case of pure series-connection) when more than one MOST are in the transient state. It is reasonable because the larger resistance of R_t can be seen as the resistance composed of more than one MOST that are in the turn-on state. Finally, we conclude that the effective resistances of transient MOST's on the discharging path can be obtained by looking up R_t-table but the one that is most near the output should be treated as the turn-on MOST and replaced by R_{on}.

6. Slope estimation

If the output waveform can be treated as a simple RC waveform, then the parameter T in Eqs. 6 and 7 can be calculated by the equation: $T=0.5(t_{90\%}-t_{50\%})=0.294RC$. In BTS we defined ***slope*** as the time spent by the signal voltage dropping one volt, i.e. in units of time/volt, and then T=$ when V_{dd}=5V. Under this condition, T is the time that signal voltage drops one volt. The equivalent RC time constant of N tree can be computed by the equations as follows and by the recursive way while traversing the whole RC tree [6], [7].

(1) leaf node:

$$R_{eq}=R_t \text{ or } R_{on}$$
$$C_{eq}=C \qquad (10)$$
$$\tau_{eq}=R_{eq}C_{eq}$$

(2) for series connection:

$$R_{eq}=R_{eq1}+R_{eq2}$$
$$C_{eq}=C_{eq1}+C_{eq2}$$
$$\tau_{eq1}=R_{eq1}C_{eq1} \qquad (11)$$
$$\tau_{eq2}=R_{eq2}C_{eq2}$$
$$\tau_{eq}=\tau_{eq1}+\tau_{eq2}+R_{eq1}C_{eq2}$$

Note: R_{eq2} is nearer to the output node than R_{eq1}.

(3) for parallel connection:

$$R_{eq}=R_{eq1} \parallel R_{eq2}$$
$$C_{eq}=C_{eq1}+C_{eq2} \quad (12)$$
$$\tau_{eq}=R_{eq}(\tau_{eq1}/R_{eq1}+\tau_{eq2}/R_{eq2})$$

, where τ_{eq} is the equivalent RC time constant of subtree. Because the falling signal of output is affected by not only the N tree but also some internal nodes connected to output node in the P tree, the slope algorithm should consider both and then estimate the total effects.

7. Results

This method has been tested by some basic modules such as counters, decoders, adders, and ALU's. We take a 3-input NAND gate as example to show the results of delay estimation, which are shown in Fig. 8. Next, Fig. 9 shows the waveforms comparisons of a decoder

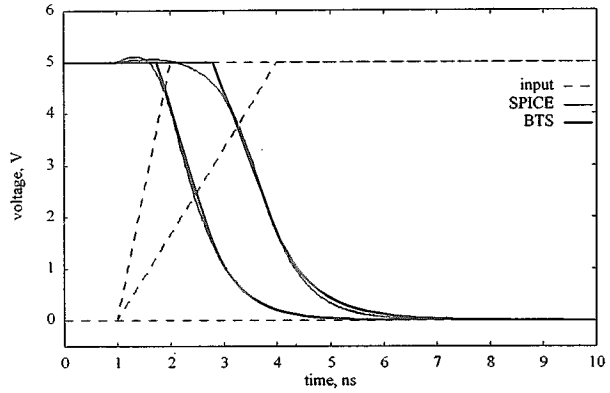

Figure 8 : Simulated waveforms of a 3-input NAND gate.

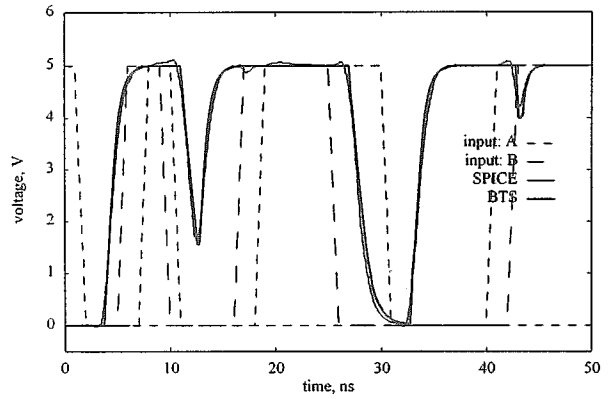

Figure 9 : Simulated waveforms of a 3-to-8 decoder SN74138.

SN74138 with SPICE results. In this circuit, four stages are passed from input to output, and any error of delay estimation will be accumulated to the next stage. Furthermore, the internal charges [3],[9] will also affect the delay and slope, and it is not the thing we can grasp easily and estimate accurately. Therefore, a slightly large amount of error is seen in Fig. 9. The speed of BTS is two to three orders of magnitude faster than that of SPICE for circuits with hundreds of transistors, and the speed ratio is expected to be even more significant for large-scale circuits. The CPU time comparisons are summarized in Table 1, and the speed ratio shows the advantage of BTS, especially for large scale circuits

8. Conclusion

A new approach for estimating the delay time of CMOS circuits is presented. This method modified the drawback in the previous version of BTS that converted the effect of overshoot to the turn-on-time of MOST (the value of V_T is shifted to 3.1V), and can offer a better adaptability for a wide range of circuit and input specification.

Table 1 : Comparisons between BTS and Spice

Circuit	MOS no.	CPU time on PC (DX4-100), secs BTS	Pspice	Speed ratio	Primary input event no.
complex gate (Fig. 1(a))	12	0.11	2.53	0.043	2
74138	88	0.33	102.22	0.0032	11
7483	258	0.99	774.64	0.0013	13
74381	584	1.10	1670.98	0.00066	14

References

[1] R. E. Bryant, "A Switch level model and simulator for MOS digital systems," IEEE Computers, vol. C-33, pp. 160-177, Feb. 1894.

[2] C. J. Terman, "RSIM - A Logic-Level Timing Simulator," Proceedings of the IEEE International Conference on Computer Design, New York, pp. 437-440, November 1983.

[3] J. H. Wang, Molin Chang, and W. S. Feng, "Binary-tree timing simulation with consideration of internal charges," IEE Proceedings-E, vol. 140, No.4, pp.211-219, July 1993.

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

[5] W. C. Elmore, "The transient response of damped linear networks with particular regard to wide-band amplifiers," J. Appl. Phys., vol. 19,no. 1,pp. 55-63, Jan. 1948.

[6] T. M. Lin, and C. A. Mead, "Signal delay in general RC networks," IEEE Trans. on Computer-Aided Design, vol. CAD-3, No.4, pp.331-349, October 1984.

[7] J. -P. Caisso, E. Cerny, and N. C. Rumin, "A recursive technique for computing delays in series-parallel MOS transistor circuits," IEEE Trans. on Computer-Aided Design, vol. 10, No.5, pp.589-595, May 1991.

[8] F.C.Chang, C.F.Chen, and P.Subramaniam, "An accurate and efficient gate level delay calculator for MOS circuits," Proceedings of 25th ACM/IEEE conference on Design automation, Anaheim, CA, USA, pp.282-287, June 1988.

[9] J. H. Wang, Molin Chang, and W. S. Feng, "The effects of internal charges to waveform calculation," ASIA-PACIFIC Conference on Circuits and Systems, Australia, pp.271-276. Dec., 1992.