MODELING THE DYNAMIC BEHAVIOR OF SERIES–CONNECTED MOSFETS FOR DELAY ANALYSIS OF MULTIPLE–INPUT CMOS GATES

L. Bisdounis and O. Koufopavlou

VLSI Design Laboratory, Department of Electrical & Computer
Engineering, University of Patras, GR-26500 Patras, Greece.

e-mail: bisdouni@ee.upatras.gr

ABSTRACT

In this paper the dynamic behavior of series-connected MOSFETs is studied, in order to compute the propagation delay of multiple-input static CMOS gates. A method for the reduction of series-connected MOSFETs to a simple MOSFET with the same behavior is proposed. The effective width of the equivalent transistor is not constant as in some previous works. So all cases of input slopes, the load capacitance, the number and the position of the switching inputs, and the body effect, are considered in order to determine the equivalent transistor's width. Along with the reduction process, an accurate analytical inverter timing model is used to compute the propagation delay of multiple-input static gates. The produced results are in very good agreement with SPICE simulations.

1. INTRODUCTION

Many different techniques for timing modeling of VLSI circuits have been proposed in order to improve the speed of circuit simulators. The reliability of these approaches depends on the accuracy with which the transient response and consequently the propagation delay of basic circuits can be evaluated.

In [1] we introduced an accurate analytical timing model for the CMOS inverter. The analytical nature of this model results in high computational speed, provided that an accurate and fast method to analyze multiple-input gates is to reduce them to equivalent inverters. Many reduction techniques [2]–[8] have been proposed in the literature. Traditionally, the equivalent width of series-connected transistors is approximated by W/N (W: channel width of one transistor, N: number of series-connected transistors) [2],[3], without taking into account the effects of the load and the input transition time. This is the main reason of inaccuracy, because it is valid only for step input waveforms, or when all transistors operate in the linear region. Another source of delay errors in some existing methods [3],[4] is the use of simple quadratic equations for the transistor currents, which are not valid for the recent sub-micron technologies.

Recently, a reduction technique in order to determine the effective width of serial-connected transistors has been proposed in [5], where several empirical parameters are used. However, the conventional model (W_{eff} = W/N) is used when the inputs are fast or all inputs are switched together. In [6], an approach to generalize an inverter-based timing model to multiple-input gates is proposed, where the serial transistors are handled by repeated dc analyses using SPICE every time a transition sets a new path. However, the modeling of transient phenomena by dc analysis results in inaccuracies. A number of dc analyses is also required in the reduction technique presented by Kong et al. [7], in order to determine the effective transconductance of series-connected transistors. Moreover, in [7] the current of the short-circuiting block is neglected in the delay calculation. Both techniques [6],[7] are limited to single input switching. The reduction technique suggested in [8], is demonstrated only for step inputs, and the calculation of the transistors' effective resistance is not discussed.

In this paper, a new reduction technique based on the dynamic behavior of series-connected MOSFETs, is presented. In the determination of the equivalent transistor's width, the influences of the output load, the input transition time, the number and the position of the switching inputs, and the body effect, are taken into account. Also, the additional delay due to the internal and coupling capacitances is included. Along with the reduction process, an accurate inverter timing model [1] which includes most of the factors that influence the inverter operation such as the currents through both transistors, the load capacitance, the input slope, and the input-to-output coupling capacitance, is used to compute the propagation delay of multiple-input CMOS gates. For the transistor currents, the α-power MOS model [9] is used, in order to include the velocity saturation effect of recent short-channel devices.

2. REDUCTION TECHNIQUE

Consider the multiple-input NAND gate shown in Fig.1. In the following we describe the reduction technique for series-connected nMOS transistors. The treatment of pMOS transistors is analogous. For the transistor drain currents, the following expressions of the α-power law MOSFET model [9] are used.

$$I_D = \begin{cases} P_C\,(W/L)\left(V_{GS} - V_{TH}\right)^{\alpha}, & V_{DS} \geq V'_{DO}, \quad \text{Saturation} \\ P_L\,(W/L)\left(V_{GS} - V_{TH}\right)^{\alpha/2} V_{DS}, & V_{DS} < V'_{DO}, \quad \text{Linear} \end{cases},$$

where $V'_{DO} = P_V\left(V_{GS} - V_{TH}\right)^{\alpha/2}$ is the drain saturation voltage, and $P_L = P_C / P_V$. P_C, P_V and α (velocity saturation index) are extracted from the I-V static characteristics of the device. For the determination of the device threshold voltage (V_{TH}) a linear approximation of the body effect is used [6], which results to the following simple formula

$$V_{TH} = V_{TO} + \gamma_1 V_{SB},$$

where V_{TO} is the zero-bias threshold voltage, and γ_1 is the body effect coefficient.

0-7803-4455-3/98/$10.00 © 1998 IEEE

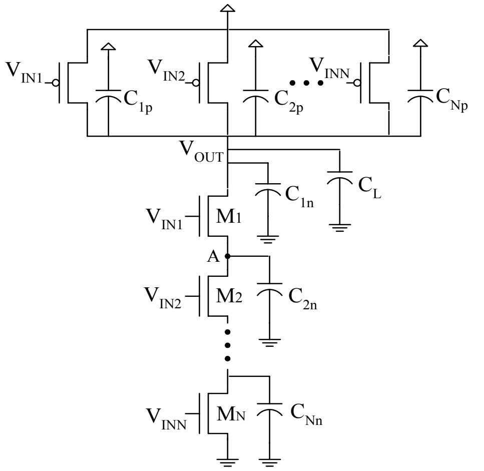

Fig.1: Multiple-input NAND gate

First, the case when all the transistors are switching together, which is the worst case scenario, is analyzed. The input voltages are assumed to be ramps, with input rise time τ. When the serial array performs the discharge operation, initially the topmost transistor operates in the saturation region, while the rest operate in the linear region. In the case when the topmost transistor is still saturated after the end of the input transition (fast input transition compared with the output one), its source node (A) is charged up to a plateau voltage [4] and maintains that voltage level until all transistors enter in the linear region (Fig.2). Then it follows the output voltage of the gate to ground. Since, the transistors M_2 to M_N operate in the linear region they approximated by an equivalent transistor M_K with channel width equal to W_K

$$\frac{1}{W_K}=\frac{1}{W_2}+\frac{1}{W_3}+\cdots+\frac{1}{W_N}.$$

When the voltage of node A is in the plateau region, no current flows in or out the internal capacitance of the node. Thus, the drain currents of the transistors M_1 and M_K are equal,

$$I_{D1} = I_{DK},$$

$$P_C(W_1/L)\left[V_{DD}-V_{TO}-\left(1+\gamma_1\right)V_P\right]^{\alpha_1} =$$

$$P_L(W_K/L)\left(V_{DD}-\overline{V}_{TK}\right)^{\alpha k/2}V_P, \quad (1)$$

where V_P is the plateau voltage. $\overline{V}_{TK}$ is the average value of the threshold voltages for the transistors M_2 to M_N, an approximated value of which is given by

$$\overline{V}_{TK}=\left(\sum_{i=1}^{N-1}V_{THi}\right)\Big/\left(N-1\right), \quad \text{where} \quad V_{THi}=V_{TO}+\gamma_1 V_{Si}, \quad \text{and}$$

$$V_{Si}=\frac{\left(N-i\right)\left(V_{DD}-V_{TO}\right)}{2N}.$$

Note, that the parameter α_k is extracted using the channel width W_K because it depends on the transistor width [9]. In order to solve (1), a Taylor series expansion of its left part around the point $V_P = [(N-1)\ V_{DD} / 2N]$, up to the second order coefficient is used. After that, V_P becomes the root of a simple quadratic equation.

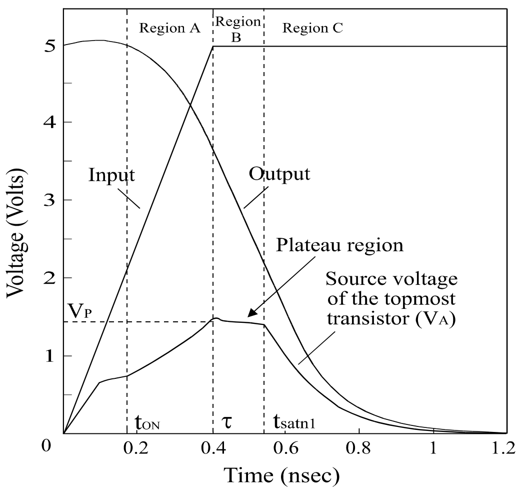

Fig.2: Voltage waveforms for fast inputs

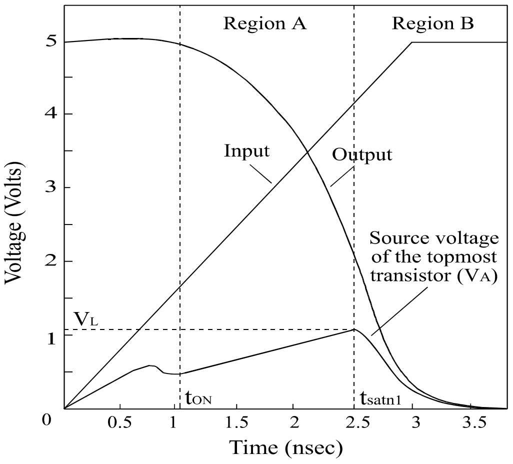

Fig.3: Voltage waveforms for slow inputs

In the following, the voltage at node A is considered linear for the interval between the time t_{ON} and τ. t_{ON} is the time where the chain of the serial transistors starts conducting. It is calculated by analyzing the influence of the gate-drain and gate-source capacitances on the chain operation [10]. After the calculation of the time t_{ON} and the value of V_A at this time, the slope of the voltage waveform at node A can be determined.

In the case when the topmost transistor enters in the linear region before the end of the input transition (slow input transitions), its source node voltage exhibits a peak value (V_L, see Fig.3) lower than the plateau one, before the input reaches its final value. This peak value occurs when the topmost transistor is entering the linear region. SPICE simulations indicate that the slope of V_A in this case is approximately the same with that calculated assuming the existence of plateau region. For slow inputs, V_A is considered linear between the time t_{ON} and t_{satn1}. t_{satn1} is the time when the topmost transistor is entering the linear region (Fig.3).

The discharge current through the serial array when the topmost transistor operates in the saturation region, is given by

$$I_D = P_C(W_1/L)\left[V_{IN} - V_{TO} - \left(1+\gamma_1\right)V_A\right]^{\alpha_1},$$

$$I_D = P_C(W_1/L)\left[1 - \frac{\left(1+\gamma_1\right)V_A}{V_{IN} - V_{TO}}\right]^{\alpha_1}\left(V_{IN} - V_{TO}\right)^{\alpha_1}.$$

The above equation has the same form as the current equation in the saturation region of a single transistor, the equivalent width of which is given by

$$W_{eq} = W_1\left[1 - \frac{\left(1+\gamma_1\right)V_A}{V_{IN} - V_{TO}}\right]^{\alpha_1}. \qquad (2)$$

Similarly, when the topmost transistor operates in the linear region the equivalent transistor's width becomes

$$W_{eq} = \frac{W_1}{N}\left[1 - \frac{\left(1+\gamma_1\right)\left(N-1\right)V_{OUT}}{N\left(V_{IN} - V_{TO}\right)}\right]^{\alpha_1/2}. \qquad (3)$$

Since all transistors operate in the linear mode, the array is considered as a voltage divider ($V_A = (N-1)\ V_{OUT} / N$), in the above equation. The last step is to determine the value of W_{eq} in each region of the chain operation.

In the case where the plateau region exists (Fig.2), three regions are studied.

Region A, $t_{ON} \le t \le \tau$: The equivalent width is calculated by equation (2) at the time $t = (t_{ON} + \tau) / 2$.

Region B, $\tau < t \le t_{satn1}$: The equivalent width is calculated by equation (2) for $V_{IN} = V_{DD}$, and $V_A = V_P$. t_{satn1} is calculated by equating the output voltage expression of the inverter model [1] in region 5A with the drain saturation voltage of the topmost transistor.

Region C, $t > t_{satn1}$: The equivalent width is the average between that calculated by equation (3) for $t = t_{satn1}$ and that calculated for $V_{OUT} = 0$.

In the case when the topmost transistor is entering the linear region before the end of the input transition (Fig.3), two regions are studied.

Region A, $t_{ON} \le t \le t_{satn1}$: W_{eq} is calculated by equation (2) at the time $t = (t_{ON} + t_s) / 2$. t_s is an approximation of t_{satn1}, and is calculated by equating the drain-source voltage ($V_{OUT} - V_A$) with the drain saturation voltage of the topmost transistor. The output voltage expression is determined by solving the differential equation resulting from the application of the Kirchoff's current law at the output node, with the assumption of negligible pMOS current. The use of t_s instead of t_{satn1} results in an error lower than 2% in the calculation of W_{eq}.

Region B, $t > t_{satn1}$: W_{eq} is calculated by equation (3), as in the region C of the previous case. t_{satn1} is calculated by equating the output voltage expression of the inverter model [1] in region 4 (or in region 3 for slower inputs) with the drain saturation voltage of the topmost transistor.

The output response of a multiple-input gate is a function of the number and the position of the switching transistors in the serial chain. When the input transition is sufficiently faster than the output, the topmost terminal switching shows faster discharge operation. This is because the lower transistors must discharge the upper transistors' internal capacitances. As the input transition becomes slower the lower terminal switching shows faster operation. This is because the transistor nearest to the ground has a smaller threshold voltage, while the magnitude of its gate-source voltage is greater than the other transistors in the chain. Hence, it will have a higher channel conductance than the other switching transistors, and the discharge operation will become faster.

The output waveforms for different combinations of switching inputs are translational [4], i.e. the shape of the curve is preserved except that its transition edge is shifted to the right or to the left depending on the combination of input signals. On the basis of this observation the equivalent width for each combination of switching inputs can be determined by multiplying the one of the worst case (all inputs switching together) with a single empirical factor m, during the input transition. m depends on the position and the number of the switching transistors, and on the relation between the input and the output waveforms. As a good metric of this relation, the single lumped parameter $G = (I_{DO}\ \tau) / (V_{DD}\ C_L)$ is used. I_{DO} is the drain current at $V_{GS} = V_{DS} = V_{DD}$ of a device with channel width equal to W/N (W: channel width of one transistor, N: number of series-connected transistors). Simulation results show that m changes exponentially with respect to G. This enables us to use the following equation for the determination of the coefficient m

$$m = m_{vf} + (m_{vs} - m_{vf})[1 - e^{-d\,(G-0.2)}], \qquad (4)$$

where m_{vf} is the weight coefficient for very fast inputs, m_{vs} is the weight coefficient for very slow inputs, and d is a constant. m_{vf} and m_{vs} are given by a look-up table (part of which is given in Table I). The values of the look-up table are extracted from SPICE simulations by adjusting the transistor size of an inverter to give the same output with the multi-input gate in the vicinity of V_{DD} / 2, for all the combinations of switching inputs. The values of Table I was obtained using a 0.8-micron technology, for G = 0.2 (very fast inputs) and G = 10 (very slow inputs). The constant d is equal to 0.42 for the used technology process. This value is almost independent of N, at least for the range $2 \le N \le 5$.

In the case of overlapping inputs, the existing waveform representation techniques [2],[5] can be used, which reduce the overlapping input signals of a multiple-input gate to a single effective signal. The equivalent channel width of parallel-connected transistors can be extracted by adding the widths of the switching transistors, and in the case of overlapping inputs can be found as in [5].

In the series-connected transistors, since the charge variation due to each of $C_{1n}, \ldots, C_{Nn}$ capacitances (see Fig.1) occurs through a different number of channels, their contribution to the output capacitance depend on their relative position in the chain. In the parallel transistors the capacitances $C_{1p}, \ldots, C_{Np}$ are added to the output load because they tied directly to the output node.

3. RESULTS AND CONCLUSIONS

In this section we illustrate the accuracy of the proposed approach for the evaluation of the transient response and the propagation delay of multiple - input CMOS static gates. In

Table I: Coefficients m_{vf} and m_{vs}

Switching Inputs	m_{vf}		m_{vs}	
	Number of serial transistors			
	3	2	3	2
1	1.12	1.07	1.41	1.29
2	1.06	1.02	1.65	1.60
1, 2	1.03	1	1.12	1
3	1.03		2.07	
1, 3	1.09		1.15	
2, 3	1.02		1.29	
1, 2, 3	1		1	

Fig.4 the output voltage waveforms of a static 3-input NAND gate for various input terminals switching with $C_L = 0.2$ pF, $\tau = 1.5$ ns and $V_{DD} = 5$ V, are shown. A 0.8-micron technology has been used, with transistor widths $W_n = 12$ μm and $W_p = 3$ μm. The output waveforms produced by SPICE simulations are added for comparison. It can be observed that the analytical waveforms are very close to those produced by SPICE simulations. This occurs because our model for the reduction of series-connected MOSFETs includes the influences of the output load, the input transition time, the number and the position of the switching inputs, and the body effect. In Fig.4 the input slope is smaller than that of the output waveforms. Thus, as mentioned in section 2, the topmost terminal (V_{IN1}) exhibits slower operation than the last one (V_{IN3}). In Fig.5, the propagation delay at the 50% voltage level of a 4-input NAND gate with the same characteristics, is plotted as a function of the input rise time. Results derived using the conventional model (W_{eq} = W/N) are also given. It is shown that, the conventional model gives inaccurate results, especially in the cases when only one terminal is switching.

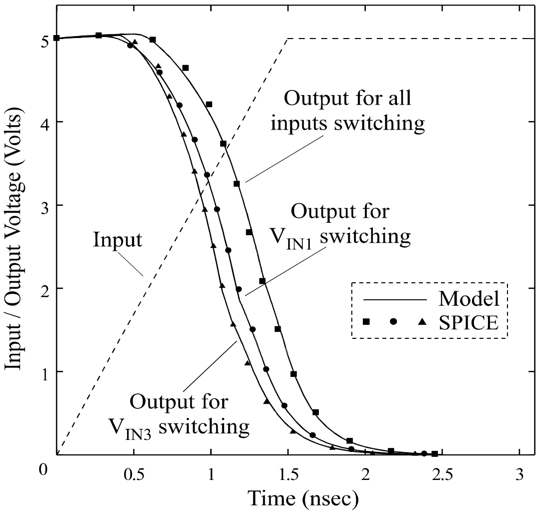

Fig.4: Output waveforms of a 3-input NAND gate

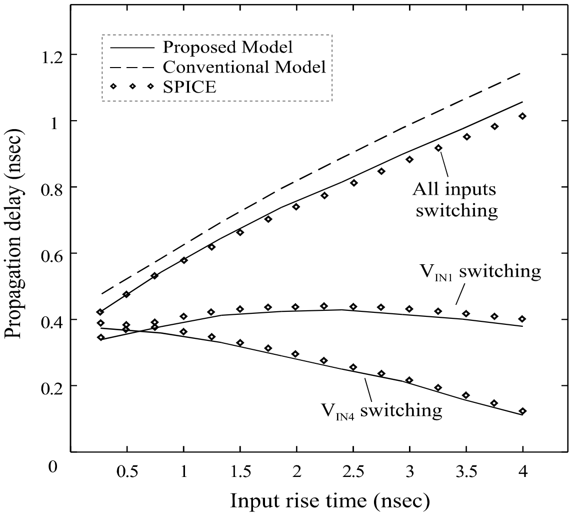

Fig.5: Propagation delay of a 4-input NAND gate

4. REFERENCES

[1] L. Bisdounis, S. Nikolaidis, O. Koufopavlou, "Analytical transient response and propagation delay evaluation of the CMOS inverter for short-channel devices", *IEEE J. Solid-State Circuits*, vol.33, pp. 302-306, February 1998.

[2] Y.H. Jun, K. Jun, S.B. Park, "An accurate and efficient delay time modeling for MOS logic circuits using polynomial approximation", *IEEE Trans. CAD*, vol.8, pp. 1027-1032, September. 1989

[3] Y.H. Shih, Y. Leblebici, S.M. Kang, "ILLIADS: A fast timing and reliability simulator for digital MOS circuits", *IEEE Trans. CAD*, vol.12, pp.1387-1402, September. 1993.

[4] S.M. Kang, H.Y. Chen, "A global delay model for domino CMOS circuits with application to transistor sizing", *International Journal Circuit Theory & Applications*, vol.18, pp.289-306, May 1990.

[5] A. Nabavi-Lishi, N.C. Rumin, "Inverter models of CMOS gates for supply current and delay evaluation", *IEEE Trans. CAD*, vol.13, pp.1271-1279, October 1994.

[6] T. Sakurai, A.R. Newton, "Delay analysis of series-connected MOSFET circuits", *IEEE J. Solid-State Circuits*, vol.26, pp.122-131, February 1991.

[7] J-T. Kong, D. Overhauser, "Methods to improve digital MOS macromodel accuracy", *IEEE Trans. CAD*, vol.14, pp.868-881, July 1995.

[8] D. Deschacht, M. Robert, D. Auvergne, "Synchronous-mode evaluation of delays in CMOS structures", *IEEE J. Solid-State Circuits*, vol.26, pp.789-795, May 1991.

[9] T. Sakurai, A.R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas", *IEEE J. Solid-State Circuits*, vol.25, pp. 584-594, April 1990.

[10]A. Chatzigeorgiou, S. Nikolaidis, "Collapsing the transistor chain to an effective single equivalent transistor", in Proc. Design, Automation and Test in Europe, February 1998.