CMOS GATE MODELING BASED ON EQUIVALENT INVERTER

A. Chatzigeorgiou, S. Nikolaidis [1], *I. Tsoukalas and O. Koufopavlou* [2]

Computer Science Department, [1] Department of Physics
Aristotle University of Thessaloniki, 54006 Thessaloniki, Greece
[2] Department of Electrical and Computer Engineering
University of Patras, 26500 Patras, Greece

ABSTRACT

A method for modeling complex CMOS gates by the reduction of each gate to an effective equivalent inverter is introduced. The conducting and parasitic behavior of parallel and serially connected transistors is accurately analyzed and an equivalent transistor is extracted for each case, taking into account the actual operating conditions of each device in the structure. The accuracy of the method is validated by the results for two submicron technologies and its efficiency as a technique that can improve existing timing simulators is demonstrated.

1. INTRODUCTION

As the size of integrated circuits is increasing continuously, the demand for fast and accurate simulation is also growing. CAD tools which are based on numerical methods for timing and power analysis such as SPICE are prohibitively slow for large designs and consequently analytical methods that can provide the same level of accuracy at much lower time are required. Extensive research has been conducted on the modeling of the CMOS inverter leading to efficient analytical expressions for its performance [1], [2], [3].

However little has been done on more complicated gates such as NAND/NOR gates mainly because of the intrinsic difficulties in the analysis of multinodal circuits. All previous works on the modeling of CMOS gates can be divided in two main categories : The first one aims at the solution of the differential equations that describe the operation of a gate and are based on the fact that within a transistor chain some transistors operate always in the linear region [4]. However, simplifying assumptions were used, such as transistor replacement by resistors, long channel transistor current models, negligible body effect and step inputs resulting in limited accuracy. The second attempt [5] seeks an equivalent inverter for each gate that will have the same output response with that of the complete gate. Unfortunately, the different modes of operation of the transistors within a transistor chain have not been captured accurately by the single equivalent transistor. It should be noted that in commercial or university dynamic timing simulators, transistor chains are treated by merging the serially connected transistors in the conventional way (i.e. the transconductance of the equivalent transistor is reduced by the number of the transistors in the chain) leading in errors up to 100 % [6].

In this paper a method for modeling NAND/NOR CMOS gates by an effective equivalent inverter which is calculated taking into account the different modes of operation of the transistors in the gate, is introduced. Serial and parallel combinations of transistors are analyzed for both operating conditions, conducting and short-circuiting behavior. More complex gates are treated by collapsing them to an equivalent NAND/NOR gate structure [6]. The proposed model incorporates short-channel effects, the influence of body effect and is developed for non-zero transition time inputs.

2. MODELING OF SERIAL AND PARALLEL TRANSISTOR STRUCTURES

Since CMOS gates consist of parallel and serially connected transistors, each of these structures has to be modeled accordingly. In order to analyze the operation of the transistor chain in a NAND gate when it is conducting, let us consider the circuit in Fig. 1a assuming that an input ramp with transition time τ is applied to the gates of all transistors in the chain. In case the applied ramps to a gate are not normalized, that is when they have different starting points and transition times, an input mapping algorithm such as the one mentioned in [7] should be employed. The influence of the pMOS transistors will be considered later on in the analysis.

The α-power law model proposed in [1], is used for the transistor currents :

$$I_D = \begin{cases} 0 & V_{GS} \le V_{TN} : \textit{cutoff} \\ k_l\left(V_{GS}-V_{TN}\right)^{\alpha/2} V_{DS}, & V_{DS} < V_{D-SAT} : \textit{linear} \\ k_s\left(V_{GS}-V_{TN}\right)^{\alpha} & V_{DS} \ge V_{D-SAT} : \textit{saturation} \end{cases} \quad (1)$$

The above model takes into account the carrier velocity saturation effect of short-channel devices. $V_{D\text{-}SAT}$ is the drain saturation voltage, k_l , k_s are the transconductance parameters, α is the carrier velocity saturation index and V_{TN} is the threshold voltage expressed by its first order Taylor series approximation as $\tilde{V}_{TN} = \theta + \delta V_{SB}$, where V_{SB} is the source-to-substrate voltage.

The topmost transistor in the chain, (M_n), begins its operation in saturation and enters the linear region when $V_{DS_n} = V_{D-SATN}$ while all other transistors operate always in the linear region. When the input reaches V_{DD} and until transistor M_n exits saturation, its current and therefore the internal node voltages remain constant. During this time interval which is known as the "plateau" state [4], the parasitic currents are eliminated and consequently the same current flows through all transistors in the chain. The plateau state is apparent only for fast input transitions. Fast and slow inputs are determined according to the position of time point t_2, when the top transistor in the chain exits saturation : in case $t_2<\tau$, the input is slow, otherwise it should be considered fast. The plateau voltage, which is the voltage of the source node of the topmost transistor in the chain during the plateau state, can be calculated according to [8].

In addition, the source voltage of the top transistor in the chain, V_M, is considered linear between time t_1 where the chain starts conducting and time τ (fast inputs) or time t_2 (slow inputs) where the top transistor exits saturation [8]. Since time t_1 and $V_M[t_1]$ can be estimated and for fast inputs the plateau voltage occurs at time τ, the slope of V_M can also be obtained [8].

0-7803-5474-5/99/$10.00 ©1999 IEEE

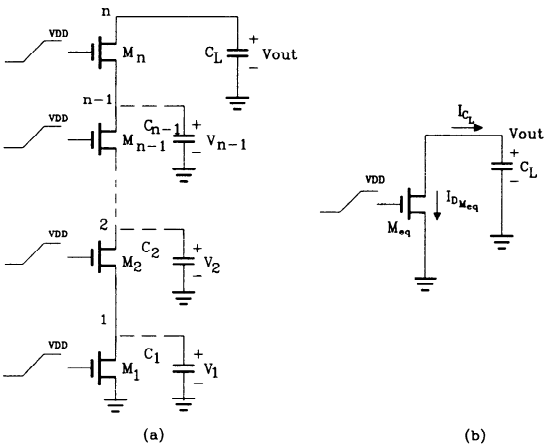

Fig. 1: a) Complete transistor chain, b) single equivalent transistor

It is obvious that a single equivalent transistor will have the same output response with the complete chain, if it successfully manages to reproduce the combined behavior of the nonsaturated devices with the dual operation of the topmost transistor, in saturation and the linear region. While the top transistor is saturated, the current through that transistor is the bottleneck for the current that is flowing through the chain. Thus, in order to obtain the width (W_{eq}) of the single equivalent transistor (M_{eq}) (Fig. 1b) the currents through transistor M_n of the complete chain and transistor M_{eq} are set equal :

$$I_{M_n} = I_{M_{eq}} \Rightarrow$$

$$P_s \frac{W_n}{L}\left(V_{in} - \theta - (1+\delta)\, V_M\right)^a = P_s \frac{W_{eq}}{L}\left(V_{in} - V_{TO}\right)^a \quad (2)$$

The above equation can be solved for several values of t yielding corresponding W_{eq} values. In order to find an average effective value for W_{eq}, time point t_2 has to be calculated. The output voltage expression can be obtained by solving the following differential equation at the output node:

$$C_L \frac{dV_{out}}{dt} = -I_{M_n} = -k_s\left(V_{in} - \theta - (1+\delta)V_M\right)^a \quad (3)$$

Time t_2 is calculated by equating the drain saturation voltage to the actual drain-to-source voltage of transistor M_n:

$$V_{D-SATN}[t_2] = V_{out}[t_2] - V_M[t_2] \quad (4)$$

However, in a real CMOS gate the effect of the short-circuit current must also be taken into account. Therefore, the next step is to incorporate in the analysis the short-circuit current of a parallel pMOS transistor structure whose influence on the estimation of t_2 has been neglected. This parasitic current acts as an additional charge, resulting in an extension of the saturation region and thus affecting the effective value of W_{eq} in this region. In order to obtain higher accuracy in the estimation of t_2, the expression of the pMOS current should be inserted in eq. (3), increasing the mathematical complexity of the proposed method.

However, for simplicity, t_2 is calculated by eqs. (3), (4) and the effective value of W_{eq} is selected in such a way, so that the errors which are introduced by the underestimation of the bound of the saturation region be compensated. A very good approximation for W_{eq} which is valid for a wide range of input slopes and load capacitances, is to calculate its value from eq. (2) at $t = t_2$ for fast inputs and at $t = (t_1 + 3.3\, t_2)/4$ for slow inputs. The calculated value of W_{eq} in this region of operation will be referred to as W_{sat}.

When all transistors operate in the linear region, the transistor chain can be considered as a voltage divider with a uniform distribution of the output voltage among all drain/source nodes. According to this, the width of the equivalent transistor for this region can be calculated as $W_{lin} = W/n$, in case of non-tapered transistor chains.

Consequently the chain can be modeled by a single equivalent transistor whose width from time t_1 to time t_2 is W_{sat} and for the rest of the time equal to W_{lin}. However, since the aim was to provide an equivalent width, the above two width values should be efficiently merged into one. This can be accomplished by calculating the fraction of charge (Q_{sat}) that is discharged to ground during the time in which the top transistor in the chain operates in saturation over the total charge ($Q_{total}=C_L V_{DD}$) that is stored initially in the output load. Q_{sat} is calculated as: $Q_{sat}=Q_{total} - Q[t_2] = C_L V_{DD} - C_L V_{out}[t_2]$ and the saturation coefficient c_{sat} can be defined as $c_{sat} = Q_{sat}/Q_{total}$. Thus, the corresponding coefficient when all transistors operate in linear mode, is equal to $c_{lin}=1-c_{sat}$.

Since the calculated coefficients act as the "weight" of each mode of operation on the overall output voltage temporal evolution, the width of the single equivalent transistor can be calculated as :

$$W_{eq} = c_{sat} \cdot W_{sat} + c_{lin} \cdot W_{lin} \quad (5)$$

The parallel transistor structure is less complicated than the serial one, since if an equal transistor width is assumed (as in the case of a NAND/NOR gate) and the same input ramp is applied to the gates of all transistors, the currents flowing through each transistor will be identical. Consequently, the parallel transistor structure can be replaced by a single transistor with its width multiplied by the number of the transistors.

According to the above, the output evolution of a NAND/NOR gate can be accurately captured by an equivalent inverter whose transistors (conducting and short-circuiting) replace the corresponding blocks of the gate.

An important issue in the modeling of CMOS gates is the accurate estimation of the time point when a gate starts conducting. This time point, which can be calculated according to [9], should be used when performing the transient analysis of the equivalent inverter so that the output charging/discharging begins at this point.

3. PARASITIC BEHAVIOR OF THE TRANSISTOR CHAIN

With the term parasitic behavior of the transistor chain during output switching of a CMOS gate, we refer to its parasitic effect on the output voltage evolution. The parasitic behavior results in a short-circuit current which reduces the rate of charging/discharging of the output load and increases the propagation delay.

Let us consider a NAND gate where all pMOS transistors have been replaced by an equivalent one. The parasitic behavior of the nMOS transistor chain will be modeled by an equivalent transistor. Consequently, in the case of charging output, the NAND gate diminishes to an equivalent inverter. A falling ramp input with transition time τ is considered to be applied to the gates of all transistors.

First, because the output voltage is small while the V_{GS} of the nMOS devices is large, all these transistors start their operation in linear mode. As the output voltage rises, the voltages at the internal nodes of the chain are also increasing. All nMOS transistors have almost equal V_{DS} (voltage divider) while the top is biased by the smallest V_{GS}. This means that this transistor at some time point will enter saturation and after this time, the current in the chain will decrease keeping all other devices in linear mode.

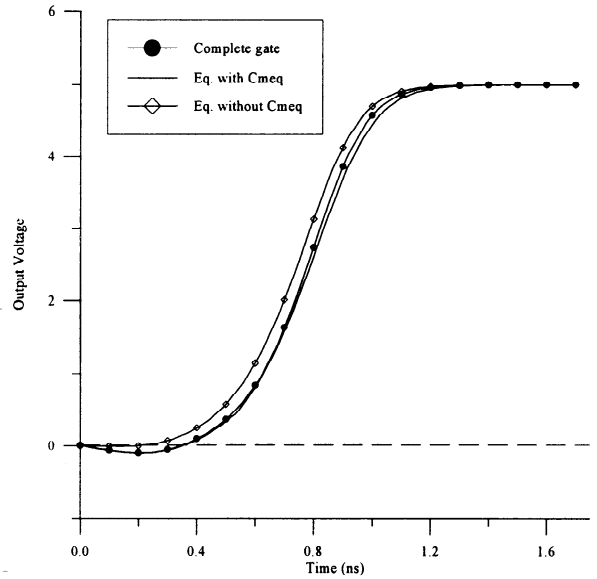

Fig. 2: Output waveform comparison between the complete gate and the equivalent inverter with and without the equivalent coupling capacitance (L=0.5 μm).

A significant amount of parasitic current is also flowing through the coupling capacitances between the gates of the nMOS transistors and the corresponding drain/source diffusion areas. In order to perform an accurate modeling of the gate when the chain behaves parasitically, an equivalent capacitance has to be inserted between the input terminal and the output node of the corresponding nMOS transistor in the equivalent inverter. Although the dual operation of the topmost transistor is also present during the parasitic operation of the chain, conventional estimation of the width of the equivalent transistor as $W_{eq} = W/n$ has been found to give sufficiently accurate results if the effect of the parasitic capacitances is modeled properly.

The output load will start being charged through the pMOS transistor by a current of the form $I_s = k_s \cdot \left(V_{GS} - |V_{TP}|\right)^{\alpha}$. If we ignore the parasitic contribution of the nMOS transistor currents, the rate of the output voltage increase during $[t_p, t_{sat}]$ is given by :

$$C_L \frac{dV_{out}}{dt} = I_s(t) \Rightarrow \frac{dV_{out}}{dt} = \frac{I_s(t)}{C_L} = I^c(t) \qquad (6)$$

where $t_p = |V_{TP}|\,\tau/V_{DD}$ is the time when the pMOS transistor starts conducting and t_{sat} is the time when the top transistor in the nMOS transistor chain enters saturation and is approximated by $\left(t_n + t_p\right)/2$, where $t_n \approx \frac{V_{DD} - V_{TN}}{V_{DD}} \cdot \tau$ is the time when the nMOS transistors cease to conduct.

Since the whole chain acts as voltage divider during $[t_p, t_{sat}]$ the slope of each internal node voltage is found assuming a uniform distribution of the output voltage slope.

The current that the coupling capacitance at each node, C_{M_i}, is drawing during time interval $[t_p, t_{sat}]$ is equal to :

$$I_i = C_{M_i} \frac{d\left(V_i - V_{in}\right)}{dt} = C_{M_i} \cdot \left(\frac{i \cdot I^c}{n} + s\right) \qquad (7)$$

where n is the number of the transistors in the chain, $s = V_{DD}/\tau$, and $\frac{i\,I^c}{n}$ the slope of the voltage waveform at the internal node i.

By summing the currents through all coupling capacitances and equating the sum with the current that must flow through the equivalent coupling capacitance, $C_{M_{eq}}$ is obtained:

$$\sum_{i=1}^{n} I_i = C_{M_{eq}} \cdot \left(I^c + s\right) \qquad (8)$$

A constant value for $C_{M_{eq}}$ can be obtained if an average value, c_r, is calculated for I^c, by integrating the pMOS current I_s over $[t_p, t_{sat}]$. This value corresponds to the average slope of the output voltage waveform until t_{sat}.

By symmetry, when the node voltages are decreasing during $[t_{sat}, t_n]$, the same slope (with opposite sign) is valid for the voltage waveforms at the internal nodes. Thus, the equivalent coupling capacitance for the two time intervals can be written as :

$$C_{M_{eq}} = C_M \cdot \frac{n \cdot c_r + (2n-1) \cdot s}{2 \cdot \left(c_r + s\right)} \quad , \left[t_p, t_{sat}\right] \qquad (9)$$

$$C_{M_{eq}} = C_M \cdot \frac{(n-1) \cdot \left(s - c_r/2\right)}{\left(c_r + s\right)} \quad , \left[t_{sat}, t_n\right] \qquad (10)$$

The improvement that is gained by inserting this coupling capacitance to the equivalent inverter model of a gate is significant as shown in Fig. 2 which is a comparison of the output response of a 4-input NAND gate and that of the corresponding equivalent inverter with and without the calculated coupling capacitance, when the transistor chain acts parasitically.

The parasitic behavior of a parallel transistor structure is modeled accurately by a single equivalent transistor whose width is equal to that of a single transistor multiplied by the number of the transistors, since the effect of the coupling capacitances is captured by the increased coupling capacitance of the equivalent transistor.

4. RESULTS

The output waveform of a 4-input NAND gate for two submicron technologies (0.5 μm HP and 0.35 μm HP) and C_L=0.1 pF is compared to that of the proposed equivalent inverter for several input transition times and is shown in Fig. 3 (a) and (c). As it can be observed, the proposed method presents very good accuracy for both submicron technologies. In Fig. 3 (e), (f) the output waveform of a 4-input NAND gate is compared to that of an equivalent inverter whose width is calculated according to the proposed method and in the conventional way (where the transconductance of the equivalent transistor for the transistor chain is reduced by the number of transistors in the chain) for two technologies (0.5 μm and 0.35 μm) and for several input transition times. The superiority of the proposed method compared to the conventional equivalent inverter is obvious.

In Fig. 3 (b) and (d), output waveform results from a widely used dynamic timing simulator, ILLIADS2 [10], are also shown and compared to the actual output waveform of the NAND gate which is obtained using SPICE. In the same plots the output waveform of an equivalent inverter whose transistor widths are calculated according to the proposed method and which is simulated using ILLIADS2 is also shown. It is clear that the improvement that is gained when the proposed additional step is added in order to calculate more

efficiently the width of the equivalent transistor that replaces serially connected transistors is significant.

5. REFERENCES

[1] T. Sakurai and 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,no.2, pp.584-594,April 1990.

[2] 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, no. 2, February 1998.

[3] A. Hirata, H. Onodera and K. Tamaru, "Estimation of Short-Circuit Power Dissipation for Static CMOS Gates", IEICE Trans. Fundamentals, vol. E79-A., no. 3, March 1996.

[4] S. M. Kang and H. Y. Chen, "A Global Delay Model for Domino CMOS Circuits with Application to Transistor Sizing", Int. J. Circuit Theory and Applicat., vol. 18, pp. 289-306, 1990.

[5] A. Nabavi-Lishi and N. C. Rumin, "Inverter Models of CMOS Gates for Supply Current and Delay Evaluation", IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 13, no. 10, pp. 1271-1279, October 1994.

[6] J.-T. Kong, S. Z. Hussain and D. Overhauser, "Performance Estimation of Complex MOS Gates", IEEE Trans. on Circuits and Systems-I: Fundamental Theory and Applications, vol. 44, no. 9, pp. 785-795, September 1997.

[7] A. Chatzigeorgiou and S. Nikolaidis, "Input Mapping Algorithm for Modelling of CMOS Circuits", IEE Electronics Letters, vol. 34, no. 12, pp. 1177-1179, 1998.

[8] A. Chatzigeorgiou and S. Nikolaidis, "Analysis of the Transistor Chain Operation in CMOS Gates for Short Channel Devices", Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS), vol. VI, pp. 363-367, 1998.

[9] A. Chatzigeorgiou, S. Nikolaidis and I. Tsoukalas, "Estimating starting point of conduction of CMOS gates", IEE Electronics Letters, vol. 34, no. 17, pp. 1622-1624, 1998.

[10] A. Dharchoudhury, "Advanced Techniques for Fast Timing Simulation of MOS VLSI Circuit", Ph.D. Dissertation, Dept. Elec. Comp. Eng., University of Illinois, Urbana-Champaign, 1995.

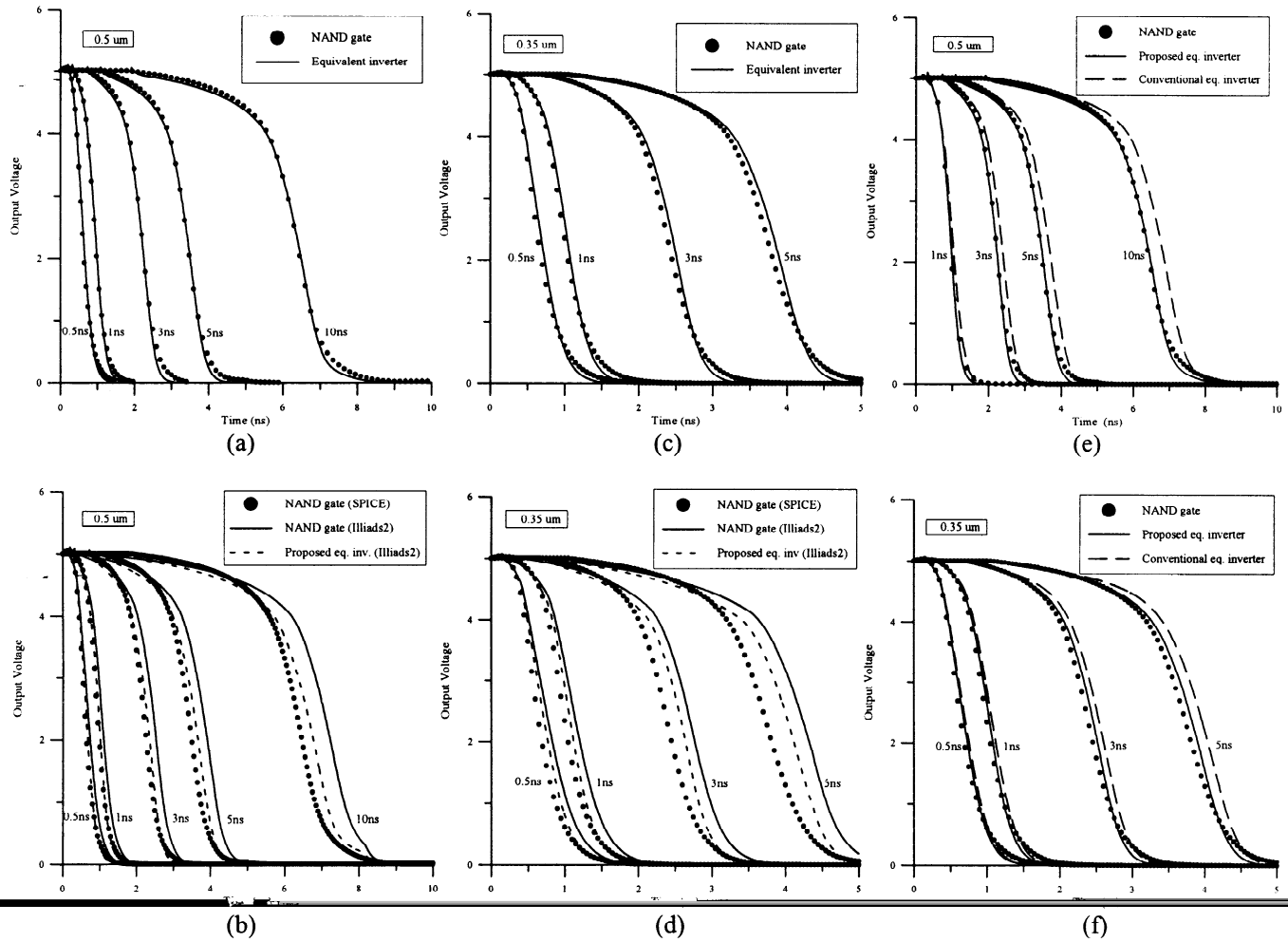

Fig. 3: (a) and (c): Output waveform comparison between the complete gate and the equivalent inverter for several technologies and input transition times, (b) and (d): Illiads2 improvement using the proposed equivalent inverter, (e) and (f): Output waveform derived according to the proposed method and in the conventional way.