Proceedings of the Design, Automation and Test in Europe Conference (DATE'98), Paris, France, 23-26 February, 1998, pp. 2-6.

Collapsing the Transistor Chain to an Effective Single Equivalent Transistor

A. Chatzigeorgiou and S. Nikolaidis*

Computer Science Department, *Department of Physics
Aristotle University of Thessaloniki,
Thessaloniki 54006, Greece

Abstract

The most common practice to model the transistor chain, as it appears in CMOS gates, is to collapse it to a single equivalent transistor. This method is analyzed and improvements are presented in this paper. Inherent shortcomings are removed and an effective transistor width is calculated taking into account the operating conditions of the structure, resulting in very good agreement with SPICE simulations. The actual time point when the chain starts conducting which influences significantly the accuracy of the model is also extracted. Finally, an algorithm to collapse every possible input pattern to a single input is presented.

1. Introduction

The development of digital integrated circuits with short design cycles requires accurate and fast timing simulation. For small designs simulation is possible by means of simulators such as SPICE which are based on numerical methods. Unfortunately, these simulators are excessively slow for the multi-million transistor designs in the submicron era. The need for analytical methods which can produce accurate results at short times is obvious and extended research has been conducted, especially for the modelling of the CMOS inverter [1, 2]. More complicated gates such as NAND/NOR structures are difficult to model because of their multinodal circuitry and multiple inputs.

The accurate modelling of the transistor chain, which is a structure of vital importance for CMOS circuits, by an equivalent transistor is examined in this paper. Conventionally, the transistor chain is modelled by a single transistor with its width reduced by the number of the transistors in the chain. In case all transistors in the chain receive the same input, this common input is also applied to the equivalent transistor. Several attempts have been made in order to model the transistor chain trying to exploit the fact that all transistors in the chain except for the top one, generally operate in linear mode. In [3] the complete chain is modelled considering a long RC chain while in [4, 5] the nonsaturated devices were replaced by an effective resistance. Sakurai and Newton [6] developed their analysis for the CMOS inverter and extension to gates was made by a delay degradation factor. In general, all previous works used simplifying assumptions for the operation of the chain, such as step inputs, resistive behaviour, long channel models and negligible body effect. Apart from the fact that it is impossible to solve analytically the differential equations which describe the operation of the chain, it is often advantageous to model the transistor chain by a single equivalent transistor, since this model can be easily applied in the delay and short-circuit power estimation of more complex gates. In addition, when a single equivalent transistor is obtained, the well defined modelling analysis of the inverter can be used in order to determine the propagation delay and power dissipation of CMOS gates, since modelling of parallel transistors can be performed successfully as in [7].

It should be mentioned that Nabavi-Lishi and Rumin [8] presented a semi-empirical method for collapsing a transistor chain to a single equivalent transistor. The equivalent width approximation, for practical inputs ends up in the conventional n-times transconductance reduction, where n is the number of transistors in the chain, resulting in limited accuracy. For slow inputs, a formula is presented in order to modify the equivalent width obtained for fast inputs, but the empirical coefficients which are used, make the application of the method impossible.

In this paper the width of the single equivalent transistor is calculated taking into account the mode of operation of the transistors in the chain and the time during which the devices operate in each mode. Moreover, the time at which the chain starts conducting is obtained overcoming a main source of errors in all existing modelling techniques. Finally, an input mapping algorithm is presented which leads to a single input signal which effectively replaces all inputs of the chain.

2. Transistor Chain Operation

Let us consider the discharging of an output load by the transistor chain consisting of NMOS devices illustrated in Fig. 1, where the parasitic internal node capacitances are also shown. The case of a PMOS charging chain is symmetrical.

A common rising input ramp with input rise time τ, is applied to all gates. The a-power model proposed in [2]

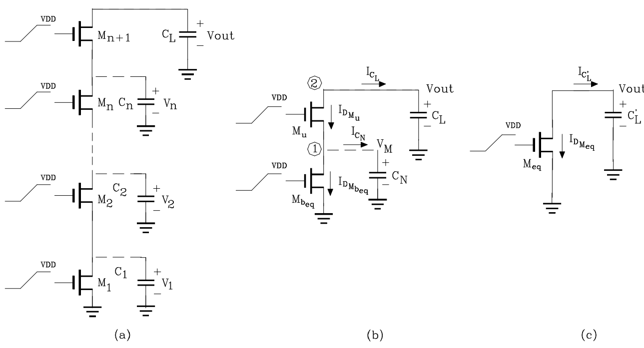


Fig. 1: **(a) Complete transistor chain, (b) two transistor equivalent circuit and (c) single equivalent transistor model**

which takes into account the carrier velocity saturation effect of short channel devices, is used for the transistor currents of the chain :

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

where $V_{D\text{-}SAT}$ is the drain saturation voltage, k_l , k_s are the transconductance parameters, a is the velocity saturation index and V_{TN} is the threshold voltage which in order to be treated mathematically, is written using a first order Taylor series approximation around V_{SB}=1V as $\widetilde{V}_{TN} = \theta + \delta V_{SB}$.

During the evolution of the load discharging, the topmost transistor in the chain (M_{n+1}) operates initially in saturation and then enters the linear region when its $V_{DS}=V_{D\text{-}SAT}$. All other transistors in the chain operate in the linear region without ever leaving this region [5]. Because the voltage at the intermediate nodes rises, the current of the nonsaturated devices increases and there will be a time point where the current of the bottom transistors will be equal to the current of the topmost transistor. The circuit remains at this "plateau" state until the topmost transistor exits saturation [5] and during this time the voltage at all drain/source nodes remains constant. The plateau voltage is apparent for fast inputs (Fig. 2). An input is considered fast (slow) when the voltage at the source of the top transistor attains its maximum value when (before) the input reaches V_{DD}.

In order to calculate the plateau voltage the nonsaturated devices are replaced by an equivalent transistor ($M_{b_{eq}}$) whose width is approximated by :

$$\frac{1}{W_{eq_{conv}}} = \frac{1}{W_1} + \frac{1}{W_2} + \ldots + \frac{1}{W_n} \tag{2}$$

The two-transistor equivalent circuit is shown in Fig. 1b. When a fast input is applied, the plateau voltage at the source of the top transistor occurs at the end of the input ramp at time t=τ. Equating the currents of the saturated top transistor (M_u) and the bottom transistor ($M_{b_{eq}}$) which operates in linear mode in the two-transistor equivalent chain for $V_{in}=V_{DD}$ results in :

$$k_s \left(V_{DD} - \theta - (1+\delta)V_p\right)^a = k_{l_{beq}} \left(V_{DD} - V_{TO}\right)^{a/2} V_p \tag{3}$$

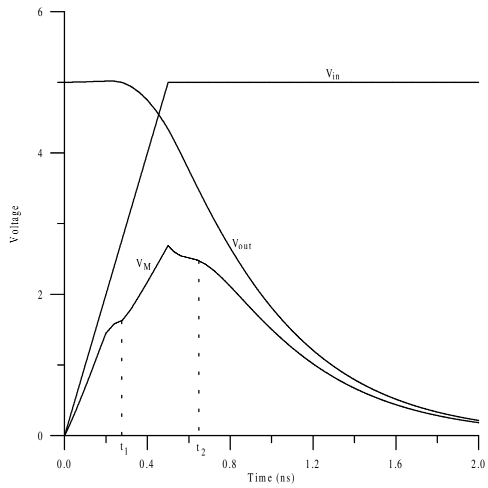


Fig. 2: **Output and intermediate node voltages of a transistor chain for a fast input ramp**

Solving the above equation using a second order Taylor series approximation gives the value of the plateau voltage with very good accuracy.

In the following analysis the voltage at the source of the top transistor in the chain, V_M, is considered linear for the interval between time point t_1, where the chain starts conducting, and time τ (fast inputs) where the input reaches its final value or time t_2 (slow inputs) where the top transistor exits saturation. This observation is based on SPICE simulations and leads to highly accurate results. Since time t_1 and $V_M[t_1]$ is known (see next section) and for fast inputs the plateau voltage occurs at time τ the slope of V_M is also known. For slow inputs it has been observed that the slope remains the same and can be calculated as if the input were fast, i.e. by assuming that at time τ, V_M reaches the plateau voltage. In order to calculate the time point (t_2), the output voltage expression has to be found by solving the following differential equation resulting from the application of Kirchhoffs's current law at the output node :

$$C_L \frac{dV_{out}}{dt} = -k_s \left(V_{in} - \theta - (1+\delta)V_M\right)^a \tag{4}$$

Time t_2 is obtained by equating the drain saturation voltage with the actual drain-to-source voltage of transistor M_{n+1} ($V_{D-SAT} = V_{out} - V_M$).

3. Starting Point of Conduction

Since each transistor in the chain has a different source voltage and threshold voltage , the time at which the condition $V_{GS} - V_{TN} = 0$ for each one is satisfied and the transistor starts conducting, is also different. Let us consider a five transistor chain which receives a common ramp input

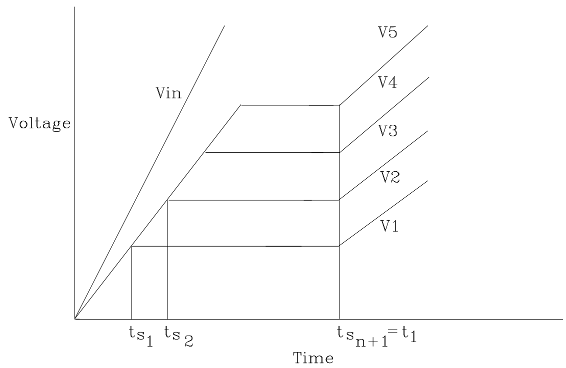


Fig. 3: **Simplified representation of intermediate node voltages**

at time t=0 and assume that all internal nodes are initially discharged. Coupling capacitance between the input terminal and the drain/source nodes forces all internal node voltages to follow the input ramp until all transistors below the node start conducting. From this point on, the node voltage is subject to two opposite trends. One is due to the coupling capacitance and tends to raise the node voltage and the other is due to the current flowing to ground which tends to pull the node voltage down. The system behaviour is modelled assuming that the two trends are counterbalanced, thus leading in a constant node voltage until the time point where all transistors above the node start to conduct, when the node voltage is pulled high again (Fig. 3). The slope of the node voltages until the transistors below the i-th node start conducting at time t_{s_i} can be found by equating the current through the coupling capacitance ($I_{C_{M_i}}$) with the charging current of the parasitic capacitance I_{C_i} : $I_{C_{M_i}} = I_{C_i} \Rightarrow$

$$C_{M_i} \frac{dV_{in} - dV_i}{dt} = C_i \frac{dV_i}{dt} \Rightarrow V_i[t] = \frac{C_{M_i}}{C_{M_i} + C_i} V_{in}[t] \quad (5)$$

The bottom transistor M_1 in the chain starts conducting at time $t_{s_1} = \frac{\theta \cdot \tau}{V_{DD}}$ when its input reaches the threshold voltage. From this time on and until all transistors further up start conducting at t=t_1, the voltage at the drain of transistor M_1 remains constant at $V_1\left[t_{s_1}\right] = \frac{C_{M_1}}{C_{M_1} + C_1} \frac{V_{DD}}{\tau} t_{s_1}$. The time at which the i-th transistor starts conducting can be calculated recursively by solving $V_{GS_i}\left[t_{s_i}\right] - V_{TN_i}\left[t_{s_i}\right] = 0$, leading to :

$$t_{s_i} = \tau \cdot \frac{\theta + (1+\delta) \dfrac{C_{M_{i-1}}}{C_{M_{i-1}} + C_{i-1}} \dfrac{V_{DD}}{\tau} t_{s_{i-1}}}{V_{DD}} \quad (6)$$

where i corresponds to the position of the transistor in the chain (t_{s_0} =0). The time at which the chain turns on is calculated using the above expression as $t_1 = t_{s_{n+1}}$.

4. Effective Width of the Equivalent Transistor

In the following analysis all internal nodes of the chain are considered to be discharged at time t=0. In case some of the nodes are initially charged, the output waveform should be shifted properly, since the charges in the internal nodes cause an additional delay in the output response [4].

It is obvious, that a single equivalent transistor (Fig. 1c) which receives the common input that is applied to all transistors in the chain, will have the same output response with the complete chain, if it successfully manages to reproduce the combined behaviour of the nonsaturated devices with the dual operation of the top transistor, in saturation and the linear region.

For the time interval where the top transistor in the chain is saturated, the current through this transistor is the bottleneck for the current that flows through the chain [5]. Therefore, in order to obtain the width (W_{eq}) of the single equivalent transistor (M_{eq}) the currents through transistor M_{n+1} of the complete chain and transistor M_{eq} should be set equal : $I_{M_{n+1}} = I_{M_{eq}} \Rightarrow$

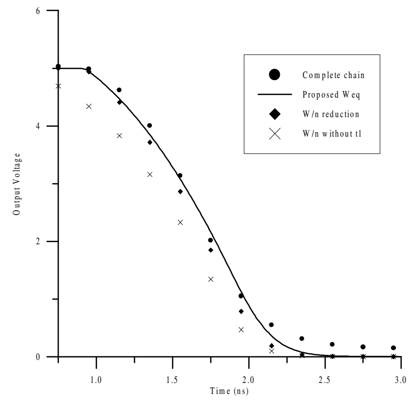
$$P_s \frac{W_{n+1}}{L} \left(V_{in} - \theta - (1+\delta)\, V_M\right)^a = P_s \frac{W_{eq}}{L} \left(V_{in} - V_{TO}\right)^a \quad (7)$$

The above equation can be solved for several values of t yielding corresponding W_{eq} values. Therefore, in order to obtain an average value, W_{eq} should be calculated from eq. (7) for time $t = \left(t_1 + t_2\right)/2$. This value of W_{eq} will be referred to as W_{sat} since it corresponds to the saturation region of the top transistor in the chain.

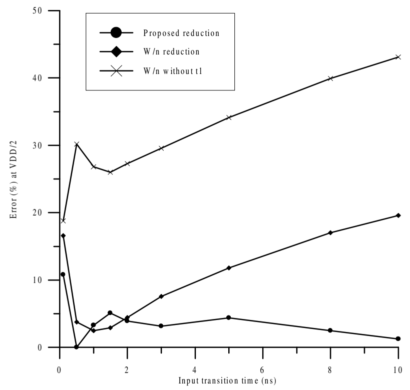
On the other hand, when all transistors in the chain operate in the linear region, the chain can be considered as a voltage divider with uniform distribution of the output voltage among all drain/source nodes. Thus, the width of the equivalent transistor can be calculated as $W_{lin} = \frac{W}{n+1}$ for non-tapered chains.

Since the transistor width for each operating condition is known, the transistor chain can be modelled 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}. But since the aim was to provide an equivalent width that would be useful for the complete region of operation in order to simplify the overall analysis, 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 interval 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 and has to be discharged. The current through the top transistor in the chain (M_{n+1}) can be integrated from time t_1 to time t_2 in order to obtain the charge Q_{sat} :

$$Q_{sat} = \int_{t_1}^{t_2} k_s \left(V_{in} - \theta - (1+\delta)\, V_M\right)^a \quad (8)$$



(a)



(b)

Fig.4: **(a) Output waveforms of the complete chain, the proposed single equivalent transistor, the conventional method and the latter without t_1 correction, for a six transistor chain and τ=3ns. (b) Percentage error at $V_{DD}/2$ between actual output response and the single equivalent transistor output response for different approaches**

The above integral will be split into two, in case the top transistor exits saturation after the input has reached its final value (fast inputs). A saturation coefficient c_{sat} can now be calculated as :

$$c_{sat} = \frac{Q_{sat}}{Q_{total}} \tag{9}$$

Consequently, the fraction of charge that is discharged through the chain during the time where all transistors operate in linear mode over the total charge is 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:

$$W_{eq} = c_{sat} \cdot W_{sat} + c_{lin} \cdot W_{lin} \tag{10}$$

It should be noted that when the input is fast and the number of transistors in the chain is large, the conventional single equivalent transistor whose width ($W_{eq_{conv}}$) is calculated using the *n*-times transconductance reduction gives acceptable results which are close to the results obtained when the width is calculated according to the proposed method in this paper. However, the time when the chain starts conducting (t_1) is crucial and in order to obtain accurate results, it should be taken into account in both cases. An example for a six transistor chain is shown in Fig. 4 together with the error (%) of the two-approaches at half-V_{DD} point of the output waveform of the complete chain.

The above analysis has also been applied for modelling of NAND/NOR gates, where the parallel transistors are replaced by an equivalent one [7]. Because of the short-circuit current, the rate of the load charging/discharging decreases, resulting in an extension in time of the saturation region of the top transistor in the chain. The new time bound of this region can be accurately estimated if the short-circuit current expression is inserted in eq. (4). However, for simplicity, t_2 is calculated for negligible short-circuit current (as previously), while the equivalent transistor width for the saturation region is calculated from eq. (7) at t_2, compensating the error from the underestimation of t_2. As it has been observed from SPICE simulations, this leads to a sufficient modelling of NAND/NOR gates by an equivalent inverter, especially for fast inputs.

5. Single Effective Input Extraction

Definitions: n+1 input ramps which are applied to the n+1 transistors of a chain and have the same transition time and starting time, will be referred to as *normalized inputs* and the single equivalent one as *normalized input.* Additionally, every set of input ramps (less than n+1) which have the same transition time and the same starting point, will be referred to as *equal ramps.*

The proposed algorithm aims to the extraction of n+1 normalized input ramps for each possible input pattern, so that when the normalized inputs are applied to the transistor gates of the chain, the chain will have the same output response with that of the actual inputs. Before the proposed algorithm can be applied, the "weight" of each transistor in the chain has to be calculated, i.e. a coefficient which corresponds to the position of each input (or combination of inputs) and becomes larger for transistors closer to the output. In order to find the weight of each position in the chain, equal ramps are applied to the transistors for which the weight coefficients are to be calculated and the inputs of the rest of the transistors are set to V_{DD}. For each case of input patterns, a coefficient is derived with whom the transition time of the applied ramp must be multiplied so that when the resulted ramp is applied to all transistors, the evolution of the output will be the same. The next three steps of the mapping algorithm should be applied for every possible input pattern :

Step 1. Inputs which efficiently act and should be treated as V_{DD} voltages have to be identified. In order to achieve this, every input ramp which at time t=t_m has a value larger than $\frac{2}{3}V_{DD}$ should be considered V_{DD} for the following steps.

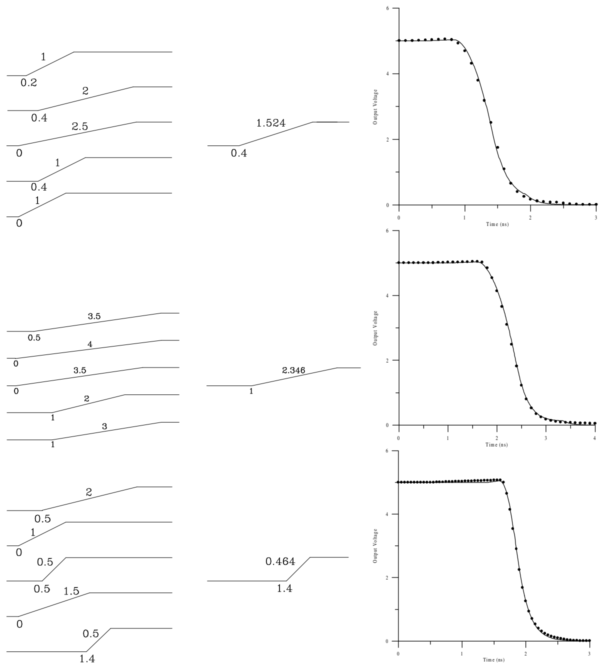


Fig. 5: **Actual set of inputs (left column) and normalized equivalent one (right column) together with the corresponding output waveform of a chain for each case. The dots correspond to the actual set of inputs**.

Time t_m occurs when the last ending input ramp reaches $V_{DD}/2$. In case two or more inputs end at the same time, t_m is measured on the one that starts last.

Step 2. The m ramp inputs that remain from step 1 have to be transformed to equal ramps so that the algorithm can proceed. The starting point of these equal ramps (t_0) is taken as $t_0=\max(t_1, t_2, ..., t_{n+1})$ where $t_1, t_2, ..., t_{n+1}$ are the starting point of **all** input ramps in the chain. This is reasonable since the chain does not conduct current before the starting point of the last changing input. In order to take into account the slope as well the time during which an input is in transition after time t_0, the transition time (T_{eq}) of the equal ramps is taken as :

$$T_{eq} = \frac{\sum_{i=1}^{m}\left[1-\frac{V_i[t_0]}{V_{DD}}\right]\left(t_{e_i}-t_0\right)}{m} \tag{11}$$

where $V_i[t_0]$ is the voltage that each input ramp has reached at the initial time and t_{e_i} is the time point at which each of the m input ramps reaches V_{DD}. Obviously, because of step 1, $t_{e_i} > t_0$ for all of the m inputs.

Step 3. The input pattern which is now applied to the chain, consists of equal ramp inputs and V_{DD} inputs. These can be mapped to normalized inputs using the weight coefficients which correspond to the transistor positions in the chain which receive a ramp input. Thus, the transition time of the normalized inputs which are applied to the chain at $t=t_0$ is :

$$T_{eff} = c_{weight} \cdot T_{eq} \tag{12}$$

The same ramp input (normalized input) is finally applied to the single equivalent transistor.

In Fig. 5, a comparison of the output responses of a five transistor chain to the actual and normalized input patterns is presented, which shows the accuracy and efficiency of the presented algorithm. The above algorithm presented very good match between the output response of the complete chain and the single equivalent transistor for a wide range of input transition times and relative distances in time of their starting points.

6. Conclusion

The widely used method for collapsing a transistor chain to a single equivalent transistor that can be used for the modelling of CMOS gates, has been analyzed and improved. Its width has been efficiently calculated taking into account the regions of operation of the structure. Several parameters such as the time when the chain starts conducting, short channel current models, body effect and the input transition time have been incorporated in the model. In addition, an efficient and accurate algorithm has been proposed in order to obtain a single equivalent input from every possible input pattern to a chain. The proposed model extracts the output waveform of a chain with very good accuracy compared to SPICE simulations.

References

[1] L. Bisdounis, S. Nikolaidis and O. Koufopavlou, "Analytical Transient Response and Propagation Delay Evaluation of the CMOS Inverter for Short-channel Devices", to appear in IEEE J. Solid-State Circuits.

[2] 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.

[3] M. Shoji, "FET Scaling in Domino CMOS Gates", IEEE J. of Solid-State Circuits, vol. SC-20, no. 5, pp. 1067-1071, October 1985.

[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] B. S. Cherkauer and E. G. Friedman, "Channel Width Tapering of Serially Connected MOSFET's with Emphasis on Power Dissipation", IEEE Trans. on Very Large Scale of Integration (VLSI) Systems, vol. 2, no.1, pp. 100-114, March 1994.

[6] T. Sakurai and A. R. Newton, "Delay Analysis of Series-Connected MOSFET Circuits", IEEE J. of Solid-State Circuits, vol. 26, no. 2, pp. 122-131, February 1991.

[7] Y.-H. Jun, K. Jun and S.-B. Park, "An Accurate and Efficient Delay Time Modeling for MOS Logic Circuits Using Polynomial Approximation", IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 8, no. 9, pp. 1027-1032, September 1989.

[8] 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.