MODELING THE OPERATION OF THE TRANSISTOR CHAIN IN CMOS GATES

S. Nikolaidis **A. Chatzigeorgiou**
Electronics & Computers Division, Department of Physics
Aristotle University of Thessaloniki
54006 Thessaloniki, GREECE
Email : snikolaid@physics.auth.gr

Abstract: This paper introduces a detailed analysis for the operation of the transistor chain in CMOS gates. The chain is modeled by a transistor pair according to the operating conditions of the structure. The system of differential equations for the derived chain model is solved and analytical expressions which accurately describe the temporal evolution of the output voltage, are extracted. For the first time a fully mathematical analysis without simplified step inputs and linear approximations of the output waveform and without resistors replacing transistors, is presented. The final results for the calculated response and the propagation delay of this structure are in excellent agreement with SPICE simulations.

Keywords
Modeling, Transistor Chain, CMOS Gates

I. INTRODUCTION

Modeling of CMOS gates has attracted the interest of many researchers during the last years, mainly because speed and power estimation in the early phases of a design is becoming increasingly important. Much effort has been devoted to the investigation of the behaviour of the CMOS inverter and analytical expressions for its output response have been derived [1,2]. In spite of this work, little has been done about more complicated gates because of their multinodal circuitry and multiple inputs.

Series connected MOSFETs form a basic structure in NAND/NOR gates and their operation is substantially more complicated than that of parallel transistors. The analysis of a transistor chain is intricated by the fact that differential equations must be solved for several nodes and for different modes of operation for the transistors according to their position in the chain and the input waveforms. Transistor models that accurately describe the characteristics of a submicron device [1] are difficult to handle within a system of differential equations. On the other hand, simple models fail to predict the response of a transistor within acceptable limits. Therefore, the inevitable engineering compromise between accuracy and simplicity has to be made, if analytical expressions are required.

A qualitative description of the behaviour of serially connected transistors applied to domino CMOS gates was given by Shoji [3]. Pretorius et al [4] simplified a part of the transistor chain by a resistor, thus limiting the accuracy. Kang and Chen [5] developed more accurate expressions for the output waveform but used linear approximations and only step inputs. Applying the *n*th power law for submicron devices Sakurai [6] developed expressions for a CMOS inverter and extended the model to gates proposing a delay degradation factor. Cherkauer and Friedman [7] performed their analysis using simplified long channel models and applying step inputs in order to optimize channel widths for low power. Nabavi-Lishi and Rumin [8] presented a semi-empirical method for collapsing the complete transistor chain to a single equivalent transistor, resulting in limited accuracy. By the same way Daga et al [9] developed their analysis for an inverter and extension to gate is performed by defining an equivalent drivability factor for the case of a transistor chain, which uses simplified assumptions for the modes of operation of the transistors.

In this paper analytical expressions for the output response of a MOSFET chain to input ramps are being derived, without the simplifications of previous works. The MOSFET model for short channel devices proposed in [10] is slightly modified for the linear region, improving the accuracy. The transistor chain is reduced to two serially connected transistors, where the one closer to the output remains unchanged and the other is the equivalent of the rest of the transistors. In this way, differential equations can be solved analytically, obtaining very good agreement between simulated and calculated results. This is the first time transistors are treated without replacement by resistors and for real input ramps.

II. TRANSISTOR CHAIN MODEL

Our analysis is performed for a chain of serially connected NMOS transistors, as shown in Fig. 1a. Therefore, the temporal evolution of the output voltage across a load capacitance that discharges through the chain, is examined. The case of a PMOS chain is symmetrical. At each node, the parasitic capacitance, formed by the diffusion regions of the transistors, is shown. Instead of the simplified input pattern that other authors have used, ramp inputs applied to the gate of all transistors are considered, which corresponds to the worst case (slower) for the output response. It is assumed, without affecting the final results, that all internal node capacitances are discharged.

The topmost transistor in the chain (M_0), begins its

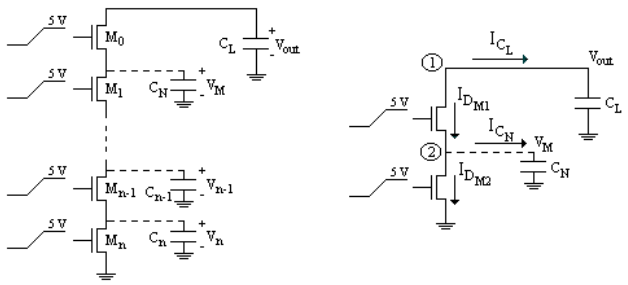


Fig. 1 (a) Complete chain of NMOS transistors and (b) proposed equivalent chain

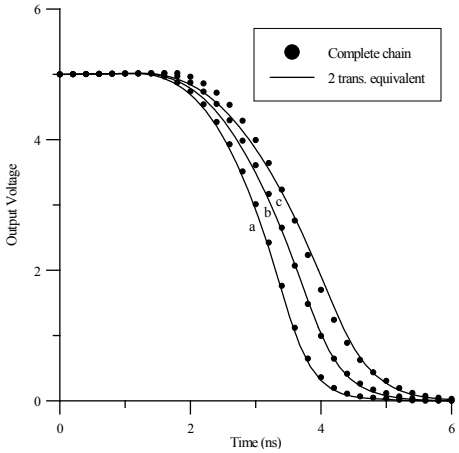


Fig.2 Output waveform comparison between the complete transistor chain and the equivalent transistor pair for 3 (a), 4 (b) and 5 (c) transistors

operation in the saturation region, since its drain to source voltage V_{DS} is initially V_{DD}. As the load capacitance (C_L) discharges and the internal node capacitance C_N charges, transistor M_0 enters the linear region when $V_{DS} = V_{D\text{-}SATN}$, where $V_{D\text{-}SATN}$ is its drain saturation voltage. The rest of the transistors operate in the linear region without ever leaving this region. That is because their V_{DS} never exceeds the drain saturation voltage, for real inputs [7]. Since the current of the transistors that operate in the linear region increases as the voltage at the intermediate nodes rises and the output voltage decreases, there will be a time point where the current of the saturated top transistor will be equal to the current of the bottom transistors. From this time on, the structure remains at this state until the charge across the load capacitance is no more adequate to keep the topmost transistor in saturation. During this period, the voltage at the source of all transistors remains constant. This is the state which Kang and Chen [5] refer to as the "plateau" voltage and is apparent for very fast inputs since intermediate nodes remain at this potential for a reasonable time (Fig.3b), but has a very short duration for slower inputs. Its significance is that it represents the inertia of the circuit at the point where upper and lower transistor at each node, supply the same current without leaving this state.

Since the number of differential equations that have to be solved in order to obtain an analytical expression for the output waveform of a transistor chain is prohibitive, the number of transistors has to be reduced. A good approximation is to replace all transistors that operate in the linear region by an equivalent one and to solve the problem for the case of two transistors (Fig. 1b) where the upper operates in the saturation and linear region and the bottom only in the linear region. We have found that for n serially connected transistors operating in the linear region, the equivalent transistor width is given by :

$$\frac{1}{W_{eq}} = \frac{1}{W_1} + \frac{1}{W_2} + \ldots + \frac{1}{W_n} \qquad (1)$$

The response of the equivalent circuit matches very well the output waveform of the complete chain as confirmed by SPICE simulations (Fig. 2). Error less than 6% has been observed for chains up to 5 transistors.

Therefore, our problem can be focused on a two node-analysis which decreases the complexity of the solution significantly.

Sufficiently accurate models have been proposed for short channel devices [1] but their expressions are difficult to handle within differential equations. As the current dependance on the gate-to-source voltage tends to be linear for deep submicron devices, the current model proposed in [10] has been chosen for this analysis. The current expression for the linear region is slightly modified by neglecting the quadratic term of V_{DS}, leading in lower complexity and increased accuracy.

NMOS transistor currents are given by the following equations :

$I_n = 0$, $\quad V_{GS} < V_{TN}$, Cutoff (2)

$I_n = \beta_n \, k_s \left(V_{GS} - V_{TN}\right)$ $\quad V_{DS} > V_{D\text{-}SATN}$, Saturation (3)

$$I_n = \frac{\beta_n}{1 + k_1 V_{DS}} \left(V_{GS} - V_{TN}\right) V_{DS} \quad V_{DS} \le V_{D\text{-}SATN}, \text{ Linear} \qquad (4)$$

where β_n is the NMOS device gain factor, V_{TN} is the threshold voltage and $k_{s,1}$ are constants which specify the effects of carriers velocity saturation. These are calculated for a given technology, by measurements on the I-V_{DS} characteristics in SPICE.

The threshold voltage is expressed by :

$$V_{TN} = V_{TO} + \gamma\left(\sqrt{2\varphi_F + V_{SB}} - \sqrt{2\varphi_F}\right) \qquad (5)$$

where V_{TO} is the zero bias threshold voltage, γ is the constant that describes the body effect, φ_F is the bulk potential and V_{SB} is the source to substrate voltage. In order to transform the above expression into a simplified one that can be treated mathematically, a first order Taylor series approximation around V_{SB}=1V is satisfactory (approximation error less than 7%) :

$$\widetilde{V}_{TN} = V_{TN}\big|_{V_{SB}=1} + \left(V_{TN}\right)'\Big|_{V_{SB}=1}\left(V_{SB} - 1\right) = \theta + \delta V_{SB} \qquad (6)$$

The NMOS saturation voltage is now calculated by equating equations (3) and (4). The two currents are equal when the transistor leaves the linear region and enters

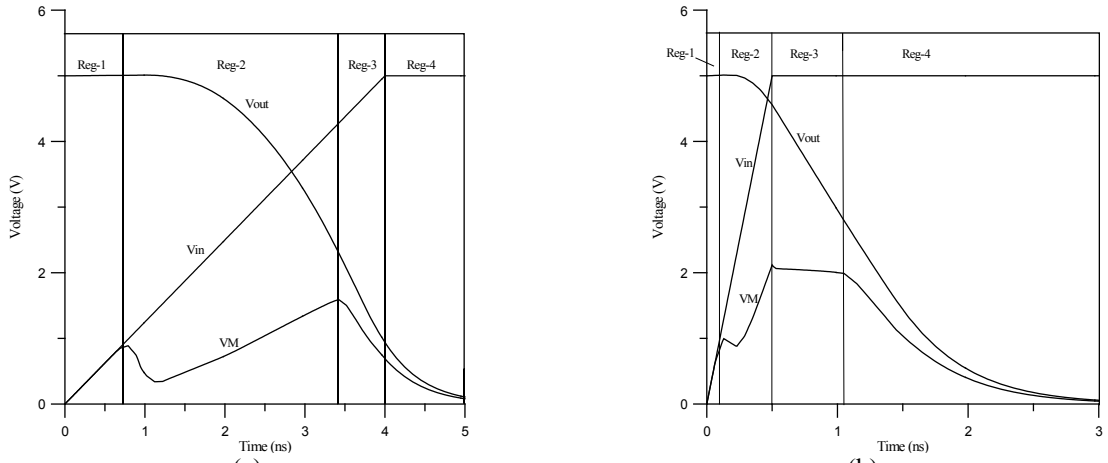


Fig.3 Regions of operation for slow (a) and fast (b) inputs

saturation for $V_{DS} = V_{D\text{-}SATN}$. This yields : $V_{D-SATN} = \frac{k_s}{1\text{-}k_1 k_s}$, which is valid with reasonable accuracy for submicron devices.

III. OUTPUT WAVEFORM ANALYSIS

The input applied to the gate of the transistors is assumed to be a ramp :

$$V_{in} = \begin{cases} 0, & t \le 0 \\ V_{DD} \cdot \frac{t}{\tau}, & 0 < t \le \tau \\ V_{DD}, & t > \tau \end{cases} \quad (7)$$

where τ is the input rise time. The differential equations that describe the operation of the circuit in Fig. 1b are derived by applying Kirchhoff's current law at node 1 and 2 :

$$I_{C_L} = -I_{D_{M1}} \Rightarrow C_L \frac{dV_{out}}{dt} = -I_{D_{M1}} \quad (8)$$

$$I_{D_{M1}} = I_{D_{M2}} + I_{C_N} \Rightarrow -C_L \frac{dV_{out}}{dt} = I_{D_{M2}} + C_N \frac{dV_M}{dt} \quad (9)$$

where V_M is the voltage at the intermediate node and C_N which is the diffusion capacitance between two transistors, can be calculated as a function of "base" area and "sidewall" periphery.

Two cases, slow and fast input ramps will be considered. For slow (fast) inputs the intermediate node attains its maximum value before (after) the input ramp reaches V_{DD}. One typical waveform example of each case, is shown in Figure 3.

Slow inputs

Region 1. The circuit operates in this region until the input ramp reaches the threshold voltage at time $t_1 = \frac{V_{TO}\,\tau}{V_{DD}}$. Both transistors are off and output voltage remains at V_{DD}.

Region 2. In this region the upper transistor is saturated and the bottom operates in the linear region. It extends from time t_1 to the time point where the top transistor exits saturation (t_2). This is the region where the transistor chain remains for most of the time. For node 2, differential equation (9), becomes :

$$\beta_u k_s \left(\frac{V_{DD}}{\tau} t - \theta - (1+\delta) V_M \right) = \frac{\beta_b}{1+k_1 V_M} \left(\frac{V_{DD}}{\tau} t - V_{TO} \right) V_M + C_N \frac{dV_M}{dt} \quad (10)$$

where β_u and β_b are the transistor gain factors of the upper and bottom transistors respectively. Since the system of differential equations for the two nodes, cannot be solved analytically in this region, we assume that V_M is linear, which is a good approximation even when the number of the transistors in the chain is large, except for a small discrepancy close to the starting point of the region, which does not have any significant effect on the final solution. Setting $V_M=0$ for $t=0$:

$$V_M = a \cdot t \quad (11)$$

Substituting eq. (11) into eq. (10) and solving the resulting equation for α, gives α as a function of time. Equation (10) should be validated for every value of t and a reasonable approximation is to set $t=\tau/2$ in order to obtain the slope of V_M.

Differential equation (8) at the output node, when V_M is substituted, has the solution :

$$V_{out} = \frac{q_1}{2} \cdot t^2 + q_2 \cdot t + V_{DD} \quad (12)$$

where $q_1 = \frac{\beta_u k_s}{C_L} \left[(1+\delta) \cdot a - \frac{V_{DD}}{\tau} \right]$ and $q_2 = \frac{\beta_u k_s}{C_L}\,\theta$

The limit of this region (t_2) is computed by :

$$V_{D-SATN} = V_{out}[t_2] - V_M[t_2]$$

Very good estimation of the output response is achieved for region 2. This is crucial and significantly determines the accuracy of the overall solution as region 2 has the longest duration.

Region 3. Both transistors are in the linear region and the input is still a ramp. Therefore, this region lasts until V_{in} reaches its final value at time τ. The differential equation at the output node becomes :

$$C_L \frac{dV_{out}}{dt} = \quad (13)$$

$$-\frac{\beta_u}{1+k_1\left(V_{out}-V_M\right)}\left[V_{in}-\theta-\left(1+\delta\right)V_M\right]\left(V_{out}-V_M\right)$$

Since the system of differential equations that governs the operation of the circuit cannot be solved analytically, some approximations have to be made. V_{in} has almost reached its final value and it can be replaced by the average value $\widehat{V}_{in}=\frac{V_{in}\big|_{t=t_2}+V_{DD}}{2}$. The term ($V_{out}$-$V_M$) in the denominator can be substituted by its value at $t=t_2$ since V_{out} and V_M are known from the previous region, and similarly V_M in the (V_{in}-θ-(1+d)V_M) term by its value at $t=t_2$ which is also the maximum attainable voltage for the intermediate node and corresponds to the "plateau" voltage V_p. Thus, setting $h_1=\frac{\beta_u}{1+k_1\left(V_{out}-V_M\right)\big|_{t=t_2}}$ and $u=\widehat{V}_{in}-\theta-\left(1+\delta\right)\cdot V_p$ the solution of the above differential equation for V_M is :

$$V_M=\frac{C_L}{h_1\cdot u}\cdot\frac{dV_{out}}{dt}+V_{out} \qquad (14)$$

The differential equation at node 2 is :

$$-C_L\frac{dV_{out}}{dt}=\frac{\beta_b}{1+k_1V_M}\left(V_{in}-V_{TO}\right)V_M+C_N\frac{dV_M}{dt} \qquad (15)$$

By the same way, substituting $V_p=V_M[t_2]$ for V_M in the denominator, $\widehat{V}_{in}$ for V_{in} and eq. (14) into eq. (15), we obtain :

$$V_{out}=c_1\cdot e^{\frac{-g_2+\sqrt{g_2^2-4g_1g_3}}{2g_1}t}+c_2\cdot e^{\frac{-g_2-\sqrt{g_2^2-4g_1g_3}}{2g_1}t} \qquad (16)$$

where: $g_1=\frac{C_N\cdot C_L}{h_1\cdot u}$, $g_2=C_L+C_N+\frac{h_2\cdot\left(\widehat{V}_{in}-V_{TO}\right)\cdot C_L}{h_1\cdot u}$,

$$g_3=h_2\cdot\left(\widehat{V}_{in}-V_{TO}\right),\ h_2=\frac{\beta_b}{1+k_1V_p}$$

Since the exponent of the second term of the above expression for the V_{out} is very small, it can be neglected without affecting the output waveform. Constant c_1 that remains, is calculated by equating the modified eq. (16) with (12), for time $t=t_2$ (boundary between region 2 and 3). The above equation gives waveforms very close to SPICE simulations, which indicates the validation of the above approximations.

Region 4. In this region V_{in} has reached its final value and both transistors operate in the linear region. The same type of approximations as in the previous region can be made. Thus, solving for V_M the differential equation at the output node gives :

$$V_M=\frac{C_L}{f_1}\frac{dV_{out}}{dt}+V_{out} \qquad (17)$$

where $f_1=\frac{\beta_u\cdot\left[V_{DD}-\theta-\left(1+\delta\right)\frac{V_p}{2}\right]}{1+k_1\frac{V_p}{2}}$ and V_p is the "plateau" voltage. The differential equation at the intermediate node has the solution :

$$V_{out}=c_3\cdot e^{\frac{-p_2+\sqrt{p_2^2-4p_1p_3}}{2p_1}t}+c_4\cdot e^{\frac{-p_2-\sqrt{p_2^2-4p_1p_3}}{2p_1}t} \qquad (18)$$

$$p_1=\frac{C_L\cdot C_N}{f_1},\ p_2=C_L+C_N+\frac{f_2\cdot C_L}{f_1},\ p_3=f_2,$$

where: $f_2=\frac{\beta_b\cdot\left(V_{DD}-V_{TO}\right)}{1+k_1\frac{V_p}{2}}$

Again, the second term of the solution can be neglected since its exponent is extremely small. c_3 is calculated by equating the modified eq. (16) and (18) for $t=\tau$.

Fast inputs

Region 1. As in the case of slow inputs the output voltage remains at V_{DD} until the input ramps reach the threshold voltage.

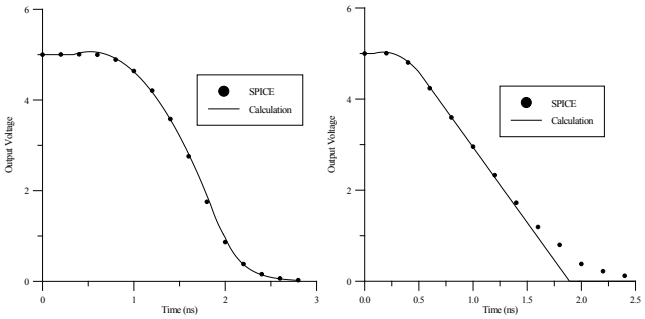
Region 2. The top transistor is saturated and the bottom is in the linear region. In this case, the limit of this region is the end of the input ramp at time τ. At this time, when input has reached V_{DD}, the intermediate node reaches its maximum voltage, V_p, which is the "plateau" voltage. The system of differential equations can be solved as in the case of slow inputs, approximating the voltage at the intermediate node by a linear function of time. The expression of V_M can be substituted in the differential equation at the output node, resulting in a similar expression for V_{out}.

Region 3. The circuit remains in this region until the top transistor exits saturation. During this time, the intermediate node voltage is constant and equal to the "plateau" voltage. The solution of the differential equation at the output node is shown below :

$$V_{out}=c-\frac{\beta_u k_s\left[V_{DD}-\theta-\left(1+\delta\right)V_p\right]}{C_L}\cdot t, \qquad (19)$$

where c is found by setting eq. (19) equal to $V_{out}[\tau]$, which is taken from the solution in the previous region. The limit of this region (t_2) is given by $V_{D-SATN}=V_{out}\left[t_2\right]-V_M\left[t_2\right]$, when the top transistor exits saturation.

Region 4. Both transistors operate in the linear region and $V_{in}=V_{DD}$. Since the time range of this region is wider than for slow inputs and the approximations used there do not lead to results with acceptable accuracy, we extend the solution of the previous region linearly. The accuracy of this linear approximation is confirmed from the very good agreement between calculated and simulated values, down to very small output voltage values. This region lasts until the calculated V_{out} crosses the time axis. From this time on, the output voltage is zero.



(a) (b)

Fig. 4 Output waveform comparison between simulated and calculated results
(a) Slow input, $\tau = 2$ ns (b) Fast input, $\tau = 0.5$ ns

Whether an input ramp is slow or fast can be determined by solving $V_{D-SATN} = V_{out}[t_2] - V_M[t_2]$, in the second region. If the top transistor exits saturation before the input reaches its final value ($t_2 < \tau$), the input is slow, otherwise it should be considered fast.

Results and Delay Comparison

The expressions that have been calculated for the output waveform of the transistor chain, match the SPICE simulation results very well, as shown in Figure 4, which is a comparison for slow and fast inputs between calculated and simulated output voltage values, for the AMS 0.8μm technology. The small error that can be observed, proves the accuracy of the extracted expressions and the validity of the proposed reduction of the transistor chain to two equivalent transistors, according to their mode of operation.

Since the output expression for each of the above regions of operation is known, propagation delay for the discharging case (t_{PHL}) can be calculated as the time from the half-V_{DD} point of the input to the half-V_{DD} point of the output. In the charging case, the delay t_{PLH} is defined in the same way. The region in which $V_{DD}/2$ of the output occurs, can be found by comparing it with $V_{out}[t_2]$ and $V_{out}[\tau]$. Using this definition, delay results for several input waveforms and transistor chains, compared with simulation results, are presented in Table I. It is observed that in all cases the delay computed using the analytical expressions is within 2.5 % of the delay computed by SPICE.

IV. CONCLUSION

A detailed analysis of a transistor chain as it appears in CMOS gates has been introduced. Accounting for real operation conditions, analytical expressions for the output response of a discharging chain have been derived. Using a model that reduces the number of transistors in the chain to two, it has been possible to solve the differential equations that describe the system without simplified approximations. Output voltage and propagation delay results derived by the proposed analytical method, match very well SPICE simulation results.

Table I. Delay comparison (in ns) between calculated and simulated values for several transistor chains and input slopes

	2 Transistors		3 Transistors		4 Transistors	
	Calculation	SPICE	Calculation	SPICE	Calculation	SPICE
$\tau = 0.5$	0.465	0.455	0.596	0.602	0.770	0.780
$\tau = 1$	0.537	0.538	0.688	0.688	0.842	0.847
$\tau = 1.5$	0.615	0.611	0.790	0.780	0.947	0.940
$\tau = 2$	0.670	0.655	0.845	0.858	1.048	1.032

V. 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, April 1990, pp. 584-594.

[2] L. Bisdounis, S. Nikolaidis, O. Koufopavlou, C. Goutis, "Accurate Timing Model for the CMOS Inverter", in Proc. of ICECS, vol. 1, 1996, pp. 89-92.

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

[4] J. A. Pretorius, A. S. Shubat and C. A. T. Salama, "Analysis and Design Optimization of Domino CMOS Logic with Application to Standard Cells", IEEE J. Solid-State Circuits, vol. SC-20, no. 2, April 1985, pp. 523-530.

[5] 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, 1990, pp. 289-306.

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

[7] 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, March 1994, pp. 100-114.

[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, October 1994, pp. 1271-1279.

[9] J. M. Daga, S. Turgis and D. Auvergne, "Design Oriented Standard Cell Delay Modelling", in Proc. of PATMOS, 1996, pp. 265-274.

[10] Y. Tsividis, "Operation and Modeling of the MOS transistor", McGraw-Hill International, 1988.