ANALYSIS OF THE TRANSISTOR CHAIN OPERATION IN CMOS GATES FOR SHORT CHANNEL DEVICES

A. Chatzigeorgiou and S. Nikolaidis [1]

Computer Science Department, [1]Department of Physics
Aristotle University of Thessaloniki
54006 Thessaloniki, Greece

ABSTRACT

A detailed analysis of the transistor chain operation in CMOS gates is presented. The chain is diminished to a transistor pair taking into account the actual operating conditions of the structure. The output waveform is obtained analytically, without linear approximations of the output voltage and for ramp inputs. The α-power transistor current model which takes into account second order effects of submicron devices is used, while previous inconsistencies in the chain currents are eliminated by introducing a drain-to-source voltage modulation factor. The exact time when the chain starts conducting is efficiently calculated removing a major source of errors. The calculated output waveform results according to the proposed model are in excellent agreement with SPICE simulations.

1. INTRODUCTION

Since the need for analytical methods which can accurately perform timing simulations of digital integrated circuits is growing as the minimum feature sizes decrease and the number of transistors per chip increases, modeling of CMOS gates is becoming important. It has been extensively pointed out, that simulators such as SPICE which are based on numerical methods, are excessively slow for large designs. Motivated by the previous observations, much research effort has been devoted to the investigation of the behavior of the CMOS inverter and well defined expressions for its output response have been obtained [1], [2], [3].

However little has been done on more complicated gates such as NAND/NOR gates because of their multinodal circuitry and multiple inputs. Modeling of these gates is intricated mainly by the operation of the transistor chain through which the output load is discharged (NAND) or charged (NOR). Since the timing behavior of such a chain cannot be obtained by solving a differential equation at each node of the structure, the inherent properties and operating conditions of the chain have to be exploited. All previous attempts to model the transistor chain can be categorized in two main groups :

The most usual one is the replacement of the complete chain by a single equivalent transistor. As a rule of thumb, the width of the equivalent transistor is calculated by a single *m*-times transconductance reduction, where *m* is the number of the devices in the chain. Although attempts have been made in order to improve the efficiency of this model incorporating parasitic capacitances [4], the single equivalent transistor replacement generally fails to reproduce the output waveform of the chain, since it does not take into account the actual operating conditions of the structure.

The next step that has been taken in search for a better modeling technique was to replace a part of the transistor chain, namely those devices which operate always in the linear region, by an equivalent resistor. Such models have been presented by [5], [6]. However these techniques are based on simplified approximations and lead to prohibitively inaccurate results.

It should be mentioned that all previously reported methods ignore second order effects that are present in submicron devices, assume only step inputs and present inconsistency in the chain currents, which is the main error in existing modeling techniques [7].

In this paper, a different approach is followed, overcoming the inaccuracies of all previous works. Nonsaturated devices are replaced by an equivalent transistor whose width is calculated efficiently without leading to inconsistent currents. The method is presented for non-zero transition time inputs, short channel transistor current models and the exact time point when the chain starts conducting is calculated, eliminating another main source of errors.

2. TRANSISTOR CHAIN OPERATION

In order to study the operation of the transistor chain in CMOS gates, let us consider the circuit of Fig. 1a where the discharging of a load capacitance (C_L) through the NMOS transistor chain is examined. Charging through a PMOS chain is symmetrical. The parasitic capacitances formed by the drain/source diffusion areas are also shown. A common ramp input is applied to the gates of all transistors in the chain :

$$V_{in} = \begin{cases} 0 & t < 0 \\ \dfrac{V_{DD}}{\tau} \cdot t & 0 \le t \le \tau \\ V_{DD} & t \ge \tau \end{cases} \qquad (1)$$

where τ is the input transition time. All internal nodes are considered to be initially discharged. In case the nodes are charged at t=0, the output waveform can be obtained by shifting it in time according to the charge that was initially stored in all nodes [5] and will not be discussed here.

In order to take into account second order effects of submicron devices, the *a*-power model [2] has been used for the transistor currents :

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

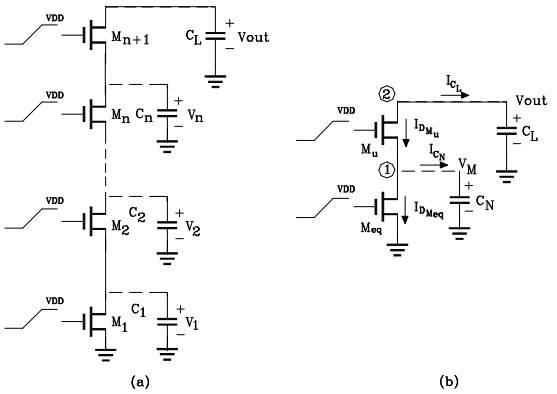


Fig. 1: (a) Complete transistor chain and (b) two-transistor equivalent chain

where V_{D-SAT} is the drain saturation voltage, k_l , k_s are the transconductance parameters which depend on the width to length ratio of a transistor, a is the carrier velocity saturation index and V_{TN} is the threshold voltage which is approximated by its first order Taylor series approximation around V_{SB}=1V, $\widetilde{V}_{TN} = \theta + \delta \cdot V_{SB}$.

The topmost transistor in the chain (M_{n+1}) operates initially in saturation since its drain-to-source voltage (V_{DS}) is higher than the drain-to-source saturation voltage ($V_{D\text{-}SAT}$). As the output load capacitance discharges and the internal node voltages rise, transistor M_{n+1} will enter the linear mode of operation when V_{DS}=$V_{D\text{-}SAT}$. All other transistors of the chain operate always in linear mode, since after time t_1 when the chain starts conducting their V_{DS} never exceeds the drain saturation voltage [6].

From the time point τ when the input reaches its final value and until the time point t_2 when the topmost transistor exits saturation (in case $t_2 > \tau$), all node voltages remain constant. That is because if the node voltages were decreasing, the saturation current of the topmost transistor would increase, thus increasing the node voltages. On the other hand, if the node voltages were increasing the current of the topmost transistor would decrease thus decreasing the node voltages. Consequently, all node voltages remain at their initial potential at time τ, and this state which is known as the "plateau" state [5] is apparent for fast inputs or large output loads (Fig. 2a). During the plateau state all parasitic currents at the internal nodes are eliminated since the voltages remain constant. In this way the currents of all transistors in the chain are equal.

In order to calculate the plateau voltage of the chain, let us consider the circuit of Fig. 1a and assume that the same ramp input is applied to all transistors. Although the analysis here refers to fast input ramps where the plateau state appears, the derived results are also valid for slow inputs. A first approximation is used for the width W'_{eq} of the equivalent transistor M_{eq} in Fig. 1b, which replaces all the nonsaturated transistors and is given by :

$$\frac{1}{W'_{eq}} = \frac{1}{W_1} + \frac{1}{W_2} + \ldots + \frac{1}{W_n} \tag{3}$$

The plateau voltage at the source of the top transistor, V_p, occurs at the end of the input ramp (V_{in}=V_{DD}) where the current of the top transistor ceases to increase. Thus, V_p can be calculated by setting the saturation current of the top transistor (M_u) equal to the current of the bottom transistor (M_{eq}) which operates in linear mode :

$$k_s\left(V_{DD} - \theta - (1+\delta)V_p\right)^a = k_{l_{eq}}\left(V_{DD} - V_{TO}\right)^{a/2} V_p \tag{4}$$

The above equation can be solved with very good accuracy using a second order Taylor series approximation around V_p=1 V.

The approach of previous works is based on the assumption that there is a uniform distribution of the source voltage of the top transistor among the drain/source nodes of the rest transistors in the chain operating in linear mode. However, this is not a valid assumption as the gate-to-source voltage and the threshold voltage of each transistor in the chain are different and consequently they would not be able to drive the same current if they had equal drain-to-source voltages. For example, equating the currents through the two closer to ground transistors (for the same transistor width) for V_{in}=V_{DD} and setting the same V_{DS} for each transistor gives :

$$I_1 = I_2 \Rightarrow$$
$$k_l\left(V_{DD} - \theta\right)^{a/2} V_{DS} = k_l\left(V_{DD} - \theta - (1+\delta)V_1\right)^{a/2} V_{DS} \tag{5}$$

which results in $(1+\delta)V_1 = 0$ where V_1 is the drain voltage of the bottom transistor. This is an invalid expression, because always δ>0. Trying to keep the current of each transistor in the chain constant, the reduction in V_{GS} and the increase in V_{TN} of a transistor closer to the output is compensated by an increase in its V_{DS}. Considering a gradual increment of V_{DS} by a constant factor v (v>1), called *drain-to-source voltage modulation factor*, as we are moving closer to the output, results in very good agreement with SPICE simulations. This means that for two adjacent transistors it is $V_{DS_{(j+1)}} = v \cdot V_{DS_{(j)}}$, where the index shows the position of the transistor in the chain (Fig. 1a). In this way, equation (5) can be rewritten as :

$$k_l\left(V_{DD} - \theta\right)^{a/2} V_{DS_1} = k_l\left(V_{DD} - \theta - (1+\delta)V_{DS_1}\right)^{a/2} v\, V_{DS_1} \tag{6}$$

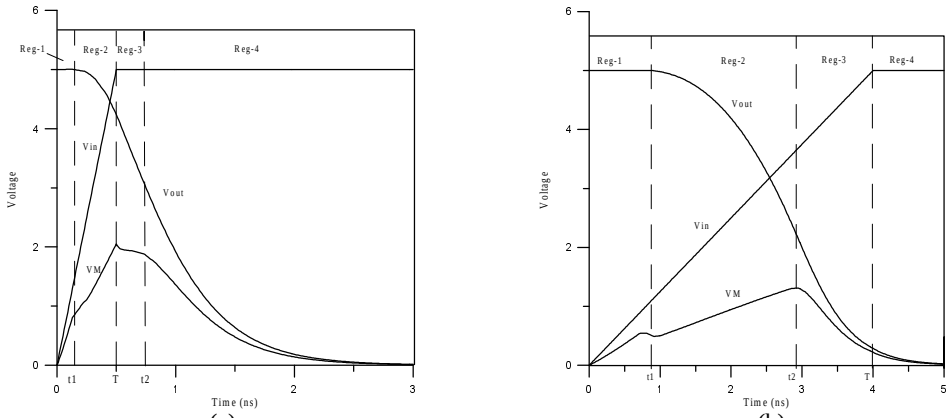
In order to solve the above equation, a first order approximation of the V_{DS_1} term inside the parenthesis in the right hand side of eq. (6) is used. Considering the part of the transistor chain which contains the nonsaturated devices as a voltage divider, that term V_{DS_1} can be set equal to V_p / n (for the case that all transistors have the same width) and eq. (6) can be solved for v resulting in :

$$v = \left[\frac{V_{DD} - \theta}{V_{DD} - \theta - (1+\delta)\left(V_p / n\right)}\right]^{a/2} \tag{7}$$

Consequently, the plateau voltage of the chain is : $V_p = \left(1 + v + \ldots + v^{n-1}\right) \cdot V_{DS_1}$. Equating the current that flows through the equivalent transistor (M_{eq} in Fig. 1b) with the current through the closest to the ground transistor of the chain (M_1 in Fig. 1a), the final width of the equivalent transistor is obtained:

$$W_{eq} = \frac{W_1}{1 + v + \ldots + v^{n-1}} \tag{8}$$

which is used in the mathematical analysis.



(a) (b)

Fig. 2: Regions of operation for (a) fast and (b) slow input ramps

The accuracy of the proposed width for the equivalent transistor is validated by comparison between the output responses of the complete chain and the two transistor chain model, as shown in Fig. 3 for an HP 0.5 μm technology. Also, a comparison with the output response, when the equivalent transistor width is calculated in the conventional way described by eq. 3 and when the nonsaturated devices are replaced by a resistor [5] is also presented in Fig. 4. The superiority of the proposed method is obvious. Consequently, the multinodal analysis problem is now diminished to a two node-analysis which decreases the complexity of the solution significantly.

3. OUTPUT WAVEFORM ANALYSIS

Because of coupling capacitance (C_M) between transistor gates and the drain/source nodes, drain voltages tend to follow the input ramp until all lower transistors start conducting. Until the time point where the transistor below a node starts conducting, the voltage waveform of that node, as it is isolated between two cut-off transistors, is derived by equating the current due to the coupling capacitance of the node , $I_{C_{M_i}}$, to the charging current of the parasitic node capacitance I_{C_i} :

$$I_{C_{M_i}} = I_{C_i} \Rightarrow$$

$$C_{M_i}\frac{dV_{in}-dV_i}{dt} = C_i\frac{dV_i}{dt} \Rightarrow V_i[t] = \frac{C_{M_i}}{C_{M_i}+C_i}V_{in}[t] \quad (9)$$

After the time at which all transistors below the *i*-th node start to conduct (t_{s_i}) and until the time at which the complete chain starts to conduct (t_1), this node is subject to two opposite trends. One tends to pull the voltage of the node high and is due to the coupling capacitance between input and the node and is intense for fast inputs and high coupling to node capacitance ratio. The other tends to pull its voltage down because of the discharging currents through all lower transistors and is more intense for nodes closer to the ground. For simplicity, here, the two trends are considered to be counterweighted which gives good results in most practical cases. Therefore, the voltage of each node after the time where all the lower transistors start conducting and until time t_1, is considered to be constant and equal to the node voltage at the beginning of this time interval.

By solving $V_{GS_i} - V_{TN_i} = 0$ for each transistor in the chain, the time at which the *i*-th transistor starts conducting (t_{s_i}) is given by the recursive expression :

$$t_{s_i} = \tau \cdot \frac{\theta + (1+\delta)\dfrac{C_{M_{i-1}}}{C_{M_{i-1}}+C_{i-1}}\dfrac{V_{DD}}{\tau}t_{s_{i-1}}}{V_{DD}} \quad (10)$$

where the index *i* corresponds to the position of the transistor in the chain and starts counting (*i*=1) from the bottom transistor. ($t_{s_0} = 0$). From the above expression, the time at which the chain starts conducting $t_{s_{n+1}} = t_1$, can be easily obtained.

It has been observed by SPICE simulations that the voltage (V_M) at the source of the top transistor is almost linear between time t_1 and time τ. According to the above, V_M will have a value V_s at time t_1 and V_p at time τ. Thus, V_M for the time interval t_1-τ can be expressed as : $V_M[t] = V_a + m \cdot t$,

where $V_a = V_s - \dfrac{V_p - V_s}{\tau - t_1}t_1$ and $m = \dfrac{V_p - V_s}{\tau - t_1}$.

Although the slope of V_M was calculated for fast inputs, it can be found exactly in the same way for slower inputs [8].

The differential equations that describe the operation of the circuit in Fig. 1b are derived by applying Kirchhoff's current law at nodes 2 and 1:

$$I_{C_L} = -I_{D_{M_u}} \Rightarrow C_L\frac{dV_{out}}{dt} = -I_{D_{M_u}} \quad (11)$$

$$I_{D_{M_u}} = I_{D_{M_{eq}}} + I_{C_N} \Rightarrow -C_L\frac{dV_{out}}{dt} = I_{D_{M_{eq}}} + C_N\frac{dV_M}{dt} \quad (12)$$

where V_M is the voltage at the intermediate node and C_N is the lumped capacitance of all diffusion capacitances of the internal nodes in the chain. Each node capacitance, C_{node} , is calculated as a function of "base" area and "sidewall" periphery [9].

The above differential equations are solved resulting in the expressions for the output voltage waveform for each operating region of the transistors in the chain.

Two cases, fast and slow input ramps are considered. For the fast (slow) case, the intermediate node voltage V_M attains its maximum value when (before) the input ramp reaches V_{DD}.

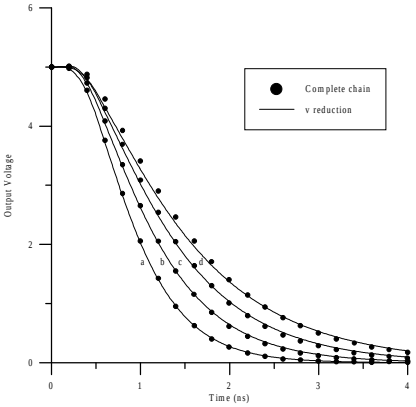


Fig. 3: Output waveform comparison between complete chain and two transistor chain, for a=3, b=4, c=5, d=6 transistors in the chain

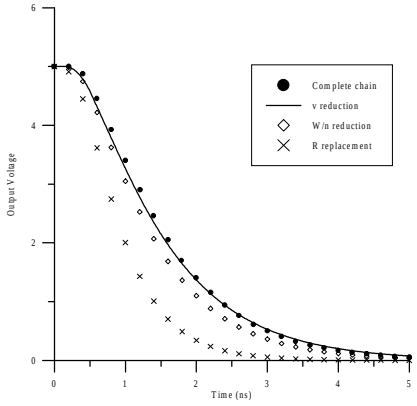


Fig. 4: Comparison between the output waveform of the complete chain and the two transistor chain model using the *v* factor, the *n*-times transconductance reduction and replacement by a resistor, for a 6 transistor chain

A. Fast input ramps

Region 1. The top transistor M_u is cut off. This region extends from time t=0 until $t=t_1$ when transistor M_u starts conducting and enters saturation. The output voltage remains at V_{DD} (Fig. 2a). This is also validated by SPICE simulations: no overshoot is observed because of the very small gate-to-drain coupling capacitance of a transistor in cut-off.

Region 2. The upper transistor is saturated and the bottom operates in linear mode. This region extends from time t_1 until $t=\tau$ when the input reaches its final value. Since the system of differential equations that describes the operation of the circuit cannot be solved analytically, V_M is considered to be linear.

Substituting $V_M[t] = V_a + m \cdot t$ into eq. (11) and solving the resulting equation gives :

$$V_{out} = c_1 + \left(q_1 \cdot t - q_2\right)^a \frac{k_s}{C_L\left(1+a\right)}\left[\frac{q_2}{q_1} - t\right] \quad (13)$$

where $q_1=(V_{DD}/\tau)-(1+\delta)m$, $q_2=\theta+(1+\delta)V_a$ and $c_1 \approx V_{DD}$.

Region 3. The input ramp has reached V_{DD}, the top transistor is in saturation and the bottom in the linear mode of operation. The limit of this region is time t_2 when the top transistor exits saturation and until that time, the intermediate node remains at the plateau voltage. Since $V_M=V_p$, differential eq. (11) gives :

$$V_{out} = c_2 - \frac{k_s}{C_L}\left[V_{DD} - \theta - \left(1+\delta\right) V_p\right]^a \cdot t \quad (14)$$

where $c_2 = V_{out}\big|_{t=\tau} + \frac{k_s}{C_L}\left[V_{DD} - \theta - (1+\delta)V_p\right]^a \cdot \tau$.

The limit of this region is computed by solving $V_{D-SATN}[t_2] = V_{out}[t_2] - V_p$ for the upper transistor, where $V_{D-SATN}[t] = \frac{k_s}{k_l}\left(V_{GS} - V_{TN}\right)^{a/2}$ according to [2].

Region 4. Both transistors operate in linear mode. The system of differential equations becomes :

$$C_L \frac{dV_{out}}{dt} = -k_{lu}\left(V_{DD} - \theta - \left(1+\delta\right) V_M\right)^{a/2}\left(V_{out} - V_M\right) \quad (15)$$

$$-C_L \frac{dV_{out}}{dt} = k_{lb}\left(V_{DD} - V_{TO}\right)^{a/2} V_M + C_N \frac{dV_M}{dt} \quad (16)$$

where k_{lu}, k_{lb} specify the linear region transconductances for the upper and bottom transistors respectively. Since the above system cannot be solved analytically, V_M in eq. (15), in the term that is powered to $a/2$, is replaced by its average value $V_p/2$. Solving eq. (15) for V_M, substituting the resulting expression in eq. (16), and setting $g_1 = k_{lu}\left(V_{DD} - \theta - \left(1+\delta\right)\frac{V_p}{2}\right)^{a/2}$ and $g_2 = k_{lb}\left(V_{DD} - V_{TO}\right)^{a/2}$ results in a second order differential equation which has the solution :

$$V_{out} \cong c_3 \cdot e^{\frac{-p_2+\sqrt{p_2^2-4p_1p_3}}{2p_1}t} \quad (17)$$

where $p_1 = \frac{C_N \cdot C_L}{g_1}$, $p_2 = \frac{C_L \cdot g_2}{g_1} + C_L + C_N$, $p_3 = g_2$

and c_3 is calculated by equating the above equation for $t=t_2$ with $V_{out}[t_2]$ which is obtained from the previous region.

B. Slow input ramps

For slow input ramps the analysis can be performed in the same way, except for region 3 ($t_2<t<\tau$) since the top transistor exits saturation before the input reaches V_{DD} (Fig. 2b). For this time interval the input has to be approximated by its average value and the analysis can proceed as in region 4 for fast inputs.

Whether an input ramp is slow or fast can be determined by solving $V_{D-SATN}[t_2] = V_{out}[t_2] - V_M[t_2]$ 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.

The previous analysis was based on the assumption that normalized inputs, i.e. inputs which have the same starting point and transition time are applied to the transistors of the chain. In case non-normalized inputs are applied, an input mapping algorithm [8] can be employed in order to map every possible input pattern to a set of normalized inputs.

4. RESULTS AND DELAY CALCULATION

The calculated output waveforms of the two transistor equivalent chain, match very well the SPICE simulation results

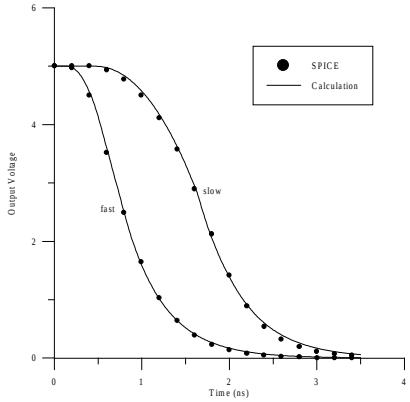


Fig. 5: Output waveform comparison between simulated and calculated values for fast and slow input ramps and for a 0.5μm HP technology

of the complete chain, as shown in Fig. 5. A comparison of the chain output response calculated according to the proposed method to that produced by the approach of [4], where the chain is replaced by a single transistor with its transconductance reduced by the number of the transistors in the chain is also included. In Table I, approximation errors in the calculation of the output waveforms for the two approaches at half-V_{DD} point when the same ramp input is applied to all transistors are presented. Moreover, a comparison for the case of tapered chains is also given. From this comparison it is obvious that the proposed two-transistor equivalent chain models the behavior of the complete chain with excellent accuracy and is much more reliable than the replacement by a single transistor : not only the average error of the proposed approach (4.1 %) is much smaller than the average error in the simple *n*-times transconductance reduction (15.5 %), but furthermore the latter method presents a higher error variance.

Since the output waveform expression for each of the 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. 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 have been obtained and compared with simulation results. It was observed that in all cases the propagation delay computed using the analytical expressions is within 4 % of that computed by SPICE when the same ramp input was applied to all transistors.

5. CONCLUSION

A detailed analysis for the operation of the transistor chain in CMOS gates was introduced. All nonsaturated devices in the chain are replaced by an equivalent transistor whose width is efficiently calculated taking into account the operating conditions of the structure. The exact time when the transistor chain starts conducting is obtained and analytical expressions for the output response to non-zero transition time inputs are extracted using short channel transistor current models which take into account second order effects of submicron devices. The calculated output waveform and delay results present very small errors compared to SPICE simulation values.

6. REFERENCES

[1] L. Bisdounis, S. Nikolaidis and 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, pp. 302-306, February 1998.

[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] N. Hedenstierna and K. O. Jeppson, "CMOS Circuit Speed and Buffer Optimization", *IEEE Trans. Computer-Aided Design*, vol. CAD-6, no. 2, March 1987.

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

[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, pp. 289-306, 1990.

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

[7] J.-T. Kong and D. Overhauser, "Methods to improve digital MOS macromodel accuracy", *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 14, pp. 868-881, July 1995.

[8] A. Chatzigeorgiou and S. Nikolaidis, "Collapsing the Transistor Chain to an Effective Single Equivalent Transistor", *Proc. Design Automation and Test in Europe Conference* (*DATE*), Paris, France, February 1998.

[9] J. M. Rabaey, "*Digital Integrated Circuits : A Design Perspective*", Upper Saddle River, NJ : Prentice Hall, 1996.

Table I: Approximation error (%) in calculation of a 4-transistor chain output response for the two-transistor and single-transistor equivalent approaches, at $V_{DD}/2$. *L* and *W* are given in μm.

		τ =0.5ns		τ =1ns		τ =2ns	
L	*W*	Prop.	Conv.	Prop.	Conv.	Prop.	Conv.
0.5	4.5	4.751	7.852	5.769	5.897	7.168	1.477
	9	4.200	18.202	5.794	16.887	7.655	21.204
1	12	0.979	20.533	5.534	21.637	6.502	19.316
	18	1.771	42.511	3.059	39.580	4.446	36.564
0.5, a=0.7	W_b=9	1.996	6.347	1.169	4.344	3.089	0.938
1, a=0.7	W_b =18	2.072	3.780	4.032	6.652	5.000	5.980