Modeling the Transistor Chain Operation in CMOS Gates for Short Channel Devices

Spiridon Nikolaidis, *Member, IEEE,* and Alexander Chatzigeorgiou, *Student Member, IEEE*

***Abstract*—A detailed analysis of the transistor chain operation in CMOS gates is introduced. 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 width of the equivalent transistor that replaces all nonsaturated devices is efficiently calculated, eliminating previous inconsistencies in chain currents. A mapping algorithm for all possible input patterns to a scheme that can be handled analytically is also derived. The final results for the calculated response and the propagation delay of this structure are in excellent agreement with SPICE simulations.**

***Index Terms*—Modeling, simulation, timing analysis, transistor chain.**

I. Introduction

EFFICIENT design of digital integrated circuits requires the advance estimation of gate delays. Circuit simulators, such as SPICE, that can provide a detailed and accurate analysis are based on numerical methods and, therefore, are prohibitively slow for large designs. The alternative is to use analytical expressions which take into account the most critical factors that influence the system behavior and are orders of magnitude faster than SPICE. Much research effort has been devoted during recent years to the modeling of the CMOS inverter behavior [1]–[3], but little has been done on more complicated gates because of their multinodal circuitry and multiple inputs. In this work, series connected MOSFET's, which form a basic structure in digital circuits since they are used in the implementation of NAND/NOR gates, are examined. Their operation is substantially more complicated than that of parallel transistors and is complicated by the fact that differential equations that govern the behavior of the circuit must be solved for several nodes and input patterns.

A qualitative description of the behavior of serially connected transistors in domino CMOS gates was given by Shoji [4]. An attempt to study the MOSFET chain was made, considering a long RC chain and without taking into account any second-order effects. Pretorius *et al.* [5] simplified the nonsaturated transistors of the chain by an equivalent resistor which fails to reproduce their characteristics, thus limiting the accuracy. Moreover, gate delay is calculated by assuming step inputs. In [6] pull-down delays of nFET chains are also determined using an RC tree model as a modeling technique, based on the Elmore delay formula. Kang and Chen [7] used linear approximations for the output voltage waveform of the transistor chain, attempting to model the propagation delay in domino gates, and only step inputs and long channel devices were considered. Additionally, the n-times transconductance reduction for the equivalent transistor, which later is replaced by a resistor, results in inconsistency of the currents, as will be shown in this paper. Applying the nth power law for submicron devices, Sakurai and Newton [8] developed expressions for a CMOS inverter. Extension to gates was made either by fitting models to all possible compound I–V curves of the transistor chain in order to extract the corresponding effective parameters, or by proposing a delay degradation factor which states that the ratio of the delay of a transistor chain to the delay of a single MOSFET can be calculated as the ratio of the corresponding drain currents for $V_{\rm GS} = V_{\rm DS} = V_{\rm DD}$. Cherkauer and Friedman [9] performed their analysis, using a simplified long-channel model and applying step inputs in order to optimize channel widths for low power consumption. Effective resistance for each of the nonsaturated devices is calculated, assuming negligible body effect and a uniform distribution of the voltage across a voltage divider, which results in inconsistent currents. Nabavi-Lishi and Rumin [10] presented a semi-empirical method for collapsing the complete transistor chain to a single equivalent transistor. The equivalent transistor width approximation is based on a simple n-times transconductance reduction, resulting in limited accuracy. In the same manner, Daga *et al.* [11] developed their analysis for an inverter macromodel and gates were treated by defining an equivalent drivability factor, using simplified assumptions for the operation of the transistors in the chain.

Some of the secondary effects which are present in the operation of the transistor chain have been mentioned in [12], where a chain collapsing technique based on a nonlinear macromodel is proposed. However, parameters are extracted from dc analyses and applied on transient phenomena. Moreover, a simplified theoretical analysis is used for the validation of the proposed effective transconductance model.

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 transistor chain is reduced to an equivalent circuit consisting of two serially

Manuscript received October 21, 1997; revised October 28, 1998. This paper was recommended by Associate Editor M. Glessner.

S. Nikolaidis is with the Department of Physics, Aristotle University of Thessaloniki, Thessaloniki 54006, Greece.

A. Chatzigeorgiou is with the Computer Science Department, Aristotle University of Thessaloniki, Thessaloniki 54006, Greece.

Publisher Item Identifier S 1057-7122(99)08104-0.

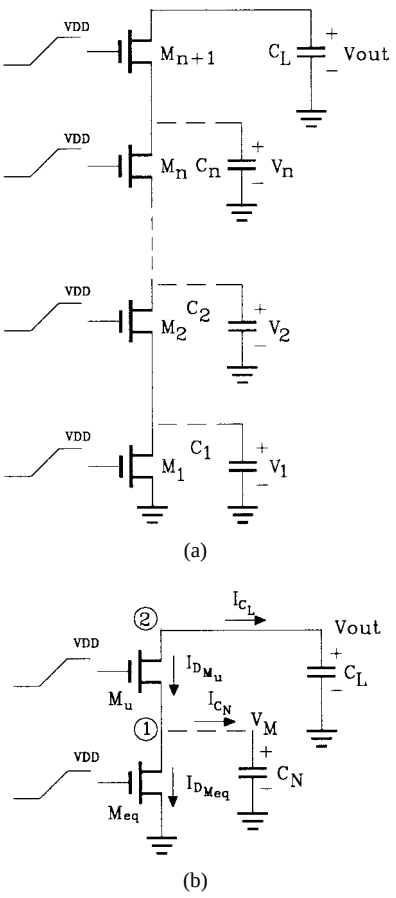

Fig. 1. (a) Complete transistor chain and (b) two-transistor equivalent chain.

connected transistors, where the one closer to the output remains unchanged and the other is the equivalent of the rest of the transistors in the chain. 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 replacing them by resistors and for inputs with nonzero transition time. In addition, all possible input patterns that can be applied to the gates of the transistors in the complete chain are mapped onto the two gate inputs of the equivalent circuit.

The two-transistor equivalent circuit, which gives the opportunity to treat the transistor chain equations analytically, is presented in Section II. The mathematical expressions which describe the output waveform evolution are derived in Section III, while in Section IV, the calculated results are compared with SPICE simulations. The input mapping algorithm for transforming all possible input patterns to normalized input ramps, which are handled in the mathematical analysis, is described in Section V. Section VI is dedicated to conclusions.

II. Transistor Chain Model

Our analysis is performed for a chain of serially connected NMOS transistors, as shown in Fig. 1(a), where the capacitances attached to the intermediate nodes correspond to the parasitic capacitances formed by the diffusion region of the transistors. The temporal evolution of the output voltage across a load capacitance that discharges through the chain, is examined. The case of a PMOS transistor chain is symmetrical. Instead of the simplified step-input pattern used in all previous works, a common ramp input, applied to the gates of all transistors, is considered which corresponds to the worst case (slower) for the output response. The a-power model proposed in [2], which takes into account the carrier velocity saturation effect of short channel devices, is used for the transistor currents of the chain

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

where $V_{D\text{-SAT}}$ is the drain saturation voltage, k_l, k_s are the transconductance parameters which depend on the width to length ratio of a transistor, α is the velocity saturation index, and $V_{\rm TN}$ is the threshold voltage expressed by

$$V_{\rm TN} = V_{TO} + \gamma\left(\sqrt{2\varphi_F + |V_{\rm SB}|} - \sqrt{2\varphi_F}\right) \tag{2}$$

where V_{TO} is the zero bias threshold voltage, γ is the body-effect coefficient, φ_F is the bulk potential, and $V_{\rm SB}$ is the source to substrate voltage. In order to transform the expression for the threshold voltage into a simplified one that can be treated mathematically, a first-order Taylor series approximation around $V_{\rm SB} = 0.2V_{\rm DD}$ is satisfactory

$$\begin{aligned} \tilde{V}_{\rm TN} &= V_{\rm TN}|_{V_{\rm SB}=0.2V_{\rm DD}} + (V_{\rm TN})'|_{V_{\rm SB}=0.2V_{\rm DD}} \\ &\quad \cdot (V_{\rm SB} - 0.2V_{\rm DD}) \\ &= \theta + \delta V_{\rm SB}. \end{aligned} \tag{3}$$

The topmost transistor in the chain (M_{n+1}) begins its operation in saturation mode, since its drain to source voltage $V_{\rm DS}$ is initially $V_{\rm DD}$. As the load capacitance (C_L) discharges and the internal node capacitance C_n charges, transistor M_{n+1} enters the linear region when $V_{\rm DS} = V_{D\text{-SAT}N}$. The rest of the transistors operate in the linear region without ever leaving this region. That is because after the chain starts conducting, their $V_{\rm DS}$ never exceeds the drain saturation voltage [9]. Since the current of the transistors that operate in linear mode increases as the voltage at the intermediate nodes rises, there will be a time point where the current of the bottom transistors will be equal to the current of the saturated top transistor. From this time on, the structure remains at this state until the charge across the load capacitance is no longer adequate to keep the topmost transistor in saturation. During this time interval, the voltage at the source of all transistors remains constant. This is the state which Kang and Chen [7] refer to as the plateau voltage and is apparent for fast input transitions, since intermediate nodes remain at this potential for a reasonable time [Fig. 5(a)].

Since the number of differential equations that must be solved in order to obtain an analytical expression for the output waveform of a transistor chain is prohibitive, the number of transistors must be reduced. A good approximation is to replace all transistors that operate in linear mode by an equivalent one and to solve the problem for the case of two

transistors [Fig. 1(b)], where the upper operates initially in saturation and then in linear mode and the bottom only in linear mode.

In order to calculate the plateau voltage of the chain, let us consider the circuit of Fig. 1(a) 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 $W_{\rm eq}$ of the equivalent transistor $M_{\rm eq}$ in Fig. 1(b), which replaces all nonsaturated transistors (their number is denoted as n) and is given by

$$\frac{1}{W'_{\rm eq}} = \frac{1}{W_1} + \frac{1}{W_2} + \cdots + \frac{1}{W_n}. \tag{4}$$

The plateau voltage at the source of the top transistor, V_p, occurs at the end of the input ramp ($V_{\rm in} = V_{\rm DD}$) as it is explained in the next section. 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 ($M_{\rm eq}$), which operates in linear mode

$$k_s(V_{\rm DD} - \theta - (1+\delta)V_p)^a = k_{l_{\rm eq}}(V_{\rm DD} - V_{TO})^{a/2}V_p. \tag{5}$$

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, where transistors are replaced by resistors, 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 of the 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, the transistors would not be able to drive the same current if they had equal drain-to-source voltages. This is the primary source of errors in existing modeling techniques [13]. For example, equating the currents through the two closer to ground transistors (for the same transistor width) for $V_{\rm in} = V_{\rm DD}$ and setting the same $V_{\rm DS}$ for each transistor gives

$$\begin{aligned} I_1 = I_2 &\Rightarrow k_l(V_{\rm DD} - \theta)^{a/2}V_{\rm DS} \\ &= k_l(V_{\rm DD} - \theta - (1+\delta)V_1)^{a/2}V_{\rm DS} \end{aligned} \tag{6}$$

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 $\delta > 0$. Trying to keep the current of each transistor in the chain constant, the reduction in $V_{\rm GS}$ and the increase in $V_{\rm TN}$ of a transistor closer to the output is compensated by an increase in its $V_{\rm DS}$. Considering a gradual increment of $V_{\rm DS}$ by a constant factor $v(v > 1)$, called the *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. 1(a)]. In this way, (6) can be rewritten as

$$\begin{aligned} &k_l(V_{\rm DD} - \theta)^{a/2}V_{DS_1} \\ &\quad = k_l(V_{\rm DD} - \theta - (1+\delta)V_{\rm DS_1})^{a/2}v \cdot V_{\rm DS_1}. \end{aligned} \tag{7}$$

In order to solve the above equation, a first-order approximation of the V_{DS_1} term inside the parenthesis on the right-hand side of (7) 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 (when all transistors have the same width) and (7) can be solved for v resulting in

$$v = \left[\frac{V_{\rm DD} - \theta}{V_{\rm DD} - \theta - (1+\delta)\dfrac{V_p}{n}}\right]^{a/2}. \tag{8}$$

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

$$W_{\rm eq} = \frac{W_1}{1 + v + \cdots + v^{n-1}} \tag{9}$$

which is used in the mathematical analysis. In case of a tapered transistor chain, the v factor and the width of the equivalent transistor can be easily extracted following the above procedure.

It should be mentioned that the drain-to-source voltage modulation factor v is not constant. This has been observed from SPICE simulations and can be explained as follows. The factor v increases for nodes closer to the output since a further increase in the source voltage of a transistor requires a further increase in its drain-to-source voltage, in order to keep the current through the transistor constant. For operating regions away from the plateau state, the factor v reduces. After the plateau state, the closer to the ground transistors have to conduct larger currents, due to the discharging of the internal capacitances and the current sourced by the coupling capacitances between input and each internal node, resulting in an increase in their $V_{\rm DS}$. Thus, the value of v is reduced, compared to the plateau state. The opposite should happen during the charging of the internal nodes before the plateau state. However, charges are injected to each internal node which, if the effect of the coupling capacitances is intense, not only compensate for the charging currents of the parasitic node capacitances, but also contribute to the currents flowing through the lower transistors in the chain. Again, since each transistor below a node must also conduct these extra currents, its $V_{\rm DS}$ is increased, resulting in a reduction in the value of v. The estimation of v, using the equations which describe the current through the two bottom transistors at the plateau state, gives an average value which is sufficiently valid for the complete region of operation of the chain.

The SPICE circuit model used for simulating the two serially connected transistors so that the bottom transistor always operates in linear mode, independent of the intermediate node voltage V_M, is shown in Fig. 2. Since in the current expression for the linear region of a transistor, it is equivalent to reduce the k_l term or the $V_{\rm DS}$ term, the width of the bottom transistor is kept unchanged W_1 and its $V_{\rm DS}$ is reduced, respectively, by setting the controlled voltage source, shown in Fig. 2. Since the transistor chain starts conducting later than the two-transistor equivalent circuit, for proper simulation, the input to

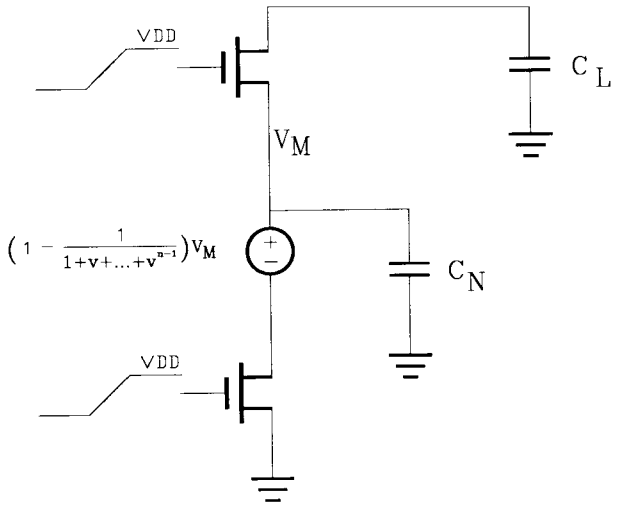

Fig. 2. SPICE equivalent circuit model of the complete transistor chain.

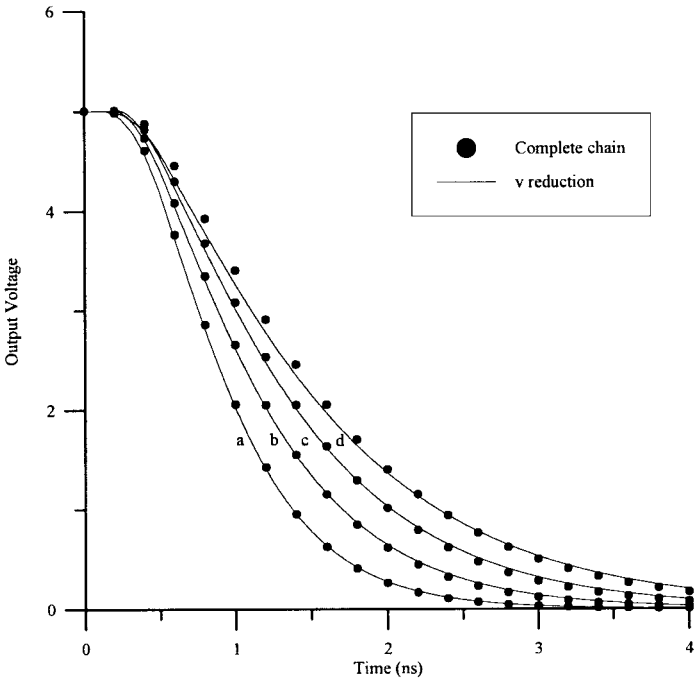

Fig. 3. Output waveform comparison between complete chain and two-transistor chain model, for $a = 3$, $b = 4$, $c = 5$, $d = 6$ transistors in the chain.

be applied to both transistors of the equivalent chain should remain at 0 V until the transistor chain starts conducting at time t_1 and then abruptly rise up and coincide with the input applied to the nonsaturated devices of the chain. In addition, the voltage at the source node of the top transistor in the chain (M_{n+1}) at time t_1 should be set as an initial condition to the intermediate node of the circuit in Fig. 2, and simulation should be performed thereafter.

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. In addition, a comparison with the output response when the equivalent transistor width is calculated in the way described

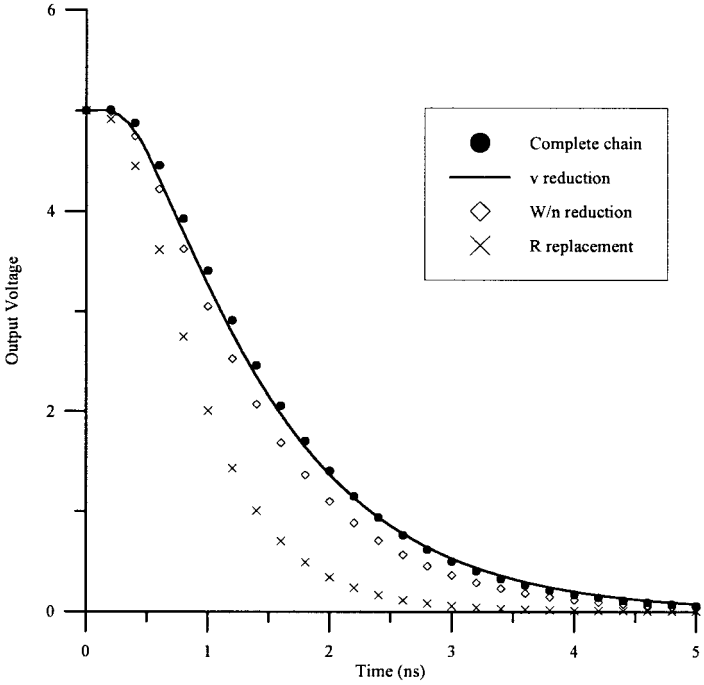

Fig. 4. Comparison between the output waveform of the complete chain and the two-transistor chain model where the nonsaturated devices are replaced using the v factor, the n-times transconductance reduction and a resistor, for a six-transistor chain.

by (4) and when the nonsaturated devices are replaced by a resistor [7], are 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.

III. Output Waveform Analysis

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

$$V_{\text{in}} = \begin{cases} 0, & t \le 0 \\ \dfrac{V_{\text{DD}}}{\tau} \cdot t, & 0 < t \le \tau \\ V_{\text{DD}}, & t > \tau \end{cases} \tag{10}$$

where τ is the input rise time. In the following analysis, all internal nodes of the chain are considered to be discharged at time $t = 0$. The effect of initially charged nodes in the chain operation is discussed in Appendix A. The differential equations that describe the operation of the circuit in Fig. 1(b) are derived by applying Kirchhoff's current law at nodes two and one

$$I_{C_L} = -I_{D_{M_u}} \Rightarrow C_L \frac{dV_{\text{out}}}{dt} = -I_{D_{M_u}} \tag{11}$$

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

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} can

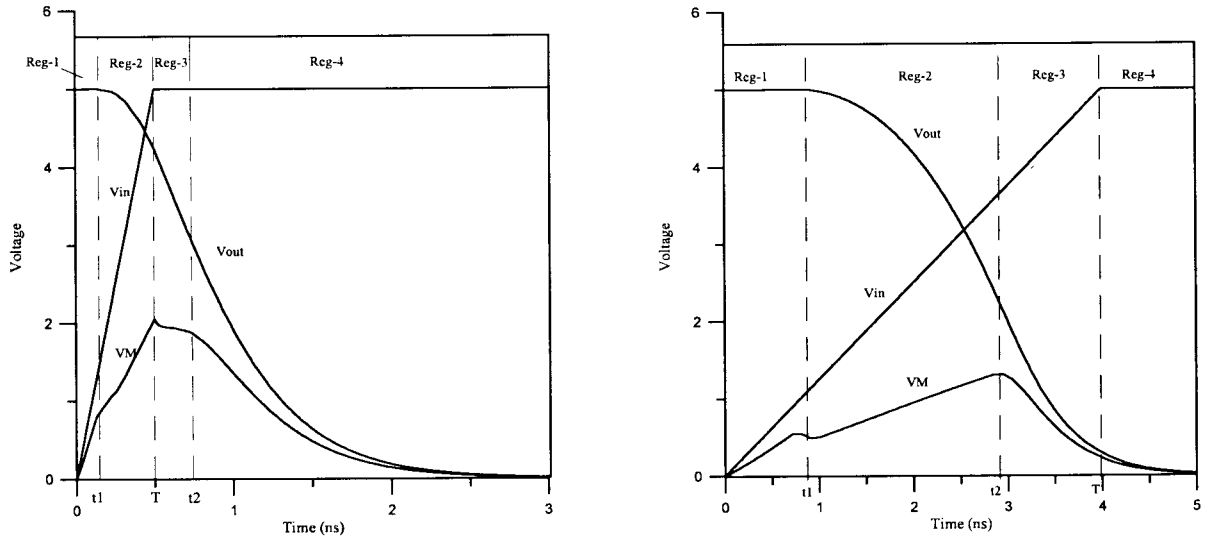

Fig. 5. Regions of operation for (a) fast and (b) slow input ramps.

be calculated as a function of the base area and sidewall periphery [14]

$$C_{\text{node}} = W \cdot d \cdot CJ + 2 \cdot d \cdot CJSW \tag{13}$$

where W is the transistor width, d is the spacing between two adjacent polysilicon lines that form the gates, and CJ, $CJSW$ are the SPICE parameters for the base and sidewall capacitance, respectively.

The lumped capacitance C_N in the equivalent circuit in Fig. 1(b) is calculated so that its charge will be equivalent to the overall charge, which is stored in all intermediate nodes of the complete chain at any time and is given by the following equation:

$$C_N = \sum_{i=1}^{n} C_i \frac{\sum_{j=0}^{i-1} v^j}{1 + v + \cdots + v^{n-1}} \tag{14}$$

where i corresponds to the internal nodes of the circuit in Fig. 1(a) (numbering starts from the drain of the bottom transistor), C_i is the diffusion capacitance at each node of the chain and the term $(\sum_{j=0}^{i-1} v^j / 1+v+\cdots+v^{n-1})$ corresponds to the ratio of the voltage at each node to that at the source of the top transistor. However, it should be mentioned that the influence of this lumped capacitance for short channel devices is not significant.

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}.

A. Fast Input Ramps

Region One: 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. Time t_1 is calculated by solving the equation $V_{GS_u}[t_1] = V_{TN_u}[t_1] \Rightarrow V_{\text{in}}[t_1] - V_M[t_1] = \theta + \delta \cdot V_M[t_1]$ where V_M is the voltage at the source of the top transistor and is considered to be linear, as is explained in the next region. The output voltage remains at V_{DD} [Fig. 5(a)]. 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 cutoff or in saturation [15].

A more precise estimation of t_1, which takes into account the effect of coupling capacitance between the transistor gates and the intermediate nodes of the chain, is given in Appendix B.

Region Two: 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, which is a valid assumption as confirmed by SPICE simulations.

The plateau voltage V_p will occur when the input reaches its final value, where the current of the top transistor ceases to increase. For very fast input ramps and because of system inertia, the plateau state occurs after this time point. However, in this case the effect of coupling capacitance between input and internal nodes becomes significant (see Appendix B) imposing V_M to obtain its maximum value at $t = \tau$. Although a voltage overshoot appears, considering the plateau voltage to extend from this time point results in a very good approximation. Thus, V_p can be calculated from (5), which is obtained from (12) for time $t = \tau$ where $dV_M/dt = 0$. Since V_M is linear, it can be written as $V_M[t] = m \cdot t$, where $m = V_p/\tau$.

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

$$V_{\text{out}} = c_1 + (q \cdot t - \theta)^a \frac{k_s}{C_L(1+a)} \left[\frac{\theta}{q} - t \right] \tag{15}$$

where $q = (V_{\text{DD}}/\tau) - (1+\delta)m$ and $c_1 \approx V_{\text{DD}}$.

Region Three: The input ramp has reached V_{DD}, the top transistor is in saturation, and the bottom is 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 equation (11) gives

$$V_{\text{out}} = c_2 - \frac{k_s}{C_L}\left[V_{\text{DD}} - \theta - (1+\delta)V_p\right]^a \cdot t \qquad (16)$$

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

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

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

$$C_L \frac{dV_{\text{out}}}{dt} = -k_{lu}(V_{\text{DD}} - \theta - (1+\delta)V_M)^{a/2} \cdot (V_{\text{out}} - V_M) \qquad (17)$$

$$-C_L \frac{dV_{\text{out}}}{dt} = k_{lb}(V_{\text{DD}} - V_{TO})^{a/2}V_M + C_N \frac{dV_M}{dt} \qquad (18)$$

where k_{lu}, k_{lb} specify the linear region coefficients for the upper and bottom transistors, respectively. Since the above system cannot be solved analytically, V_M in (17) in the term that is powered to $\alpha/2$ is replaced by its average value $V_p/2$. Solving (17) for V_M, substituting the resulting expression in (18), and setting $g_1 = k_{lu}(V_{\text{DD}} - \theta - (1+\delta)(V_p/2))^{a/2}$ and $g_2 = k_{lb}(V_{\text{DD}} - V_{TO})^{a/2}$ results in a second-order differential equation which has the solution

$$V_{\text{out}} \cong c_3 \cdot \exp\left(\frac{-p_2 + \sqrt{p_2^2 - 4p_1p_3}}{2p_1}\, t\right) \qquad (19)$$

where $p_1 = C_N \cdot C_L/g_1$, $p_2 = (C_L \cdot g_2/g_1) + C_L + C_N$, $p_3 = g_2$ and c_3 is calculated by equating the above equation for $t = t_2$ with $V_{\text{out}}[t_2]$, which is obtained from the previous region.

B. Slow Input Ramps

Region One: As in the case of fast inputs, the output voltage remains at V_{DD} until time $t = t_1$ where the top transistor enters saturation [Fig. 5(b)].

Region Two: The top transistor is saturated and the bottom operates in linear mode. This region extents from time t_1 until the top transistor exits saturation at time $t_2 < \tau$. In order to solve the system of differential equations, V_M is again assumed to be linear. This time, the plateau voltage cannot be calculated as previously and the currents of the top and bottom transistor cannot be set equal for $V_{\text{in}} = V_{\text{DD}}$. However, it has been found that when the output load is sufficiently increased (which corresponds to the case of a fast input ramp where plateau voltage occurs), the slope of V_M does not change significantly (Fig. 6). Therefore, considering a larger load capacitance, V_p would occur at time $t = \tau$ and would be calculated, as previously, by (5) where there is no dependence on the load. In this way, it is possible to obtain the slope of V_M for the case of a slow input. The differential equation at the output node is the same as in the case of fast inputs and V_{out} can be obtained in the same way.

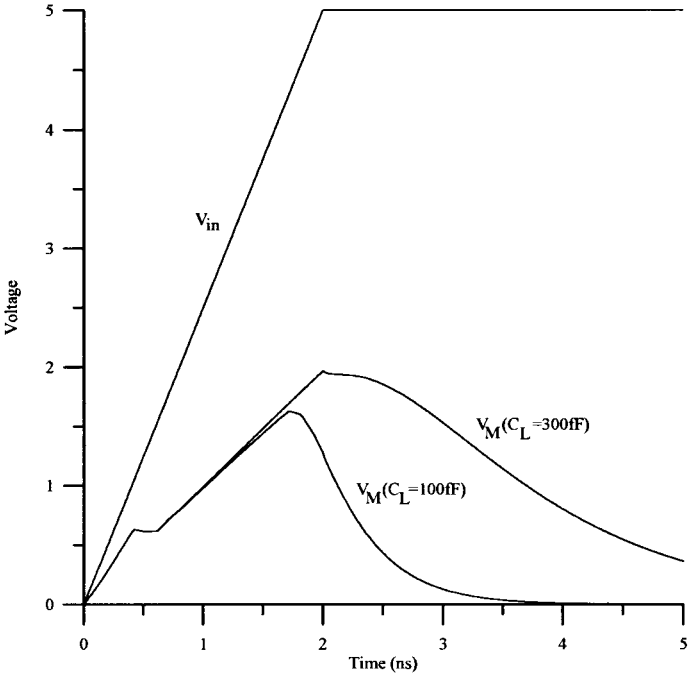

Fig. 6. Intermediate node voltage waveforms for the same input and different output loads.

The limit of this region (t_2) can be calculated by solving $V_{\text{D-SATN}}[t_2] = V_{\text{out}}[t_2] - V_M[t_2]$, using a second-order Taylor series approximation around $t = \tau$.

Region Three: Both transistors operate in linear mode and the input is still a ramp. The system of differential equations at nodes two and one, becomes

$$C_L \frac{dV_{\text{out}}}{dt} = -k_{lu}(V_{\text{in}} - \theta - (1+\delta)V_M)^{a/2} \cdot (V_{\text{out}} - V_M) \qquad (20)$$

$$-C_L \frac{dV_{\text{out}}}{dt} = k_{lb}(V_{\text{in}} - V_{TO})^{a/2}V_M + C_N \frac{dV_M}{dt}. \qquad (21)$$

In order to solve the system, the input V_{in} is replaced by its average value $\tilde{V}_{\text{in}} = (V_{\text{in}}|_{t=t_2} + V_{\text{DD}})/2$ and V_M in the term of (20) that is powered to $\alpha/2$ is replaced by its value at $t = t_2$, since the duration of this region is short. Setting $h_1 = k_{lu}[\tilde{V}_{\text{in}} - \theta - (1+\delta)V_M[t_2]]^{a/2}$ and $h_2 = k_{lb}(\tilde{V}_{\text{in}} - V_{TO})^{a/2}$ (20), (21) result in

$$V_{\text{out}} \cong c_4 \cdot \exp\left(\frac{-r_2 + \sqrt{r_2^2 - 4r_1r_3}}{2r_1}\, t\right) \qquad (22)$$

where $r_1 = C_N \cdot C_L/h_1$, $r_2 = (C_L \cdot h_2/h_1) + C_L + C_N$, $r_3 = h_2$ and c_4 is calculated by setting the above equation for $t = t_2$ equal to $V_{\text{out}}[t_2]$, which is obtained from the previous region. The limit of this region is $t = \tau$ where the input reaches V_{DD}.

Region 4: Both transistors operate in linear mode and the input is V_{DD}. The system can be solved in exactly the same way as Region Three, without approximation of the input signal. The output voltage expression is

$$V_{\text{out}} \cong c_5 \cdot \exp\left(\frac{-j_2 + \sqrt{j_2^2 - 4j_1j_3}}{2j_1}\, t\right) \qquad (23)$$

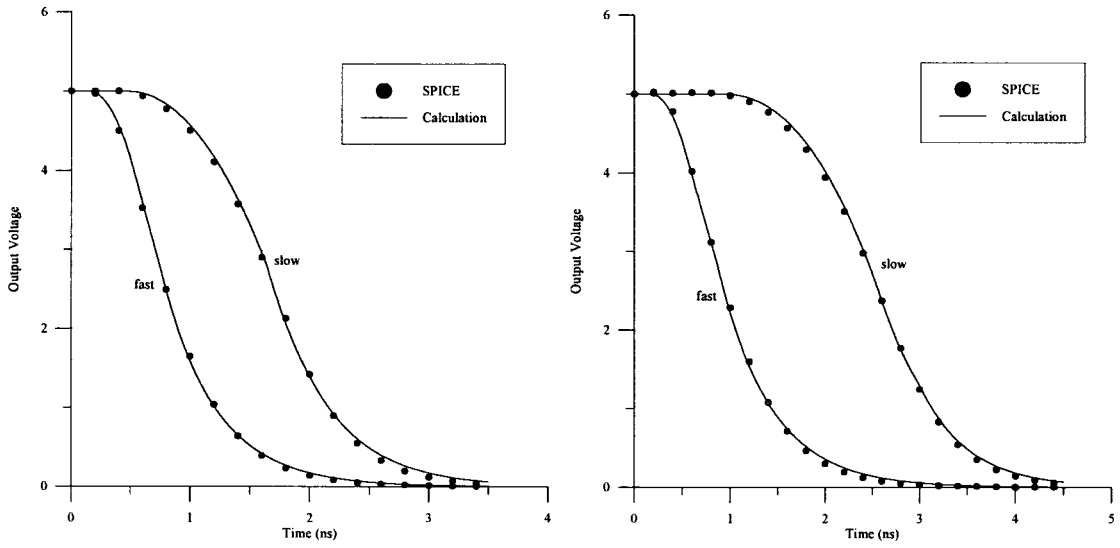

Fig. 7. Output waveform comparison between simulated and calculated values for fast and slow input ramps and for (a) 0.5- and (b) 1-μm HP technology.

where

$$j_1 = \frac{C_N \cdot C_L}{w_1}, \quad j_2 = \frac{C_L \cdot w_2}{w_1} + C_L + C_N, \quad j_3 = w_2$$

$$w_1 = k_{lu}\left(V_{\rm DD} - \theta - (1+\delta)\frac{V_M[t_2]}{2}\right)^{a/2}$$

$$w_2 = k_{lb}(V_{\rm DD} - V_{TO})^{a/2}$$

and c_5 is calculated by setting the above equation equal to $V_{\rm out}[\tau]$ for $t = \tau$.

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

The importance of the aforementioned method is that it makes it possible to reproduce the voltage evolution at each node in the circuit, thus enabling an in-depth and complete analysis of the transistor chain.

IV. Results and Delay Calculation

The calculated output waveforms of the two-transistor equivalent chain, match very well the SPICE simulation results of the complete chain, as shown in Fig. 7, which is a comparison between calculated and simulated output voltage waveforms for slow and fast input ramps, for the HP 0.5- and 1-μm 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.

A comparison of the chain output response, calculated according to the proposed method with the output response produced by the approach followed in [10] where the chain is replaced by a single transistor with its transconductance reduced by the number of the transistors in the chain (conventional method), is also given. In Table I, approximation errors in the calculation of the output waveforms at the half-$V_{\rm DD}$ point for the two approaches 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 case of a single transistor. Not only is the average error of the proposed model (4.1%) much smaller than the average error in the single-transistor model (15.5%) but, in addition, the first method presents significantly lower dispersion of error values. Another important drawback to the single-transistor chain model is that the shape of its output response deviates significantly from that of the actual chain response.

TABLE I
Approximation Error (%) in Calculation of a Four-Transistor Chain Output Response for the Two-Transistor and Single-Transistor Equivalent Approaches, at $V_{\rm DD}/2$

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

Since the output waveform expression for each of the regions of operation is known, propagation delay for the discharging case ($t_{\rm PHL}$) can be calculated as the time from the half-$V_{\rm DD}$ point of the input to the half-$V_{\rm DD}$ point of the output. The region in which $V_{\rm DD}/2$ of the output occurs can be found by comparing it with $V_{\rm out}[t_2]$ and $V_{\rm out}[\tau]$. Using this definition, delay results for several input waveforms and transistor chains have been calculated and compared with simulation results. It was observed that in practical cases, the

propagation delay computed using the analytical expressions is within 3.5% of that computed by SPICE when the same ramp input was applied to all transistors.

V. Input Mapping Algorithm

In the previous sections two transistors in series were used for analyzing the operation of a transistor chain and the same input ramp was considered to be applied to both transistor gates. Thus, in order to obtain the expressions for the output response in the general case, all possible input patterns that can be applied to the transistors of the chain must be mapped efficiently into two ramps that have the same transition time and start at the same time (normalized ramps). One method for deriving a single effective signal was given by [10] and states that for all signals that are in transition after the starting time of the latest one, the equivalent ramp starts at the initial point of the ramp that starts first and ends at the last ending point of all ramps. This scheme introduces unacceptable errors for most of the cases, especially when some signals have a much smaller transition time than others or when signals start at time points which differ significantly. In [16], for transistors connected in series, the input to be applied to a single effective transistor is chosen as the one which reaches the threshold level last and the same kind of errors as in [10] are present.

The method that is proposed in this paper for mapping transistor input ramps to two normalized ones avoids large errors in cases that deviate from the ideal ones and presents higher accuracy. The influence of each input signal depends on many factors. First of all, the starting point of the last changing input is important since the transistor chain starts conducting after this time. Consequently, the further the distance of the starting point of each input from the last starting point, the less influence this signal has on the output evolution. In addition, the influence of each input depends on the position of the transistor that it is applied to in the chain. Since the gate-to-source voltage reduces and the body effect becomes more severe further up in the chain, inputs in higher positions generally result in a slower output response, especially for submicron devices where internal node capacitances are very small. In addition, the influence of a signal depends on its slope, the relation of its slope to the slope of other signals, and its relative position in time to other signals. Obviously, in an analytical method that extracts equivalent waveforms which start at the same time, have equal slopes and can replace all inputs, it is not possible to take into account all these factors.

In Section II it was stated that the ramp input, which is applied to all transistors in the chain, is also applied to the two-transistor chain model. In this way, the problem focuses on how to map $n+1$ input signals (ramps and dc inputs) of an $(n+1)$-transistor chain to $n+1$ input ramps which start at the same time point and have the same transition time. In order to find the weight of each position in the chain, some of the inputs are set to $V_{\rm DD}$ and equal ramps (ramps which have the same starting time point and the same transition time) are applied to the rest of the transistors. For each case of input patterns, a coefficient is derived with whom 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. In this way, a look-up table is obtained by simulating each case, such as the one given in Table II for an HP 0.5-μm technology, where input numbering starts from the one closest to ground and each number declares that there is a transition in the corresponding transistor input. For the special case where only the top transistor has an input ramp applied to its gate and a dc voltage is applied to all others, no coefficient need be derived, because this case can be mapped on the same input ramp for the top transistor in the two-transistor equivalent chain and a dc voltage applied to the bottom transistor. This is a special case of the analytical solution described in Section III.

TABLE II
Weight Coefficients for a Four-Transistor Chain. The Input Numbering Starts from the One Closest to the Ground ($L = 0.5$ μm)

Changing Inputs	c_{weight}
1,2,3,4	1
2,3,4	0.93
1,3,4	0.89
1,2,4	0.85
1,2,3	0.92
3,4	0.82
2,4	0.77
1,4	0.74
2,3	0.815
1,3	0.775
1,2	0.8
4	-
3	0.67
2	0.64
1	0.6

The algorithm for mapping every input pattern to two normalized ramps consists of three steps.

Step One: In order to have an input pattern which consists only of ramp inputs and $V_{\rm DD}$ voltages, the input ramps which effectively act as dc voltages must be defined. This is obtained by examining the value of all inputs at time $t = t_m$ where the last ending input ramp reaches $V_{\rm DD}/2$. Every input ramp which at $t = t_m$ has a higher value than $(2/3)V_{\rm DD}$ should be considered as $V_{\rm DD}$. If more than one inputs end at the same time point, the input for which t_m is selected is the one that starts last.

Step Two: The m input ramps that remain from Step One should be transformed into equal ramps. The initial time of the equal ramps (t_0) is taken as $\max(t_1, t_2, \cdots, t_{n+1})$ where $t_1, t_2, \cdots, t_{n+1}$ are the starting points of **all** inputs. This is reasonable since no current flows through the chain before all transistors start conducting. The equivalent transition time for the m remaining inputs is calculated by the following formula:

$$T_{\rm eq} = \frac{\sum_{i=1}^{m}\left[1 - \frac{V_i[t_0]}{V_{\rm DD}}\right](t_{e_i} - t_0)}{m} \tag{24}$$

where $V_i[t_0]$ is the value of each of the inputs that participate in this step at the initial time and t_{e_i} is the ending time point for each of these m inputs. The above formula takes into account the time during which an input is in transition after time t_0 and also their slope. The input(s) that start at t_0 will have the major influence, since the corresponding multiplication factor $1 - (V_i[t_0]/V_{\rm DD})$ will be one.

Step Three: When Step Two is completed, the input pattern consists of inputs with equal ramps and of dc inputs. Using the coefficients which take into account the weight of each input position in the chain, this pattern can be mapped onto normalized ramps which are applied at $t = t_0$ to all transistors in the chain. Thus, the effective transition time will be

$$T_{\text{eff}} = c_{\text{weight}} \cdot T_{\text{eq}}. \tag{25}$$

The normalized ramps are finally applied to the two-transistor equivalent chain.

The above algorithm was found to be very efficient in mapping every possible input pattern of $n + 1$ inputs to a pair of input ramps which start at the same time and have the same slope. The algorithm presents accurate results, even when the transistors in a chain are tapered.

In Fig. 8, a comparison of the output response for the initial and the final input patterns obtained from the mapping algorithm is shown, validating the accuracy of the proposed algorithm. The first pattern is the actual set of inputs applied to the chain and the second is the set of the obtained normalized inputs.

VI. Conclusions

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 chain model that reduces the number of transistors in the chain to two, it has been possible to solve the differential equations that describe the structure without simplified approximations. A mapping algorithm has been developed in order to transform all possible input patterns to ramps which start at the same time and have equal transition times and which can be treated analytically. Output voltage and propagation delay results derived by the proposed analytical method, match very well SPICE simulation results.

Appendix A

In case some of the internal capacitances in the transistor chain are charged at time t_0 when the last starting input is applied, the overall propagation delay of the chain increases. However, the shape of the output waveform remains almost unchanged compared to that of a chain without charged internal nodes, which receives the same input pattern except that its transition edge is shifted [7]. This time shift t_e corresponds to the overhead time required for the internal nodes to be discharged and is estimated as follows. Let t_{tr} be the output transition time, i.e., the time needed for the charge Q_L stored in the load capacitance to be discharged through the chain. t_{tr} is calculated by connecting the 10 and 90% points of the output waveform [2]. Also, let Q_e be the charge which, when set to the output node, requires the same time to be discharged with the time needed to remove the charge stored in the internal nodes. This effective charge Q_e is obtained as

$$Q_e = \sum_{i=1}^{n} r_i \frac{i}{(n+1)} C_i V_i[t_0] \tag{A1}$$

where i denotes the internal nodes, C_i, $V_i[t_0]$ the node capacitance and initial voltage and $i/(n+1)$ is the ratio of the devices through which each node is actually discharged over the total number of transistors in the chain. r_i is a Boolean variable which takes the value zero if transistors $i+1$ to $n+1$ receive a V_{DD} input and value one in any other case. That is because if all transistors above a node are conducting, the corresponding nodes are initially charged and this case was taken into account when the relevant weight coefficients for each position in the chain were obtained.

In this way the time shift t_e is extracted by

$$t_e = \frac{Q_e}{Q_L} t_{tr}. \tag{A2}$$

The above shifting of the output waveform gives very good results for a wide range of input transition times and transistor widths (maximum output voltage error at $V_{\text{DD}}/2$ approximately 4%). A discrepancy was observed close to the time point where the chain starts conducting. This is due to the fact that a chain with initially charged nodes starts discharging the output load later and this delay is greater than the average time shift of the output waveform.

Appendix B

When the same ramp input is applied to the gates of the transistors in a chain, each transistor starts conducting at a different time, because their source and threshold voltages are different. In order to find the exact time at which the two-transistor equivalent chain has to start conducting, the complete chain will be examined. Let us consider the example of a six-transistor chain with all internal nodes initially discharged, where the same input is applied to all transistors. Fig. 9 shows the drain voltages of the five lower transistors together with the common input, in a simplified manner. Because of coupling capacitance between transistor gates and the drain/source nodes, drain voltages tend to follow the input ramp until all lower transistors start conducting. Initially the transistors are in the cutoff region and the coupling capacitance is calculated as the sum of the gate-to-source and gate-to-drain overlap capacitances of the upper and lower transistors, respectively, in each node. These overlap capacitances are given by $C_{\text{overlap}} = W[C_{\text{gdo}} + C_{\text{gso}}]$ where W is the effective width of the transistor and C_{gdo}, C_{gso} are the gate-to-drain and gate-to-source overlap capacitances per micron, which are determined by the process technology. Until the time where the transistor below a node starts conducting, the voltage waveform of that node, as it is isolated between two cutoff transistors, is derived by equating the current due to the coupling capacitance of the node $I_{C_{M_i}}$ with the charging current of the parasitic node capacitance I_{C_i} (Fig. 10)

$$\begin{aligned} I_{C_{M_i}} = I_{C_i} \Rightarrow C_{M_i} \frac{dV_{\text{in}} - dV_i}{dt} = C_i \frac{dV_i}{dt} \Rightarrow V_i[t] \\ = \frac{C_{M_i}}{C_{M_i} + C_i} V_{\text{in}}[t]. \end{aligned} \tag{B1}$$

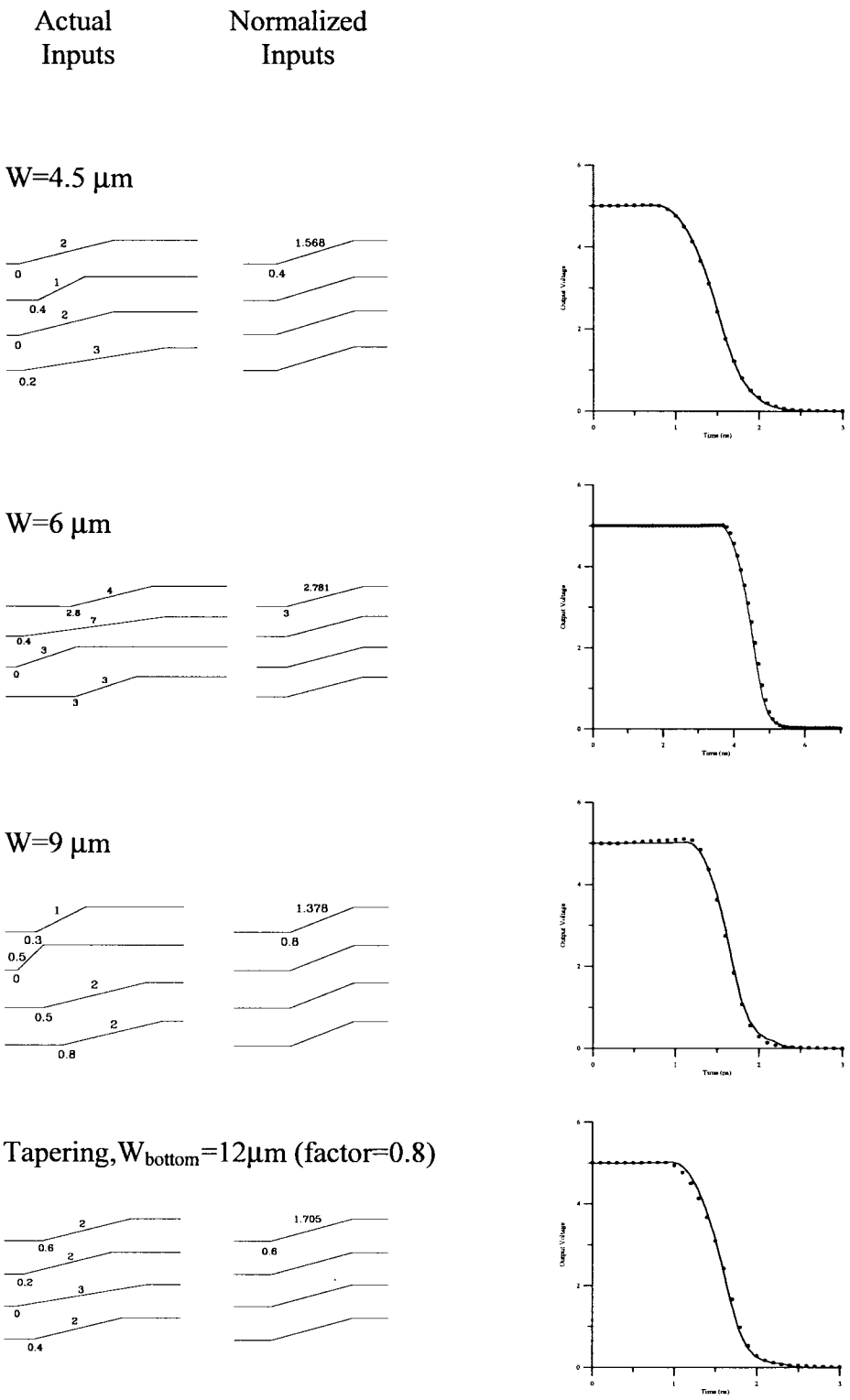

Fig. 8. Comparison between the output responses of the transistor chain ($L = 0.5$ μm) for actual inputs (dots) and for normalized ones. The starting point and the transition time of each input ramp is given in nanoseconds.

After the time at which all transistors below the ith 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. Tedious mathematical analysis is required for the extraction of the correct voltage waveform in each node. For simplicity, here the two trends are considered to be counterbalanced, which gives good results in

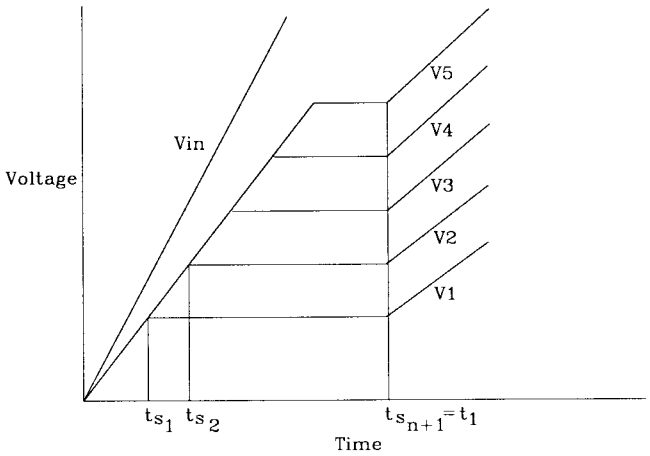

Fig. 9. Simplified representation of the intermediate node voltage waveforms.

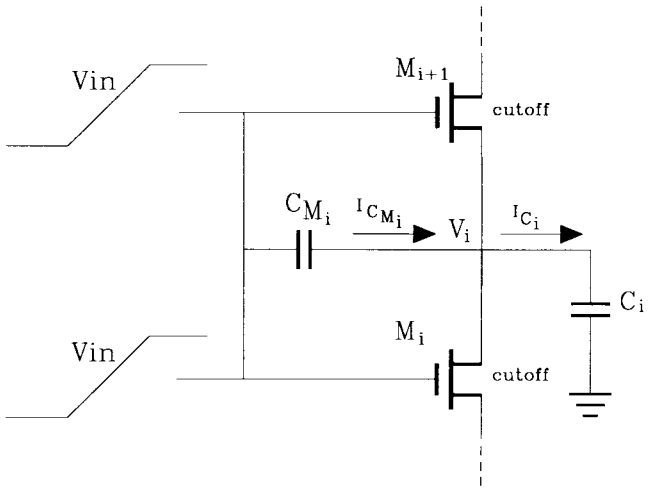

Fig. 10. Coupling and parasitic capacitances at an intermediate node.

most practical cases. This leads to the node voltage waveforms shown in Fig. 9 where the voltage of each node after the time where all lower transistors start conducting and until time t_1, is considered constant and equal to the node voltage at the beginning of this time interval.

At time $t_{s_1} = \theta \cdot \tau / V_{\rm DD}$, $(\theta \cong V_{TO})$, the bottom transistor starts conducting since its input reaches the threshold voltage. From this time on, the drain voltage of the bottom transistor remains approximately unchanged (at $V_1[t_{s_1}] = (C_{M_1}/(C_{M_1}+C_1))(V_{\rm DD}/\tau)t_{s_1}$) until the node starts charging when the complete chain has turned on at time $t = t_1$. The time at which the ith transistor in the chain starts conducting (t_{s_i}) can be calculated by solving the equation

$$V_{GS_i}[t_{s_i}] - V_{TN_i}[t_{s_i}] = 0 \Rightarrow V_{\rm in}[t_{s_i}] - V_{i-1}[t_{s_{i-1}}] - (\theta + \delta \cdot V_{i-1}[t_{s_{i-1}}]) = 0 \tag{B2}$$

which results in the recursive expression

$$t_{s_i} = \tau \frac{\theta + (1+\delta)\dfrac{C_{M_{i-1}}}{C_{M_{i-1}}+C_{i-1}}\dfrac{V_{\rm DD}}{\tau}t_{s_{i-1}}}{V_{\rm DD}} \tag{B3}$$

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.

In case time t_1 is calculated as in Appendix B, V_M will have a value V_s at time t_1 and the expression of V_M for the time interval $t_1 - \tau$ becomes $V_M[t] = V_a + m \cdot t$, where $V_a = V_s - ((V_p - V_s)/(\tau - t_1))t_1$ and $m = ((V_p - V_s)/(\tau - t_1))$.

REFERENCES

[1] N. Hedenstierna and K. O. Jeppson, "CMOS circuit speed and buffer optimization," *IEEE Trans. Computer-Aided Design*, vol. CAD-6, pp. 270–281, Mar. 1987.
[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, pp. 584–594, Apr. 1990.
[3] 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, pp. 302–306, Feb. 1998.
[4] M. Shoji, "FET scaling in domino CMOS gates," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 1067–1071, Oct. 1985.
[5] 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, pp. 523–530, Apr. 1985.
[6] S. S. Bizzan, G. A. Jullien, and W. C. Miller, "Analytical approach to sizing nFET chains," *Electron. Lett.*, vol. 28, no. 14, pp. 1334–1335, July 1992.
[7] S. M. Kang and H. Y. Chen, "A global delay model for domino CMOS circuits with application to transistor sizing," *Int. J. Circuit Theory Appl.*, vol. 18, pp. 289–306, 1990.
[8] T. Sakurai and A. R. Newton, "Delay analysis of series-connected MOSFET circuits," *IEEE J. Solid-State Circuits*, vol. 26, pp. 122–131, Feb. 1991.
[9] B. S. Cherkauer and E. G. Friedman, "Channel width tapering of serially connected MOSFET's with emphasis on power dissipation," *IEEE Trans. VLSI Syst.*, vol. 2, pp. 100–114, Mar. 1994.
[10] A. Nabavi-Lishi and N. C. Rumin, "Inverter models of CMOS gates for supply current and delay evaluation," *IEEE Trans. Computer-Aided Design*, vol. 13, pp. 1271–1279, Oct. 1994.
[11] J. M. Daga, S. Turgis, and D. Auvergne, "Design oriented standard cell delay modeling," in *Proc. Int. Workshop Power Timing Modeling, Optimization Simulation (PATMOS'96)*, 1996, pp. 265–274.
[12] J.-T. Kong, S. Z. Hussain, and D. Overhauser, "Performance estimation of complex MOS gates," *IEEE Trans. Circuits Syst. I*, vol. 44, pp. 785–795, Sept. 1997.
[13] J.-T. Kong and D. Overhauser, "Methods to improve digital MOS macromodel accuracy," *IEEE Trans. Computer-Aided Design*, vol. 14, pp. 868–881, July 1995.
[14] J. M. Rabaey, *Digital Integrated Circuits: A Design Perspective.* Upper Saddle River, NJ: Prentice-Hall, 1996.
[15] N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective*, 2nd ed. Reading MA: Addison-Wesley, 1993.
[16] 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. Computer-Aided Design*, vol. 8, pp. 1027–1032, Sept. 1989.

Spiridon Nikolaidis (S'89–M'93) was born in Kavala, Greece, in 1965. He received the Diploma and Ph.D. degrees in electrical engineering from Patras University, Patras, Greece, in 1988 and 1994, respectively.

Since September 1996 he has been with the Department of Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece, as a Lecturer in VLSI design. His research interests include CMOS gate propagation delay and power consumption modeling, high-speed and low-power CMOS circuit techniques, power estimation of DSP architectures, and design of high-speed and low-power DSP architectures.

Alexander Chatzigeorgiou (S'95) was born in Thessaloniki, Greece, in 1973. He received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Greece, in 1996, where he is currently pursuing the Ph.D. degree at the Computer Science Department, working on timing and power modeling of digital integrated circuits.

He has held internship positions at Purdue University, Lafayette, IN, the European Laboratory for Particle Physics (CERN), Geneva, Switzerland, and Imperial College, London, U.K. Since 1996 he has been with Intracom S.A., Greece, as a telecommunications software designer. His research interests include low-power VLSI design, computer architecture, and reconfigurable logic.