

ANALYTICAL ESTIMATION OF PROPAGATION DELAY AND SHORT-CIRCUIT POWER DISSIPATION IN CMOS GATES

S. NIKOLAIDIS¹ AND A. CHATZIGEORGIOU^{2,*}

¹*Department of Physics, Aristotle University of Thessaloniki, 54006 Thessaloniki, Greece*

²*Computer Science Department, Aristotle University of Thessaloniki, 54006 Thessaloniki, Greece*

SUMMARY

An efficient analytical method for calculating the propagation delay and the short-circuit power dissipation of CMOS gates is introduced in this paper. Key factors that determine the operation of a gate, such as the different modes of operation of serially connected transistors, the starting point of conduction, the parasitic behaviour of the short-circuiting block of a gate and the behaviour of parallel transistor structures are analysed and properly modelled. The analysis is performed taking into account second-order effects of short-channel devices and for non-zero transition time inputs. Analytical expressions for the output waveform, the propagation delay and the short-circuit power dissipation are obtained by solving the differential equations that govern the operation of the gate. The calculated results are in excellent agreement with SPICE simulations. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: CMOS gates; propagation delay; short-circuit power dissipation

1. INTRODUCTION

Efficient design of digital integrated circuits requires tools that can perform accurate and fast timing and power simulations. Generally, accurate simulations are obtained by means of simulators such as SPICE which are based on numerical methods for the solution of the differential equations that describe the operation of a circuit. However, numerical methods are very slow and their use for multi-million transistor designs is practically impossible. For this reason, research is being conducted on the modelling of integrated circuits in order to extract analytical expressions for the propagation delay and power dissipation of CMOS gates, which are much faster than numerical methods and close to SPICE accuracy.

During the last decade much effort has been devoted to the investigation of the CMOS inverter and analytical expressions for its performance have been obtained.^{1–5} However, research results on more complicated gates such as NAND/NOR gates are rather poor because of the intrinsic difficulties in analysing complex structures such as the transistor chain. All previously reported efforts can be divided in two main groups according to the approach that has been followed.

In the first group are all attempts that are trying to employ the analytical expressions for the inverter and are based on the hypothesis that every CMOS gate can be collapsed to an equivalent inverter which has the same performance. The accuracy of this approach is limited, mainly due to the fact that collapsing serially connected transistors to an equivalent transistor fails to model accurately the behaviour of the chain. Nabavi-Lishi and Rumin⁶ introduced a method for replacing NAND/NOR gates by an equivalent inverter using a collapsing technique which aims to the extraction of an equivalent transistor of serially connected transistors but ends up in the conventional width reduction (i.e. the equivalent transistor width is equal to the width of a transistor in the chain reduced by the number of the transistors, $W_{eq} = W/n$) and presents large errors. A more efficient method for collapsing the transistor chain to an effective single equivalent transistor

* Correspondence to: A. Chatzigeorgiou, Electronics and Computer Division, Department of Physics, Aristotle University of Thessaloniki, 54006 Thessaloniki, Greece

has been presented recently in Reference 7. Applying the n th power law for submicron devices, Sakurai and Newton⁸ developed expressions for a CMOS inverter and 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. In the same way, Daga *et al.*⁹ developed their analysis for an inverter macro-model and gates were treated by defining an equivalent drivability factor using simplified assumptions for the operation of the transistors in the chain.

The second approach corresponds to a fully mathematical analysis of the operation of the gate structure based on the fact that within a transistor chain all transistors except for the one which is attached to the output node, operate always in the linear region. Based on this observation, in References 10 and 11 the differential equations at the nodes of a transistor chain have been solved by replacing the non-saturated devices by equivalent resistors. However, resistors fail to reproduce the dynamic behaviour of the non-saturated transistors and furthermore these methods were developed for quadratic long-channel current models and for ideal step inputs which result in significant deviations from the real performance where the applied inputs have non-zero transition time.

Shih and Kang in Reference 12 presented a fully mathematical solution of general MOS circuit primitives and in Reference 13 a tool named ILLIADS has been developed on this solution. However, this method is based on quadratic form current models such as the Shichman–Hodges model and therefore depends on their deficiencies. Moreover, serially connected transistors are collapsed to an equivalent transistor using a conventional simple width reduction, resulting in limited accuracy. An improved method has been presented in Reference 14 where current expressions for short-channel devices are transformed to the required quadratic form expressions. However, their analysis requires the solution of non-linear algebraic equations in order to obtain the time point of region crossings, thereby making the method inefficient.

An efficient and accurate method for modelling CMOS NAND/NOR gates is introduced in this paper. According to the proposed analysis parallel transistor structures are replaced by a single equivalent transistor with its width equal to the width of one of the parallel transistors multiplied by their number (for equal transistor widths). Then, two cases are considered according to the role of the transistor chain in the gate operation. When the chain drives the conducting current which charges or discharges the output load, i.e. during discharging of the output load of a NAND gate, the operation of the gate is accurately captured by solving the corresponding differential equations taking into account the mode of operation of each transistor in the chain and without the simplifying approximations of previous approaches. When the transistor chain acts parasitically, i.e. during charging of the output load of a NAND gate, the transistor chain is modelled by a single equivalent transistor and an equivalent coupling capacitance. In this way the gate diminishes to an equivalent inverter and the well-known analytical expressions for the inverter can be employed. The proposed method is developed for non-zero transition time inputs and for short-channel devices. The key factors that determine the output response and the dissipated power such as the plateau voltage, the starting point of conduction and the form of the voltage waveforms at the internal nodes of a gate are considered and properly modelled. More complex gates can be treated by collapsing them to the equivalent NAND/NOR gate according to the method developed in Reference 15 while the problem of misaligned inputs can be handled using the input mapping algorithm presented in Reference 16. It should be mentioned that the proposed method has been developed for purely capacitive loads. This is reasonable since in the case of large interconnect loads the driving gate is usually a buffer and not a NAND/NOR gate. The field of CMOS gates driving RC interconnect loads needs further investigation and some previous works can be found in References 17–19.

The gate operation when the transistor chain drives the conducting current is described in Section 2 while in Section 3 the time point when the gate starts conducting is calculated. The output voltage expression is obtained by solving the corresponding differential equations in Section 4. The parasitic behaviour of the transistor chain is modelled in the next section and in Section 6 the short-circuit power dissipation of a CMOS gate is estimated. Finally we conclude in Section 7.

2. GATE OPERATION DURING TRANSISTOR CHAIN CONDUCTION

Since NAND/NOR gates consist of parallel and serial combinations of transistors, each of these structures has to be treated and modelled accordingly. It has been found that parallel connected transistors can be modelled with sufficient accuracy by an equivalent transistor with its width equal to the sum of all transistor widths. That is because parallel transistors operate under the same conditions (in case they receive the same input), and the resulting equivalent transistor also operates under the same conditions and therefore its current is the sum of all transistor currents. However, serially connected transistors present a higher complexity due to the multiple nodes and the different mode of operation of the transistors. First, the conducting behaviour of an nMOS transistor chain within a NAND gate will be examined (Figure 1(a) where the pMOS transistors have been replaced by an equivalent one), when the output load of the gate is discharged through the chain. Charging through a pMOS chain is symmetrical. Let us assume that a rising input ramp with input transition time τ is applied to the gates of all transistors:

$$V_{in} = \begin{cases} 0, & t \leq 0 \\ \frac{V_{DD}}{\tau} t, & 0 < t \leq \tau \\ V_{DD}, & t > \tau \end{cases} \quad (1)$$

In order to take into account the carrier velocity saturation effect of short-channel devices, the α -power law model² is used for the transistor currents:

$$I_D = \begin{cases} 0, & V_{GS} \leq V_{TN}: \text{cutoff region} \\ k_l(V_{GS} - V_{TN})^{a/2} V_{DS}, & V_{DS} < V_{D-SAT}: \text{linear region} \\ k_s(V_{GS} - V_{TN})^a, & V_{DS} \geq V_{D-SAT}: \text{saturation region} \end{cases} \quad (2)$$

where V_{D-SAT} is the drain saturation voltage,² k_l, k_s are the transconductance parameters which depend on the width to length ratio of a transistor, α is the carrier velocity saturation index and V_{TN} is the threshold voltage which is expressed by its first-order Taylor series approximation around $V_{SB} = 0.2V_{DD}$:

$$\tilde{V}_{TN} = V_{TN}|_{V_{SB}=0.2V_{DD}} + (V_{TN})'|_{V_{SB}=0.2V_{DD}}(V_{SB} - 0.2V_{DD}) = \theta + \delta V_{SB} \quad (3)$$

where V_{SB} is the source-to-substrate voltage.

While the input is applied and assuming that the internal node capacitances (shown in Figure 1(a) are initially discharged, the topmost transistor in the chain (M_n) begins its operation in saturation mode and then enters the linear region when its $V_{DS} = V_{D-SATN}$. The rest of the transistors operate in the linear region without ever leaving this region.¹¹ For the time interval during which the topmost transistor is in saturation and the input is rising, its current and consequently the voltages at the internal nodes are increasing. When the input reaches V_{DD} and until the topmost transistor exits saturation, its current and therefore the internal node voltages remain constant. That is because a further increase of the internal node voltages would decrease the gate-to-source voltage (V_{GS}) of the topmost transistor and therefore its current, leading in a decrease of the node voltages. On the other hand a decrease of the node voltages would increase the V_{GS} of the topmost transistor and consequently its current, leading in an increase of the internal node voltages. Therefore, these voltages remain constant until the topmost transistor exits saturation. During this time interval the parasitic currents due to drain/source node capacitances and gate-to-drain/source coupling capacitances are eliminated because the voltages at the corresponding nodes remain constant. Therefore, during this state, which is known as the 'plateau' state,¹⁰ the same current flows through all transistors in the chain. According to the previous analysis, the plateau state is apparent only for fast input transitions (Figure 2). Fast and slow inputs are determined according to the position of the time point, t_2 , when the top transistor in the chain exits saturation: in case $t_2 < \tau$, the input is slow, otherwise it should be considered fast.

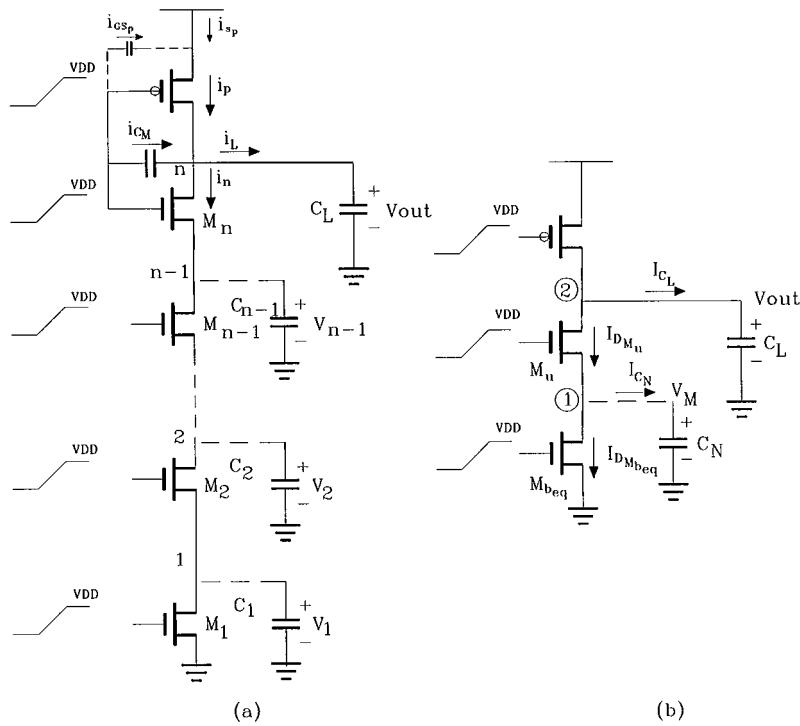


Figure 1. (a) NAND gate and (b) equivalent NAND gate with a two transistor equivalent circuit replacing the chain. The parasitic internal node capacitances are also shown

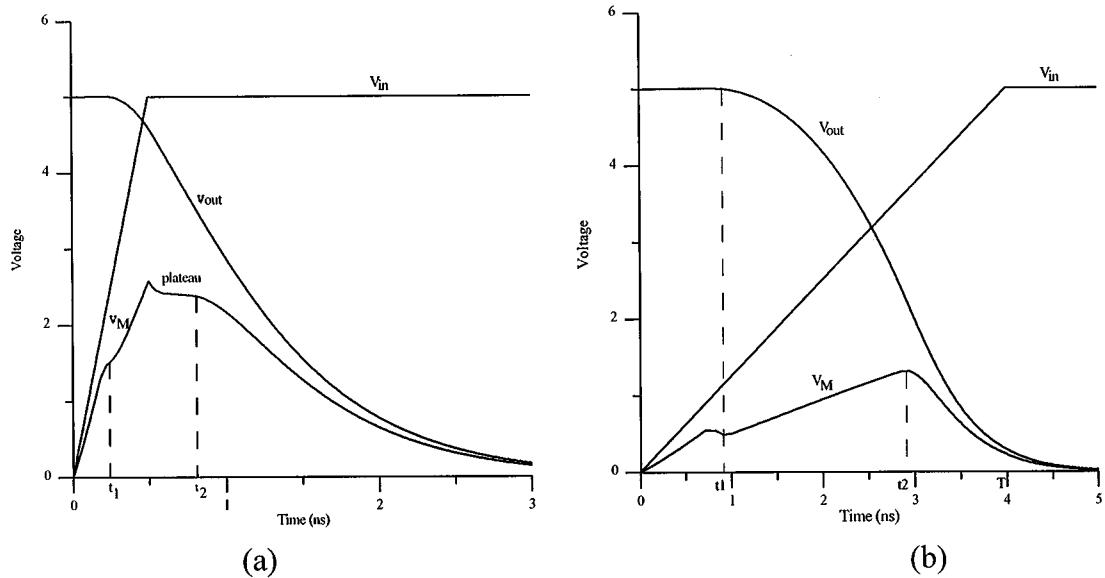


Figure 2. Output and source voltage waveform of the topmost transistor in the nMOS transistor chain of the gate in Figure 1(a), for (a) fast and (b) slow input ramp

In order to calculate the plateau voltage at the source of the topmost transistor in the chain, V_p , let us consider the circuit of Figure 1(a). 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 $W_{b_{eq}}$ of the equivalent transistor $M_{b_{eq}}$ in Figure 1(b), which replaces all non-saturated transistors of the chain and is given by

$$\frac{1}{W_{b_{eq}}} = \frac{1}{W_1} + \frac{1}{W_2} + \dots + \frac{1}{W_{n-1}} \quad (4)$$

The plateau voltage, V_p , occurs at the end of the input ramp ($V_{in} = V_{DD}$) when the discharging current ceases to increase. Thus, V_p can be calculated by setting the saturation current of transistor M_u in Figure 1(b) equal to the current of the bottom transistor ($M_{b_{eq}}$) which operates in linear mode:

$$k_{s_u}(V_{DD} - \theta - (1 + \delta)V_p)^a = k_{l_{b_{eq}}}(V_{DD} - V_{TO})^{a/2} V_p \quad (5)$$

The above equation can be solved for V_p with very good accuracy using a second-order Taylor series approximation for the left term of the equation.

In the following analysis the source voltage of the topmost transistor in the chain, V_M , is considered linear for the interval between time t_1 , where the chain starts conducting and time τ (fast inputs) or time t_2 (slow inputs) where the top transistor exits saturation. This observation is based on SPICE simulations and leads to highly accurate results (Figure 2). Time points t_1 , t_2 are calculated in Sections 3 and 4, respectively. Since time t_1 and $V_M[t_1]$ are estimated (see Section 3) and for fast inputs the plateau voltage occurs at time τ , the slope of V_M can also be estimated. For slow inputs the slope of V_M can be calculated in a similar way: Although for slow inputs the plateau voltage is not present, it has been found that if in a chain which receives a slow input the output load is increased, the slope of V_M remains almost the same. Therefore, considering a sufficiently larger load capacitance, the input would become fast, V_p would occur at time $t = \tau$ and would be calculated as previously by equation (5). Since the slope remains unchanged, the calculated slope is valid for the initial load as well. The independence of V_p on the load capacitance which is required in order for the previous proposition to be valid, is obvious from equation (5). It should be mentioned that the derivation of an expression for the voltage waveform at the source of the topmost transistor in the chain is the key point in analysing a CMOS gate.

In addition, all internal nodes of the chain are considered to be discharged at time $t = 0$. In case some of the internal nodes are initially charged, the output waveform of the gate which will result by the proposed method should be appropriately shifted,¹⁰ since the charges in the internal nodes cause an additional delay in the output response.

3. STARTING POINT OF CONDUCTION

In a transistor chain with initially discharged internal nodes and the same input applied to the gates of all transistors, the closer to the output transistors start conducting later because of a gradual increase in their source and threshold voltage. The starting point of conduction of the chain which is actually that of the topmost transistor, is estimated in this section. A first approach for the calculation of the starting point of conduction has been presented in Reference 7. This method is improved in this work and more accurate results are derived.

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. Figure 3 shows a representation of the drain voltages of the five lower transistors together with the common input. 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 cut-off 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 $C_{overlap} = C_M = W[C_{gdo} + C_{gso}]$ where

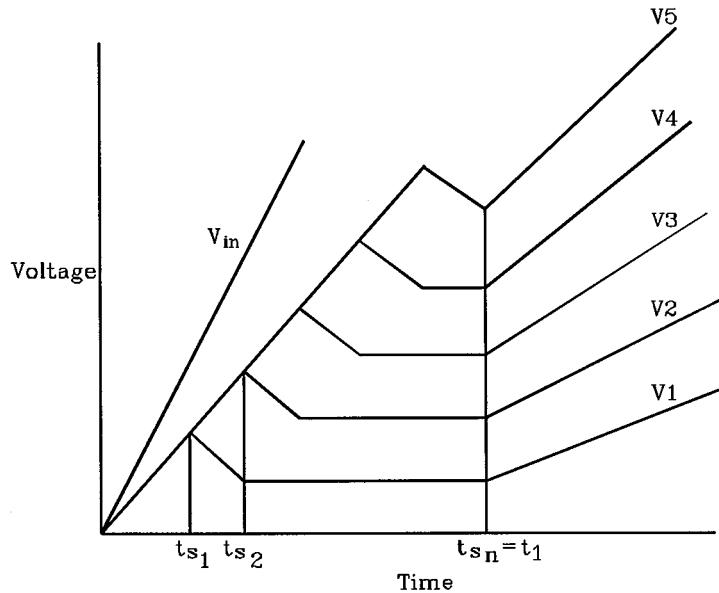


Figure 3. Intermediate node voltage waveforms until the transistor chain starts conducting

W is the transistor width 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 when the transistor below the i th node starts conducting, the voltage waveform of that node, $V_i[t]$, as it is isolated between two cut-off transistors, is derived by equating the current due to the coupling capacitance of the node, $I_{C_{\text{M}_i}}$, with the charging current of the parasitic node capacitance I_{C_i} :

$$I_{C_{\text{M}_i}} = I_{C_i} \Rightarrow C_{\text{M}_i} \frac{dV_{\text{in}} - dV_i}{dt} = C_i \frac{dV_i}{dt} \Rightarrow V_i[t] = \frac{C_{\text{M}_i}}{C_{\text{M}_i} + C_i} V_{\text{in}}[t] \quad (6)$$

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 the 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.

When a transistor starts to conduct, e.g. transistor # i , it operates initially in saturation. Therefore, since its gate-to-drain coupling capacitance is very small, the second (except for the case of very fast inputs) from the above mentioned trends dominates after time t_{s_i} and the voltage at node i decreases. This continues until time $t_{s_{i+1}}$ when transistor # $i+1$ starts conducting and enters saturation. Transistor # i , since its V_{GS} continues to increase after time t_{s_i} while its V_{DS} decreases, will enter the linear region close to $t_{s_{i+1}}$. From this point on, the gate-to-source coupling capacitance of transistor # $i+1$ increases by $\frac{2}{3}C_{\text{ox}}WL$ and the gate-to-drain coupling capacitance of transistor # i increases by $\frac{1}{2}C_{\text{ox}}WL$. Because of this increased coupling capacitance at the i th node, the two previously mentioned trends after time $t_{s_{i+1}}$ are almost counterbalanced and for simplicity the node voltage is considered constant and equal to its value at $t_{s_{i+1}}$. This observation has been verified by SPICE simulations. The node voltages start to rise again when the complete chain starts conducting at time t_1 . Additionally, the slope of the voltage waveform during $[t_{s_i}, t_{s_{i+1}}]$ is considered the same for each node and the voltage expression of node 1 during this interval can be calculated by solving the

differential equation which results from the application of Kirchhoff's current law at node 1 (Figure 4):

$$i_{n_1} = i_{C_{M_1}} - i_{C_1} \Rightarrow k_s(V_{in} - V_{TO})^a = C_{M_1} \left(\frac{dV_{in}}{dt} - \frac{dV_1}{dt} \right) - C_1 \frac{dV_1}{dt} \quad (7)$$

where the transconductance k_s is measured on the $I-V_{DS}$ characteristics for very low values of $V_{GS} (\approx V_{TO})$ and $V_{DS} (\approx (C_M/(C_M + C_n))V_{TO})$ and for simplicity the velocity saturation index α is considered one, which is a reasonable approximation for short-channel devices.

Since $t_{s_1} = V_{TO}\tau/V_{DD}$ is known and $V_1[t_{s_1}]$ is given by equation (6), the expression of $V_1[t]$ during $[t_{s_1}, t_{s_2}]$ is derived and can be used in order to calculate the time when the next transistor further up starts conducting, by solving $V_{GS_2}[t_{s_2}] - V_{TN_2}[t_{s_2}] = 0$. Having time points t_{s_1}, t_{s_2} , and the corresponding drain voltage values of the bottom transistor, $V_1[t_{s_1}], V_1[t_{s_2}]$ the slope r of each node voltage waveform during $[t_{s_1}, t_{s_{i+1}}]$ can be obtained.

According to the above analysis, the time point at which the $\#i$ transistor in the chain starts conducting can be found by solving

$$V_{GS_i}[t_{s_i}] - V_{TN_i}[t_{s_i}] = 0 \Rightarrow V_{in}[t_{s_i}] - \theta_0 - (1 + \delta_0) \left(\frac{C_{M_{i-1}}}{C_{M_{i-1}} + C_{i-1}} V_{in}[t_{s_{i-1}}] - r(t_{s_i} - t_{s_{i-1}}) \right) = 0 \quad (8)$$

which results in the following recursive expression:

$$t_{s_i} = \tau \frac{\theta_0 + (1 + \delta_0) \left(\frac{C_{M_{i-1}}}{C_{M_{i-1}} + C_{i-1}} \frac{V_{DD}}{\tau} + r \right) t_{s_{i-1}}}{V_{DD} + (1 + \delta_0) r \tau}, \quad i \geq 2 \quad (9)$$

From the above expression, the time at which the chain starts conducting $t_{s_1} = t_1$, can be easily obtained. Constants θ_0, δ_0 result from equation (3) by calculating the Taylor series approximation of the threshold voltage around $V_{SB} = V_{TO}$ for higher accuracy in this region. According to the previous analysis, the starting point of conduction of the transistor chain can be calculated with very good accuracy as shown in Figure 5

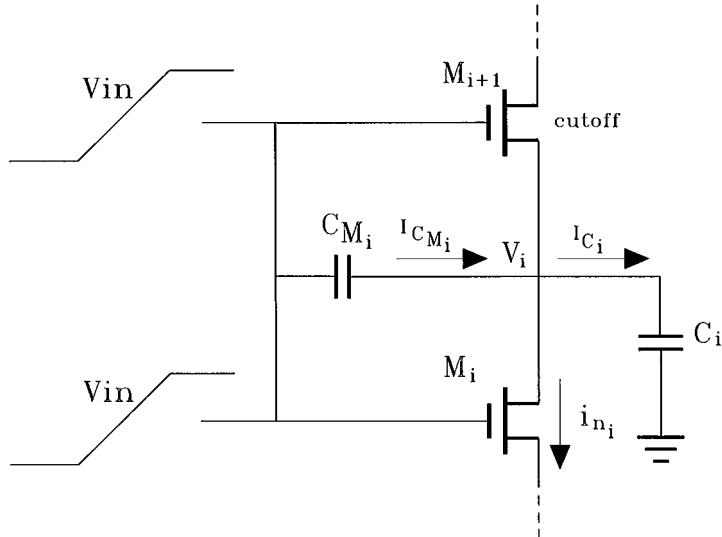


Figure 4. Currents at the i th node of the transistor chain during $[t_{s_i}, t_{s_{i+1}}]$

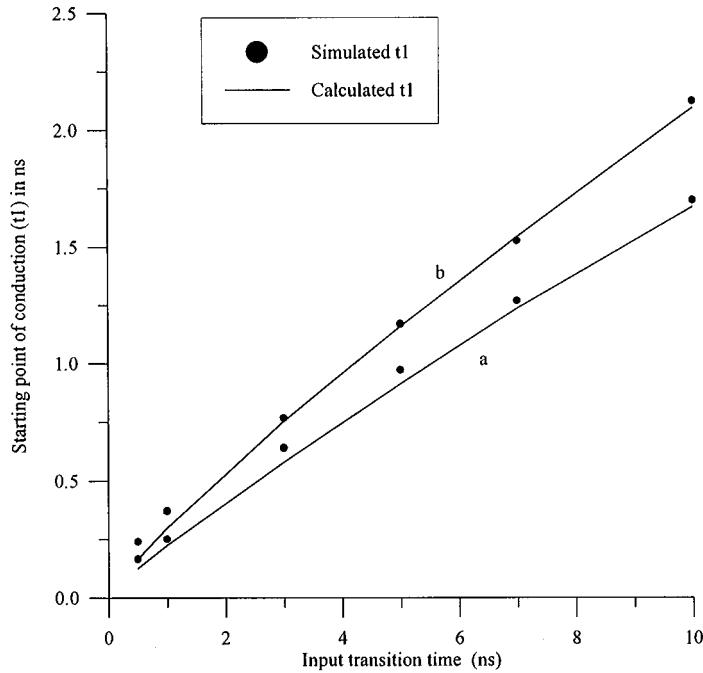


Figure 5. Comparison between the simulated and calculated starting point of conduction for several input transition times and for (a) 4-transistor chain and (b) 6-transistor chain

which is a comparison between the calculated and the actual time t_1 as it is obtained from SPICE simulations.

4. OUTPUT WAVEFORM ANALYSIS

The operation of the NAND gate during discharging of the output load is examined in this section. The output waveform of the NAND gate shown in Figure 1(a) will be extracted when the input of equation (1) is applied. The case of a NOR gate during charging of the output load is symmetrical. Although many cases can be considered according to the relative times when the topmost transistor in the chain and the pMOS transistor exit and enter saturation respectively, one case, the most common, will be presented for the sake of simplicity. The output waveform expression can be found for each of the following operating regions by solving the differential equation at the output node of the gate:

$$i_L = i_p - i_n + i_{C_M} \quad (10)$$

The coupling capacitance between input and output, C_M (Figure 1(a)), for each region is calculated taking into account the mode of operation of the pMOS and nMOS transistors attached to the output node.²⁰

Region 1. $0 \leq t < t_1$: The transistor chain is cut-off and the pMOS transistor operates in the linear region. Equation (10) can be written as

$$C_L \frac{dV_{out}}{dt} = k_{l_p} (|V_{in} - V_{DD}| - |V_{TP}|)^{a_p/2} |V_{out} - V_{DD}| + C_M \left(\frac{dV_{in}}{dt} - \frac{dV_{out}}{dt} \right) \quad (11)$$

In order to solve the above differential equation, V_{in} in the term that is powered to $a_p/2$ should be approximated by its value at $t_1/2$. The solution of the differential equation is

$$V_{out} = \frac{C_M s + k_1 V_{DD}}{k_1} + C[1] e^{-(k_1/(C_L + C_M))t} \quad (12)$$

where s is the slope of the input ($s = V_{DD}/\tau$), $k_1 = k_{l_p}(V_{DD} - st_1/2 - |V_{TP}|^{a_p/2})$ and $C[1]$ the integration constant and can be calculated easily by setting V_{out} at time $t = 0$ equal to V_{DD} .

A small voltage overshoot appears at the output node which is due to the coupling capacitance C_M , and during this overshoot current is flowing through the pMOS transistor towards V_{DD} .

The minimum value for the pMOS transistor current occurs at $t = t_1$ (Figure 6) and is given by

$$i_{p_{min}} = -k_{l_p}(|V_{in}[t_1] - V_{DD}| - |V_{TP}|^{a_p/2} |V_{out}[t_1] - V_{DD}|) \quad (13)$$

Region 2. $t_1 \leq t < t_{s-p}$: After time point t_1 the topmost transistor in the chain operates in saturation and the pMOS transistor in the linear region. According to SPICE simulations, the pMOS current can be approximated with very good accuracy by a linear function $i_p = i_{p_{min}} + c_1(t - t_1)$ as shown in Figure 6. (Reference 5). Consequently, (10) becomes

$$C_L \frac{dV_{out}}{dt} = i_{p_{min}} + c_1(t - t_1) - k_{s_n}(V_{in} - \theta - (1 + \delta)V_M)^{a_n} + i_{C_M} \quad (14)$$

where V_M according to Sections 2 and 3 can be written as $V_M = V_{M_1} + mt$ where

$$V_{M_1} = V_M[t_1] - \frac{V_p - V_M[t_1]}{\tau - t_1} t_1 \quad \text{and} \quad m = \frac{V_p - V_M[t_1]}{\tau - t_1}.$$

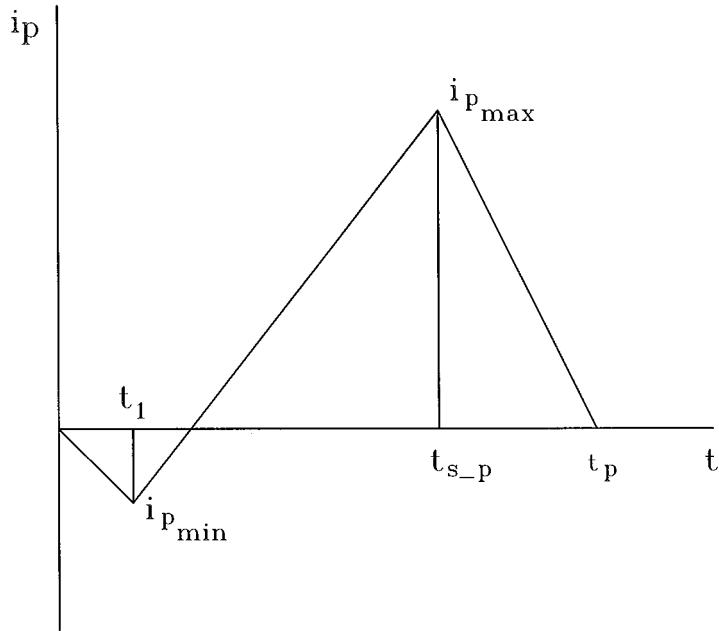


Figure 6. Representation of the pMOS transistor short-circuit current

The above equation can be solved for V_{out} as a function of c_1 , $V_{\text{out}} = f(c_1)$:

$$V_{\text{out}} = \frac{1}{C_L + C_M} \left(k_2 t + \frac{c_1}{2} t^2 + g[t] \right) + C[2] \quad (15)$$

where $k_2 = i_{p_{\min}} + C_M s - c_1 t_1$, $k_3 = s - m(1 + \delta)$, $k_4 = \theta + (1 + \delta)V_M$, $g[t] = k_{s_n}((k_3 t - k_4)^{a_n}/(k_3(1 + a_n)))$ ($k_4 - k_3 t$) and $C[2]$ is the integration constant.

Since the exact expression for the pMOS current is known, the approximated current expression can be set equal to the exact expression for one point in this region (e.g. $t_p/2$ where t_p is the time when the pMOS transistor becomes cut-off) in order to obtain the slope c_1 :⁵

$$i_{p_{\min}} + c_1 \left(\frac{t_p}{2} - t_1 \right) = k_{l_p} \left(\left| V_{\text{in}} \left[\frac{t_p}{2} \right] - V_{\text{DD}} \right| - |V_{\text{TP}}| \right)^{a_p/2} \left| V_{\text{out}} \left[\frac{t_p}{2} \right] - V_{\text{DD}} \right| \quad (16)$$

The calculated value for c_1 can be substituted in (15) in order to calculate the output voltage expression in this region.

The time point t_{s-p} when the pMOS transistor enters saturation is obtained by equating the drain saturation voltage of the pMOS transistor to its actual drain-to-source voltage:

$$V_{\text{D-SATP}} = V_{\text{DS}_p} \Rightarrow \frac{k_{s_p}}{k_{l_p}} (|V_{\text{in}} - V_{\text{DD}}| - |V_{\text{TP}}|)^{a_p/2} = |V_{\text{out}} - V_{\text{DD}}| \quad (17)$$

Now, the maximum value for the pMOS transistor current can be found as

$$i_{p_{\max}} = i_{p_{\min}} + c_1 (t_{s-p} - t_1) \quad (18)$$

Region 3. $t_{s-p} \leq t < t_2$: Both the pMOS and the topmost nMOS transistor operate in saturation. The pMOS current expression after time t_{s-p} is again assumed linear (Figure 6) and since its value at $t = t_p$ is almost zero (where $t_p = ((V_{\text{DD}} - |V_{\text{TP}}|)/V_{\text{DD}})\tau$ is the time when the pMOS transistor ceases to conduct), it can be written as

$$i_p(t) = i_{p_{\max}} - \frac{i_{p_{\max}}}{t_p - t_{s-p}} (t - t_{s-p}) = r_1 - r_2 t \quad (19)$$

where $r_1 = i_{p_{\max}} + r_2 t_{s-p}$ and $r_2 = i_{p_{\max}}/(t_p - t_{s-p})$.

Equation (10) can be written as

$$C_L \frac{dV_{\text{out}}}{dt} = i_p(t) - k_{s_n} (V_{\text{in}} - \theta - (1 + \delta)V_M)^{a_n} + i_{C_M} \quad (20)$$

which has the solution

$$V_{\text{out}} = \frac{1}{C_L + C_M} \left[(r_1 + C_M s) t - \frac{r_2}{2} t^2 + g[t] \right] + C[3] \quad (21)$$

The time limit t_2 of this region, when the topmost nMOS transistor exits saturation, can be found by solving

$$V_{\text{D-SATN}} = V_{\text{DS}_n} \Rightarrow \frac{k_{s_n}}{k_{l_n}} (V_{\text{in}} - \theta - (1 + \delta)V_M)^{a_n/2} = V_{\text{out}} - V_M \quad (22)$$

Region 4. $t_2 \leq t < t_p$: The pMOS transistor operates in saturation while all the nMOS transistors in linear mode. In order to solve equation (10) for this region, the value of the input in the i_n expression is approximated by its average value in this region. Additionally, in the i_n expression the value of the source voltage of the topmost transistor in the chain is approximated by its value at the beginning of this region. Since all transistors in the chain operate in the linear region, the transistor chain can be considered as a voltage divider and V_{DS} of the topmost transistor equal to $(1/n)V_{out}$ where n is the number of the transistors in the chain (for equal transistor widths in the chain). According to this, equation (10) can be written as

$$C_L \frac{dV_{out}}{dt} = i_p(t) - k_{l_n} \left(V_{in} \left[\frac{t_2 + t_p}{2} \right] - \theta - (1 + \delta) \frac{n-1}{n} V_{out}[t_2] \right)^{a_n/2} \frac{1}{n} V_{out} + i_{C_M} \quad (23)$$

resulting in

$$V_{out} = \frac{k_5(C_{MS} + r_1) + r_2(C_L + C_M)}{k_5^2} - \frac{r_2 t}{k_5} + C[4] e^{-(k_5/(C_L + C_M))t} \quad (24)$$

where

$$k_5 = k_{l_n} \left[\frac{s(t_2 + t_p)}{2} - \theta - (1 + \delta) \frac{n-1}{n} V_{out}[t_2] \right]^{a_n/2} \frac{1}{n}$$

Region 5. $t_p \leq t < \tau$: The input is still in transition and all transistors in the chain operate in the linear region. The pMOS device is cut-off. Equation (10) becomes

$$C_L \frac{dV_{out}}{dt} = -k_{l_n} \left(V_{in} \left[\frac{t_p + \tau}{2} \right] - \theta - (1 + \delta) \frac{n-1}{n} V_{out}[t_p] \right)^{a_n/2} \frac{1}{n} V_{out} + i_{C_M} \quad (25)$$

where the same type of approximations with that of the previous region have been made, employing the fact that all transistors in the chain operate in the linear region. The solution of the differential equation in this region is

$$V_{out} = \frac{C_{MS}}{k_6} + C[5] e^{-(k_6/(C_L + C_M))t} \quad (26)$$

where

$$k_6 = k_{l_n} \left(V_{in} \left[\frac{t_p + \tau}{2} \right] - \theta - (1 + \delta) \frac{n-1}{n} V_{out}[t_p] \right)^{a_n/2} \frac{1}{n}$$

Region 6. $t \geq \tau$: The input has reached its final value and all other conditions are as in the previous region. Consequently, equation (10) is the same with that of Region 5, without having to approximate the input voltage. Its solution is of the form

$$V_{out} = C[6] e^{-(k_7/(C_L + C_M))t} \quad (27)$$

where

$$k_7 = k_{l_n} \left(V_{DD} - \theta - (1 + \delta) \frac{n-1}{n} V_{out}[\tau] \right)^{a_n/2} \frac{1}{n}$$

The previous analysis has been performed for a specific sequence of time points t_{s-p} , t_2 , t_p and τ . However the same methodology is valid for every possible combination of input transition time, transistor widths and output load (which means different sequence of the above time points), since the differential equation at the output node can be solved in all the resulting regions, if the form of the source voltage of the topmost transistor is treated as in this paper.

In Figure 7, a comparison of the output voltage which is calculated according to the previous analysis to that obtained by SPICE simulations is shown for a 4-input NAND gate with $W_n = 8 \mu\text{m}$, $W_p = 3 \mu\text{m}$, $C_L = 100 \text{ fF}$ and a $0.5 \mu\text{m}$ HP technology for two input transition times.

Once the output voltage expression for each region is known, the propagation delay of a CMOS gate can be calculated as the time from the half- V_{DD} point of the input to the half- V_{DD} point of the output. Using this definition analytical expressions have been developed and propagation delay was calculated for NAND/NOR gates with several configurations and input transition times. It was observed that in all cases the calculated propagation delay is very close to SPICE simulation results. In Figure 8, a comparison of calculated and simulated propagation delay values is shown for a 4-input NAND gate with $W_n = 8 \mu\text{m}$, $W_p = 3 \mu\text{m}$, $C_L = 100 \text{ fF}$ and a $0.5 \mu\text{m}$ HP technology for several input transition times. In the same figure the propagation delays which are obtained when the gate is replaced by an equivalent inverter using conventional width reduction are also shown. The superiority of the proposed method is obvious.

It should be mentioned that in case the inputs which are applied to the gate are not normalised, that is if they do not have the same starting point and equal transition times, an input mapping algorithm should be applied in order to map all inputs to a set of equivalent normalised ones. Such an algorithm has been proposed in Reference 16 and presents high accuracy for a wide range of input transition times and relative distances in time of their starting points.

5. MODELLING THE PARASITIC BEHAVIOUR OF THE TRANSISTOR CHAIN

In this section the parasitic operation of an nMOS transistor chain, i.e. during charging of the output load of a NAND gate, is examined. The parasitic behaviour of the chain results in a short-circuit current which reduces the rate of charging of the output load.

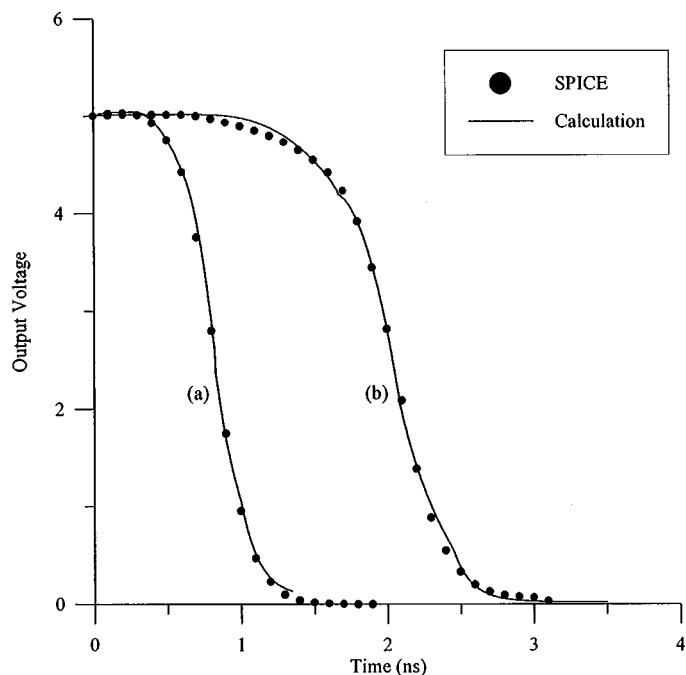


Figure 7. Output voltage comparison between calculated and simulated results for a 4-input NAND gate and (a) $\tau = 1 \text{ ns}$ and (b) $\tau = 3 \text{ ns}$

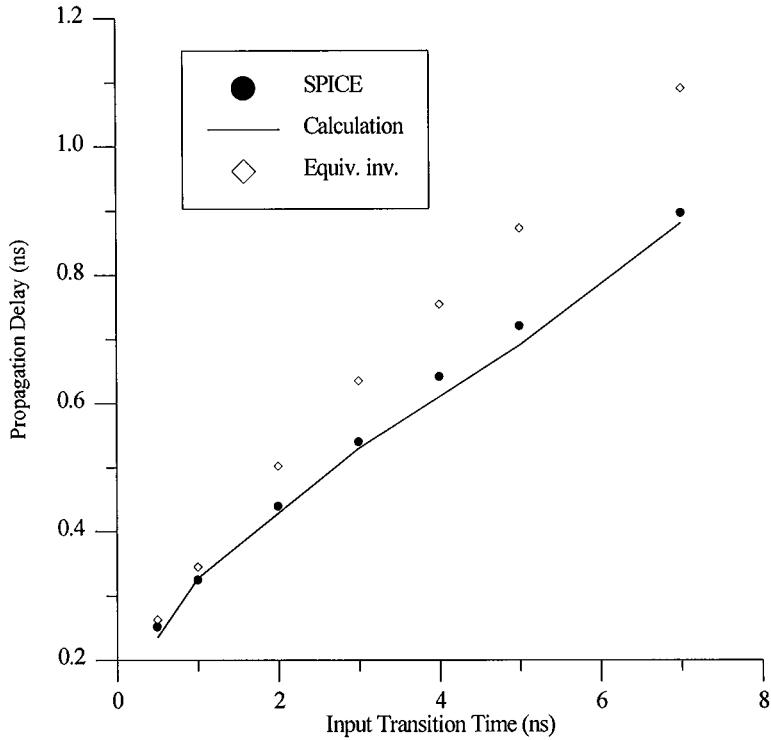


Figure 8. Propagation delays for a 4-input NAND gate measured using SPICE and calculated values using the proposed approach and that based on the conventional equivalent inverter, for several input transition times

Let us consider a NAND gate where all pMOS transistors have been replaced by an equivalent one as previously. The parasitic behaviour of the nMOS transistor chain will be modelled by an equivalent transistor. Consequently, according to the proposed method, in the case of charging output, the NAND gate diminishes to an equivalent inverter and the corresponding formulas can be used in order to calculate the output waveform, the propagation delay and the short-circuit power dissipation. A falling ramp input with transition time τ is considered to be applied to the gates of all transistors:

$$V_{in} = \begin{cases} V_{DD}, & t < 0 \\ V_{DD} - \frac{V_{DD}}{\tau} t, & 0 \leq t \leq \tau \\ 0, & t > \tau \end{cases} \quad (28)$$

The pMOS device starts conducting when the input reaches the threshold voltage ($V_{in} = V_{DD} - |V_{TPO}|$) at time $t = t_{fp}$. From this time on, current is flowing through the pMOS device and the load capacitance C_L charges. Since the nMOS devices are on when the pMOS transistor starts conducting, a short-circuit current is flowing through the gate from V_{DD} to the ground until time t_{fn} ($t_{fn} \approx ((V_{DD} - V_{TN})/V_{DD})\tau$) when the nMOS transistors cease to conduct. First, because the output voltage is small while the gate-to-source voltage of the NMOS devices is large, all these transistors start their operation in linear mode. As the output voltage rises, the voltages at the internal nodes of the chain are also increasing. All nMOS transistors have almost equal V_{DS} (voltage divider) while the topmost is biased by the smallest V_{GS} (since its source voltage has the largest value from all internal nodes). This means that this transistor at some time point will enter

saturation and after this time, the current in the chain will decrease and consequently the voltages at the internal nodes of the chain will also decrease, keeping all other devices in linear mode.

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

Since the pMOS transistor starts its operation in saturation, the output load will start to be charged by a current of the form $I_s = k_s(V_{\text{GS}} - |V_{\text{TP}}|)^a$. If we ignore the parasitic contribution of the nMOS transistor currents, the rate of the output voltage increase during $[t_{\text{fp}}, t_{\text{sat}}]$ is given by

$$C_L \frac{dV_{\text{out}}}{dt} = I_s(t) \Rightarrow \frac{dV_{\text{out}}}{dt} = \frac{I_s(t)}{C_L} = I^c(t) \quad (29)$$

where t_{sat} is the time when the top transistor in the nMOS transistor chain enters saturation and is approximated by $\frac{1}{2}(t_{\text{fp}} + t_{\text{fn}})$. It should be mentioned that the pMOS transistor until time t_{sat} operates in saturation, since the time point when it exits saturation for most of the cases is larger than t_{sat} . Considering the whole chain as a voltage divider for the interval t_{fp} to t_{sat} , the slope of each internal node voltage can also be calculated assuming a uniform distribution of the output voltage slope.

The current that each coupling capacitance is drawing during time interval $[t_{\text{fp}}, t_{\text{sat}}]$ is equal to

$$I_i = C_{M_i} \frac{d(V_i - V_{\text{in}})}{dt} = C_{M_i} \left(\frac{i \cdot I^c}{n} + s \right) \quad (30)$$

where n is the number of the transistors in the chain, s is the slope of the input and $i I^c/n$ the slope of the voltage waveform at the internal node i assuming equal transistor widths.

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

$$\sum_{i=1}^n I_i = C_{M_{\text{eq}}} (I^c + s) \quad (31)$$

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

When the node voltages are decreasing during $[t_{\text{sat}}, t_{\text{fn}}]$, the equivalent coupling capacitance can be found in a similar way. By symmetry, the same slope (with opposite sign) results for the voltage waveforms of the internal nodes.

Setting

$$c_r = \tilde{I}^c = \frac{1}{t_{\text{sat}} - t_{\text{fp}}} \int_{t_{\text{fp}}}^{t_{\text{sat}}} \frac{I_s(t)}{C_L} dt,$$

the equivalent coupling capacitance for the two time intervals can be written as (assuming equal coupling capacitances)

$$C_{M_{\text{eq}}} = C_M \frac{nc_r + (2n-1)s}{2(c_r + s)}, \quad [t_{\text{fp}}, t_{\text{sat}}] \quad (32)$$

$$C_{M_{\text{eq}}} = C_M \frac{(n-1)(s - c_r/2)}{(c_r + s)}, \quad [t_{\text{sat}}, t_{\text{fn}}] \quad (33)$$

If the contribution of the gate-to-drain coupling capacitance of the topmost transistor in the chain is neglected during $[t_{fp}, t_{sat}]$ and the average slope of the output node is taken equal to the input waveform slope (with opposite sign) the above equivalent capacitances diminish to

$$C_{M_{eq}} = C_M \frac{3(n-1)}{4}, \quad [t_{fp}, t_{sat}] \quad (34)$$

$$C_{M_{eq}} = C_M \frac{n-1}{4}, \quad [t_{sat}, t_{fn}] \quad (35)$$

which express directly their dependency on the number of transistors in the chain.

The improvement that is gained by inserting this coupling capacitance to the equivalent inverter model of a gate is significant as shown in Figure 9 which is a comparison of the output response of a 4-input NAND gate and that of the corresponding equivalent inverter with and without the calculated coupling capacitance.

It should be mentioned that in case the applied inputs to the parallel transistors, when these transistors drive the conducting current in a gate, are non-normalized, a mapping algorithm such as the one proposed in Reference 21 can be employed in order to map the applied input ramps to an equivalent normalized ramp that will be applied to the single equivalent transistor. Whenever a mapping algorithm is used for the inputs of the conducting part of a gate, the resulted input ramp should be applied to the transistor which replaces the short-circuiting part of the gate as well.

6. ESTIMATION OF THE SHORT-CIRCUIT POWER DISSIPATION

During the output switching of a gate and while both nMOS and pMOS transistor blocks are conducting, a path from V_{DD} to ground exists and causes short-circuit power dissipation. Short-circuit current and thus

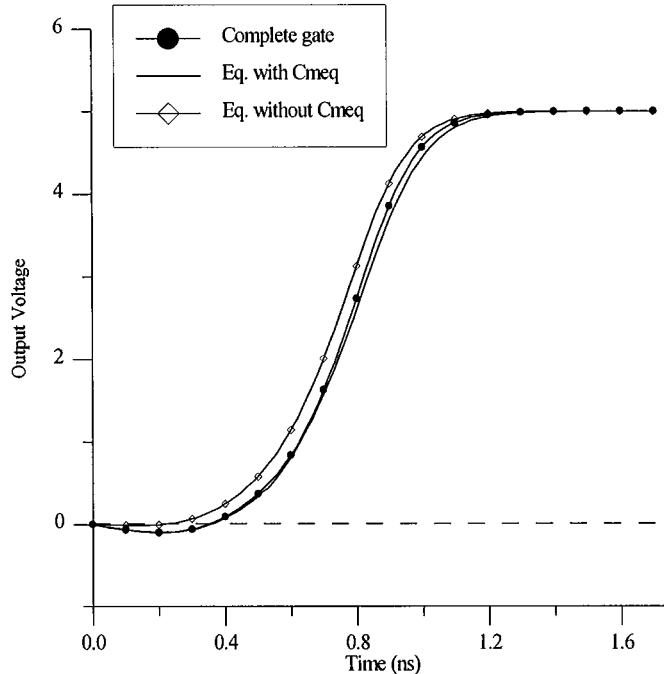


Figure 9. Output waveform comparison between the complete gate and the equivalent inverter with and without the equivalent coupling capacitance

short-circuit power dissipation appears during both charging and discharging of the output node of a gate. For complex CMOS gates these two cases (charging-discharging) are not symmetrical and consequently different approaches have to be followed for each one of them. Considering for a NAND gate the case of charging the output node, when the transistor chain operates parasitically, an equivalent inverter has been derived which models with high accuracy the behaviour of the gate. In order to estimate the short-circuit energy dissipation, E_{sc}^c , the corresponding formulas of the CMOS inverter can be directly applied.¹⁻⁴ In Figure 10 a comparison between the short-circuit energy dissipation of a 4-input NAND gate ($W_n = 8 \mu\text{m}$, $W_p = 3 \mu\text{m}$, $C_L = 100 \text{ fF}$ and $0.5 \mu\text{m}$ HP technology) when the nMOS transistor chain operates parasitically to that of the equivalent inverter is shown. The obtained accuracy is obvious.

For the case of discharging the output node, a method for calculating the short-circuit energy dissipation, E_{sc}^d , based on the analysis presented in Section 4, is proposed. The short-circuit energy dissipation begins when current starts flowing from V_{DD} towards the source node of the pMOS transistor at time point t_s because until then no current path exists from V_{DD} to ground.

In Section 4 the pMOS current was assumed linear during $[t_1, t_{s-p}]$ and $[t_{s-p}, t_p]$ and the current expression for each of these intervals was found. However, the current that is causing the short-circuit power dissipation is not the pMOS transistor current but the current that is flowing from V_{DD} towards the source of the pMOS transistor (i_{s_p}). In order to calculate this current, the Kirchhoff's current law has to be applied at the source node of the pMOS transistor (Figure 1a), which gives

$$i_{s_p} = i_p - i_{GS_p} \quad (36)$$

i_{GS_p} is the current through the gate-to-source capacitance C_{GS_p} and is given by $i_{GS_p} = C_{GS_p} dV_{in}/dt$. Since the input slope and i_p are known, the form of i_{s_p} and time points t_s and t_e , when i_{s_p} starts and ceases flowing from V_{DD} to ground ($i_{s_p} = 0$), can be obtained. The dissipated short-circuit energy during output discharging is calculated as

$$E_{sc}^d = V_{DD} \int_{t_s}^{t_e} i_{s_p}(t) dt \quad (37)$$

The calculated energy for a single output transition according to the above method lies very close to the energy which is measured from SPICE simulations. In Figure 11 a comparison of the calculated and

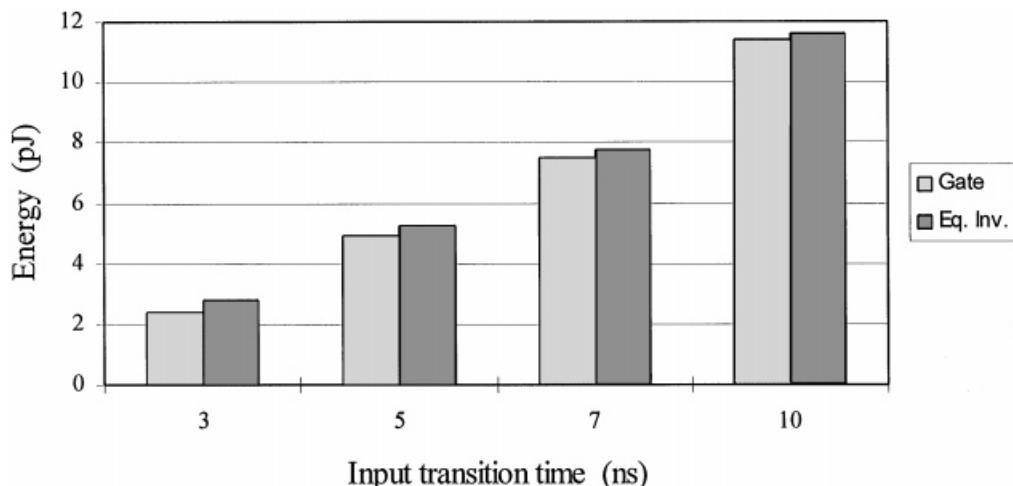


Figure 10. Comparison between short-circuit energy dissipation of a 4-input NAND gate and its equivalent inverter, when the transistor chain acts parasitically (falling input ramp), for several input transition times

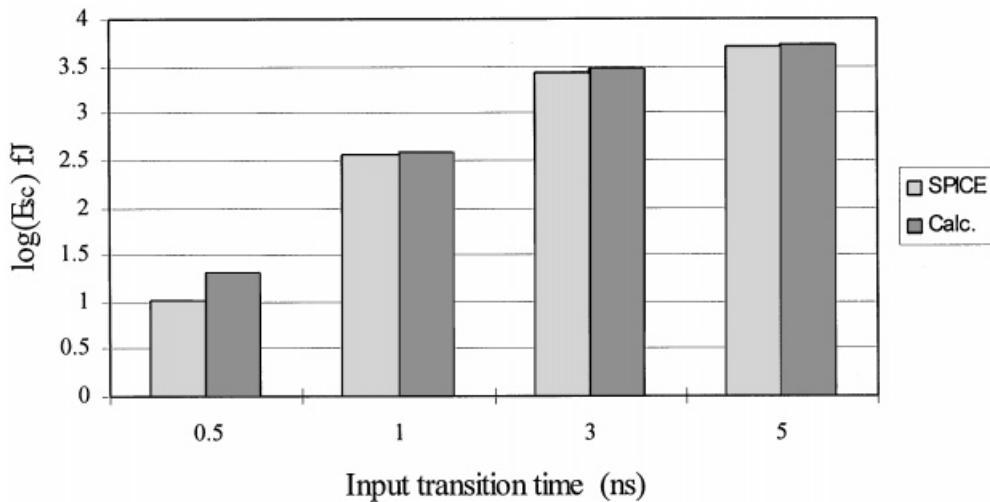


Figure 11. Comparison between calculated short-circuit energy dissipation during output discharging and that measured using SPICE for a 4-input NAND gate and for several input transition times (rising input ramp)

simulated short-circuit energy values during output discharging is shown for a 4-input NAND gate with $W_n = 8 \mu\text{m}$, $W_p = 3 \mu\text{m}$, $C_L = 100 \text{ fF}$ and a $0.5 \mu\text{m}$ HP technology for several input transition times. Using SPICE, the short-circuit power dissipation can be obtained by integrating the current at the source terminal of the pMOS transistors or by using a power meter.²²

Consequently, the short-circuit energy dissipation during a complete transition at the output node $[0 \rightarrow 1 \rightarrow 0]$ is $E_{sc} = E_{sc}^c + E_{sc}^d$ and the corresponding power can be calculated simply by multiplying the calculated energy with the frequency of transitions at the output of the gate.

According to the proposed method the output waveform, propagation delay and short-circuit power dissipation can be calculated for NAND/NOR gates. Complex gates can be treated by collapsing them to the equivalent NAND/NOR gates using algorithms such as the one proposed in Reference 15.

7. CONCLUSIONS

In this paper a complete modelling technique for the calculation of the propagation delay and the short-circuit power dissipation of CMOS gates has been presented. The proposed method examines the gate operation during output charging and discharging and takes into account second order effects such as the carrier velocity saturation of short-channel devices, the body effect and the coupling capacitance and is developed for non-zero transition time inputs. The starting point of conduction of a gate is efficiently calculated and the parasitic behaviour of the short-circuiting part of a gate is accurately modelled. In case the conducting path consists of serially connected transistors, the output waveform has been obtained through an analytical approach which leads to expressions for the estimation of the propagation delay and the short-circuit power dissipation. In case the conducting path consists of parallel connected transistors an equivalent inverter model is proposed and the corresponding formulas which have been developed for the inverter can be applied. The calculated results for the output waveform, the propagation delay and the short-circuit power dissipation match very well SPICE simulation results.

REFERENCES

1. N. Hedenstierna and K. O. Jeppson, CMOS circuit speed and Buffer optimization., *IEEE Trans. Comput. Aided Des.*, **CAD-6**(2), 270–281 (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*, **25**(2), 584–594 (1990).
3. L. Bisdounis, S. Nikolaidis and O. Koufopavlou, Propagation delay and short-circuit power dissipation modeling of the CMOS inverter, *IEEE Trans. Circuits and Systems—I: Fund. Theory Appl.*, **45**(3), 259–270 (1998).
4. A. Hirata, H. Onodera and K. Tamaru, Estimation of short-circuit power dissipation for static CMOS gates, *IEICE Trans. Fund.*, **E79-A**(3), 304–311 (1996).
5. 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*, **33**(2), 302–306 (1998).
6. A. Nabavi-Lishi and N. C. Rumin, Inverter models of CMOS gates for supply current and delay evaluation, *IEEE Trans. Computer Aided Des Integrated Circuits Systems*, **13**(10), 1271–1279 (1994).
7. A. Chatzigeorgiou and S. Nikolaidis, Collapsing the transistor chain to an effective single equivalent transistor, *Proc. Design, Automation and Test in Europe Conf. and Exhibition (DATE)*, February 1998, pp. 2–6.
8. T. Sakurai and A. R. Newton, Delay analysis of series-connected MOSFET circuits, *IEEE J. Solid-State Circuits*, **26**(2), 122–131 (1991).
9. J. M. Daga, S. Turgis and D. Auvergne, Design Oriented Standard Cell Delay Modelling, *Proc. Int. Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, 1996, pp. 265–274.
10. 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.*, **18**, 289–306 (1990).
11. 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*, **2**(1), 100–114 (1994).
12. Y.-H. Shih and S.M. Kang, Analytic transient solution of general MOS circuit primitives, *IEEE Trans. Comput. Aided Des.*, **11**(6), 719–731 (1992).
13. Y.-H. Shih, Y. Leblebici and S. M. Kang, ILLIADS: a fast timing and reliability simulator for digital MOS circuits, *IEEE Trans. Comput. Aided Des. Integrated Circuits Systems*, **12**(9), 1387–1402 (1993).
14. A. Dharchoudhury, S. M. Kang, K. H. Kim and S. H. Lee, Fast and accurate timing simulation with regionwise quadratic models of MOS I-V characteristics, *Proc. IEEE Int. Conf. on Computer-Aided Design (ICCAD)*, Nov. 1994, pp. 190–194.
15. J.-T. Kong, S. Z. Hussain and D. Overhauser, Performance estimation of complex MOS gates, *IEEE Trans. Circuits Systems—I: Fund. Theory Appl.*, **44**(9), 785–795 (1997).
16. A. Chatzigeorgiou and S. Nikolaidis, Input mapping algorithm for modelling of CMOS circuits, *IEE Electron. Lett.* **34**(12), 1177–1179 (1998).
17. J. Qian, S. Pullela and L. Pillage, Modeling the “Effective Capacitance” for the RC interconnect of CMOS gates, *IEEE Trans. Comput. Aided Des. Integrated Circuits Systems*, **13**(12), 1526–1535 (1994).
18. V. Adler and E. G. Friedman, Repeater design to reduce delay and power in resistive interconnect, *IEEE Trans. Circuits Systems — II: Analog Digital Signal Process.*, **45**(5), 607–616 (1998).
19. S. Nikolaidis, A. Chatzigeorgiou and E.D. Kyriakis-Bitzaros, Delay and power estimation for a CMOS inverter driving RC interconnect loads, *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS)*, Vol. VI, 1998, pp. 368–371.
20. J. M. Rabaey, *Digital Integrated Circuits: A Design Perspective*, Prentice-Hall, Upper Saddle River, NJ, 1996.
21. 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. Comput. Aided Des.*, **8**(9), 1027–1032 (1989).
22. S. M. Kang, Accurate simulation of power dissipation in VLSI circuits, *IEEE J. Solid State Circuits*, **SC-21**(5), 889–891 (1986).