TÍTULO DE LA TESIS
“POWER AND TIMING MODELING OF SUBMICRON CMOS GATES”

AUTOR:
Josep Lluís Rosselló

DIRECTOR:
Jaume Segura Fuster

FECHA DE LECTURA
22 de Febrero de 2003

PUBLICACIONES DERIVADAS

REVISTAS


CONGRESOS

Autores: J.L. Rosselló, E. Isern, M. Roca, E. García, J. Segura
Título: An analytical model of the charge driven by CMOS buffers
Tipo de participación: comunicación a congreso
Congreso: XII Design of Circuits and Integrated Systems Conference
Publicación: Proceedings, pp. 471-474
Lugar de celebración: Sevilla Fecha: Noviembre 1997

Título: Transient current testing based on current (charge) integration
Tipo de participación: comunicación a congreso
Congreso: International Workshop on IDDQ Testing
Publicación: Proceedings, pp. 26-30
Lugar de celebración: San Jose, California, USA Fecha: Noviembre 1998
Autores: J.L.Rosselló, E.Isern, M.Roca and J.Segura
Título: An analytical model of CMOS buffers power consumption
Tipo de participación: comunicación a congreso
Congreso: XIV Design of Circuits and Integrated Systems Conference
Publicación: Proceedings, pp. 125-130
Lugar de celebración: Palma de Mallorca, Baleares, España Fecha: Noviembre 1999

Autores: J.L.Rosselló and J.Segura
Título: A physical modeling of the alpha-power law MOSFET model
Tipo de participación: comunicación a congreso
Congreso: XV Design of Circuits and Integrated Systems Conference
Publicación: Proceedings, pp 65-70
Lugar de celebración: Montpellier, Francia Fecha: Noviembre 2000

Autores: J.L.Rosselló and J.Segura
Título: A simple power consumption model of CMOS buffers driving RC interconnect lines
Tipo de participación: comunicación a congreso
Congreso: XI International Workshop on Power and Timing Modeling
Publicación: Proceedings, Artículo no. 4.2

Autores: J.L.Rosselló and J.Segura
Título: Power-delay modeling of Dynamic CMOS gates for circuit optimization
Tipo de participación: comunicación a congreso
Congreso: the International Conference on Computer Aided Design (ICCAD’01)
Publicación: Proceedings, pp 494-499
Lugar de celebración: San José, CA, USA Fecha: Noviembre 2001

Autores: J.L.Rosselló and J.Segura
Título: Modeling the input-output coupling capacitor effects on the CMOS buffer power consumption
Tipo de participación: comunicación a congreso
Congreso: XVI Conference on Design of Circuits and Integrated Systems
Publicación: Proceedings, pp 613-617
Lugar de celebración: Porto, Portugal Fecha: Noviembre 2001

Autores: J.L.Rosselló and J.Segura
Título: A compact charge-based propagation delay model for submicronic CMOS buffers
Tipo de participación: comunicación a congreso
Congreso: XII International Workshop on Power and Timing Modeling (PATMOS 2002),
Publicación: Proceedings, pp 219-228
Lugar de celebración: Sevilla, España Fecha: Septiembre 2002
Autores: J.L. Rosselló and J. Segura
Título: A Simple Analytical Description of Power Dissipation in Submicronic CMOS Buffers
Tipo de participación: comunicación a congreso
Congreso: XVII Conference on Design of Circuits and Integrated Systems
Publicación: Proceedings, pp 627-630
Lugar de celebración: Santander, España Fecha: Noviembre 2002

Autores: J.L. Rosselló and J. Segura
Título: Power and Timing modeling of Submicron CMOS Gates
Tipo de participación: Presentación poster
Congreso: 40th Design Automation Conference (DAC’03)
Publicación: Proceedings, pp 627-630
Lugar de celebración: Anaheim, CA, USA. Fecha: Junio 2003

Autores: J.L. Rosselló and J. Segura
Título: A Compact Charge-Based Crosstalk Induced Delay Model for Submicronic CMOS Gates
Tipo de participación: Comunicación a congreso
Publicación: Proceedings, pp 627-630

Autores: J.L. Rosselló and J. Segura
Título: A physically-based nth power law MOSFET model for efficient CAD implementation
Tipo de participación: Comunicación a congreso
Congreso: XVII Design of Circuits and Integrated systems Conference (DCIS’03)
Publicación: Proceedings, pp 380-383

Autores: J.L. Rosselló and J. Segura
Título: A Compact Propagation Delay Model for Deep-submicron CMOS Technologies including Crosstalk Effects
Tipo de participación: Comunicación a congreso
Congreso: Design Automation and Test in Europe (DATE’04),
A Variable Threshold Voltage Inverter for CMOS Programmable Logic Circuits

J. Segura, J. L. Rosselló, J. Morra, and H. Sigg

Abstract — A programmable input threshold voltage inverter compatible with double gate transistors fabrication processes is presented. Such a circuit is useful as a programmable input threshold buffer for general purpose circuits that can be included in both TTL and CMOS environments, or can be used as low cost analog programmable comparator. A prototype is fabricated and measured.

Index Terms — Buffer circuits, integrated circuit design.

I. INTRODUCTION

In this work we develop a CMOS inverter compatible with double gate MOS technologies that can be programmed to different predefined threshold voltages. In the context of this work, the inverter logic threshold voltage ($V_{th}$) is defined as the static input voltage at which the inverter output is at $V_{DD}/2$. An input logic threshold programmable CMOS inverter is of high interest in applications such as CMOS design of input buffers for general purpose circuits (as microcontrollers) to be used in both CMOS and TTL applications. Once the final application or environment of the circuit is known, the input circuitry can be programmed to match the outside signal threshold voltage (TTL or CMOS), thus optimizing the noise margin and the power consumption, as will be shown. A programmable logic threshold inverter is also of interest in analog applications, to be used as a voltage comparator. Instead of using a whole analog block (as an A/D or opamp) to determine if a signal is above or below a given reference, the proposed variable logic threshold inverter can be programmed to a voltage being the reference voltage required for each application. The inverter has the advantage of a significant area reduction given its simplicity and the reduced number of transistors required. Additionally, a comparator requires two inputs, while a programmable logic threshold buffer used as a comparator would only require a single package pin.

The design proposed can be divided into two blocks: the programming circuitry and the buffer circuitry. The number of transistors of the buffer determines the different logic threshold levels of the whole inverter. A buffer with $n + 1$ transistors ($n$, N-type transistors, and one P-type MOS), can be programmed to $2^n - 1$ logic threshold voltages between $V_{th}$ and $(V_{DD} - |V_{tp}|)$. $V_{tn}$ and $V_{tp}$ being the threshold voltages of the n-MOS and p-MOS transistors, respectively. The $2^n - 1$ values that $V_{th}$ can take are determined by the relative sizing of the transistors composing the buffer.

In the next section we present the design of the inverter, while the experimental results are reported in Section III. Finally, the conclusions are reported in Section IV.

II. CIRCUIT OPERATION

The logic threshold voltage of a CMOS inverter is given by the aspect ratio $W/L$ of the N and P transistors. In general, both transistors have the same length. Thus, the $W_p/W_n$ ratio (the ratio between the transistor widths) is the value that determines the inverter logic threshold voltage. For general applications, inverters are designed to have a symmetric static transfer characteristic, i.e., the $V_{out}/V_{in}$ curve intersects the unity-gain line $V_{in} = V_{out}$ at $V_{DD}/2$. Therefore, using the Shockley model to describe the behavior of the MOS transistor, and assuming the same value of the gate oxide thickness for both transistors, then the aspect ratio that gives a symmetric transfer characteristic has the well-known expression

$$\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p} \left[ 1 - \frac{2V_{th}}{V_{DD}} \right]^2$$

where $\mu_n$ and $\mu_p$ are the electron and hole mobility, respectively, and $V_{tn}$ and $V_{tp}$ are the transistors threshold voltages.

In those cases where the transfer characteristic is not symmetric, the logic threshold voltage can be expressed as

$$V_{th} = \frac{V_{tn} + \sqrt{\frac{\mu_p W_p}{\mu_n W_n} (V_{DD} - |V_{tp}|)}}{1 + \sqrt{\frac{\mu_p W_p}{\mu_n W_n}}}$$

where it has been assumed that both transistors have the same length (i.e., $L_n = L_p$).

A. Buffer Design

A simplified schematic of a programmable logic threshold voltage inverter (with the programming circuitry not shown) is given in Fig. 1. Its operation is based on the expression stated in (2). The inverter uses a single gate p-MOS enhancement transistor and a pulldown transistors net composed of one single gate n-MOS enhancement transistor and several double gate n-MOS devices connected in parallel. The value of $W_p$ is fixed, while the effective value of $W_n$ can be changed.
The different logic threshold voltages of the programmable inverter, $V_{th_i}$, can be derived from (2) obtaining

$$V_{th_i} = V_t + \frac{\mu_p W_p}{\mu_n (W_n + \alpha \sum_{j=1}^{n-1} p_j W_{n_j})^{1/2}} \left( V_{DD} - |V_{by}| \right)$$

$$+ 1 + \frac{\mu_p W_p}{\mu_n (W_n + \alpha \sum_{j=1}^{n-1} p_j W_{n_j})^{1/2}} \left( V_{DD} - |V_{by}| \right)$$

(3)

where $\alpha = C_{ox}/C_{ox}'$ is the ratio between the gate oxide capacitance per unit area for the double and single gate transistors, $W_{n_j}$ is the width of the $j$th double gate n-MOS device, and $p_j$ is a Boolean variable associated to the $j$th double gate transistor programming state. The value of $p_j$ is one when such device is unprogrammed and zero otherwise.

Each programming configuration of double gate transistors has associated a unique set of Boolean variables $P_j$ that define a threshold level vector (TLV), $P_m$, as

$$P_m = (p_1, p_2, \cdots, p_{n-1}).$$

(4)

The subindex $m$ identifies each vector and is the decimal representation of the TLV, defined as

$$m = p_1 2^{n-2} + \cdots + p_i 2^{n-1-i} + \cdots + p_{n-1} 2^0.$$  

(5)

If $W_{n_j} \neq W_{n_j+1} \forall j \in \{1, \cdots, n-2\}$, each one of the $2^{n-1}$ TLV’s will lead to a different value of $V_{th_i}$.

The $W_p/W_n$ ratio (the width of transistors $P$ and $N_0$ in Fig. 1) is used to set the maximum value that $V_{th_i}$ can take, i.e., the logic threshold voltage associated to the TLV $P_0 = 00 \cdots 0$. Only $n-1$ of the whole $2^{n-1}$ remaining $V_{th_i}$ voltages can be independently chosen. The remaining levels will be a combination of the $n$ independent ones. We use a simple and straightforward technique to determine the size of each double gate transistor: associate each logic threshold level to one of the double gate transistors. This means that the inverter will be programmed to the logic threshold level $V_{th_i}$ using the TLV $P_i$ that has its vector components defined as

$$p_j = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \text{ for } j = 1 \text{ to } n-1.$$  

(6)

Therefore, only both the $N$ and $P$ single gate transistors and one of the double gate devices (the transistor $N_i$ for the TLV $P_i$) must be considered to calculate the width of each programmable transistor. From (3) we get

$$W_{N_i} = \frac{\mu_p W_p (V_{th_i} - V_{DD} + |V_{by}|)^2 - \mu_n W_n (V_{th_i} - V_{IN})^2}{\alpha \mu_n (V_{th_i} - V_{by})^2}.$$ 

(7)

B. General Design

The whole programmable inverter, containing both the buffer and the programming circuitry, is shown in Fig. 2. The programming circuitry consists of transistors $G_0, \cdots, G_{n-1}$ that connect or isolate the gate of the n-MOS transistors from the input buffer depending on the $N/F$ value, and transistors $N_{P_1}, \cdots, N_{P_{n-1}}$ that drive the gate programming signal $V_{PG}$. One register—the Programming Register—holds the TLV during the programming mode, and is reset in normal operation to turn off all the $N_{P_i}$ transistors. When $N/F = 1$, the inverter is in the normal mode, the input $I_{in}$ drives all the n-MOS transistors, and the equivalent active circuitry is the same as that in Fig. 1. Given that no dc current goes into the gate of the n-MOS devices, the size of the $G_i$ transistors can be minimum. When $N/F = 0$, the inverter is in the programming mode, the inverter input is disconnected from the n-MOS transistors gates, and the double gate transistors are driven by the contents of the Programming Register. In this mode, the inverter input is used to drive the drain of the double gate devices through the p-MOS transistor. During programming, the TLV is held in the Programming Register to drive the programming gate signal ($V_{PG}$) to the desired devices through the $N_{P_1}, \cdots, N_{P_{n-1}}$ transistors. The function of bit $p_0$ and transistor $G_0$ is different from the remaining $p_j$ bits and $N_{P_j}$ devices, and are used to turn off the nonprogrammable transistor $N_0$ to prevent current through this device in the programming mode.

In applications where more than one inverter is used, the programming register is shared by all the inverters of the circuit, and the input signal $I_{in}$ is used to select which inverter is being programmed.

C. A Programmable TTL/CMOS Input Inverter

A particular case of the general design is a programmable TTL/CMOS input buffer. Such a circuit would be the simpler case of a programmable logic threshold inverter given that only one double gate device is required. The width ratio of the $P$ and $N$ single gate transistors is used to set the logic threshold level of the CMOS buffer configuration, while the width of the double gate transistor adjusts the TTL configuration input logic threshold. Given that the circuit is initially unprogrammed, it is set to the TTL configuration by default. The double gate device must be programmed and turned off to switch to the CMOS configuration. The TLV for an input TTL/CMOS buffer has a single bit, $P_1$, which is set to program the CMOS configuration.
and reset for the TTL configuration. In the next section we present the results obtained from a fabricated prototype.

III. EXPERIMENTAL RESULTS

We designed and fabricated a CMOS-TTL programmable input buffer using the Philips 23G 1µ double poly single metal stacked gate EPROM process. The circuit was placed on the scribe of a wafer, and the sensitivity of the input buffer logic threshold to process variations was characterized. Only the buffer circuitry was fabricated. The programming circuitry was not included as the devices were accessed separately with probe tips for programming. Measurements were made with an HP4155 Semiconductor Parameter Analyzer. Fig. 3 shows a photograph of the circuit. The prototype was designed to have two programmable logic threshold voltages corresponding to the TTL and the CMOS switching level. The TTL configuration was designed to have \( V_{\text{TTL}} \), while the CMOS configuration was \( V_{\text{CMOS}} \).

Fig. 4 reports the voltage and current static transfer characteristics of the buffer for both the TTL and CMOS modes. The experimental value obtained for \( V_{\text{TTL}} \) is 1.35 V, while the CMOS configuration logic threshold voltage is 3 V [Fig. 4(a)]. The deviations from the theoretical values are 0.05 V for the TTL configuration and 0.3 V for the CMOS one. These small deviations are due to the use of the Shockley transistor model in (7) to derive the effective widths of programmable transistors. The use of more accurate models that take into account second order effects of the transistor would correct these small deviations.
TABLE I

<table>
<thead>
<tr>
<th>Circuit</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>Max. dev.</th>
<th>Mean dev.</th>
<th>Std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTL</td>
<td>1.35</td>
<td>1.37</td>
<td>1.39</td>
<td>1.41</td>
<td>1.28</td>
<td>1.32</td>
<td>8.5%</td>
<td>0.05</td>
<td>0.046</td>
</tr>
<tr>
<td>CMOS</td>
<td>2.99</td>
<td>3.02</td>
<td>3.02</td>
<td>3.02</td>
<td>3.02</td>
<td>3.05</td>
<td>11.4%</td>
<td>0.23</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Fig. 4(a) shows that programming the input buffer logic threshold improves the noise margin of each application. Fig. 4(b) reports that the power consumption of the buffer is optimized when the circuit is used in a CMOS environment, obtaining a reduction in the current from 60 μA to less than 1 μA (for \( V_{\text{in}} = 1 \) V). Therefore, the switching power consumption will be lowered in a CMOS environment using a programmable inverter because the maximum static current dissipation is reduced from 120 to 20 μA. Such a reduction will contribute to lower short-circuit current dissipation, and therefore the overall circuit consumption [2]–[4].

Table I shows the input buffer logic threshold values measured for different circuits. It can be observed that the TTL buffer logic threshold values have a larger variation than the CMOS buffer logic threshold values, but in both cases the variations are small.

Maximum speed transient measurements were not possible since we did not include output buffers to drive the large capacitance of the bond pads and measurement equipment [5]. The operation speed of the input buffer can be enhanced by adding normal inverter stages to obtain a tapered buffer. The number of stages and its size depends on the capacitance to be driven as has been extensively studied [6]–[8].

The penalty of the area overhead is small given that the size of the \( G_t \) transistors added to each inverter for the programming circuitry can be minimum sized. Additionally, only one programming register and one set of \( N_{P_1} \) pass transistors are required for all programmable buffers in the die. The programming register and \( N_{P_1} \) devices can be distributed anywhere in the design which contributes to compact layout.

IV. CONCLUSIONS

A programmable input logic threshold voltage inverter for double gate transistor technologies is described, along with its principle of operation and design procedure. A CMOS-TTL input buffer has been designed and fabricated showing that it can be programmed to the CMOS and TTL levels depending on the outside environment of the final application. Measurements show good agreement between predicted and obtained logic threshold voltages. Measurements also show a significant enhancement of the noise margins of the circuit. Finally, the input buffer provides the capability of reducing the input circuitry power consumption, especially for CMOS applications.

ACKNOWLEDGMENT

The authors would like to thank D. Vancil from Phillips USA for his assistance in the buffer design.

REFERENCES

Charge-Based Analytical Model for the Evaluation of Power Consumption in Submicron CMOS Buffers

José Luis Rosselló, Member, IEEE, and Jaume Segura, Member, IEEE

Abstract—The authors present an accurate analytical method for analyzing the power consumption in CMOS buffers. It is derived from the charge transferred through the circuit and makes use of the physically based MM9 MOSFET model (Velghe et al., 1994), (Foty et al., 1997) as well as a modified Sakurai alpha-power law model. The resulting analytical model accounts for the effects of input slew time, device sizes, carrier velocity saturation effects, input-to-output coupling capacitance, output load, and temperature. Results are compared to HSPICE simulations (level 50) and to other models previously published considering a large set of parameters for a 0.18 and 0.35 μm technologies, showing significant improvements.

Index Terms—Closed form expressions, deep submicron, power modeling and estimation.

NOMENCLATURE

A) Input Parameters to the Model

$\mu_0$ zero bias low field mobility;
$\theta_L$ gate field mobility reduction coefficient (MM9 parameter);
$n$, $p$ nMOS, pMOS gate field mobility reduction coefficient (MM9 parameter);
$n$, $p$ drain field mobility reduction coefficient (MM9 parameter);
$n$, $p$ nMOS, pMOS drain field mobility reduction coefficient (MM9 parameter);
$C_L$ buffer output load;
$C_{gr}$ gate oxide capacitance;
$V_{DD}$ frequency;
$K$ substrate sensitivity when depleting bulk doping (MM9 parameter);
$KO$ substrate sensitivity when depleting surface doping (MM9 parameter);
$L$ channel length;
$L_n$ channel length of nMOS;
$L_p$ channel length of pMOS;
$L_{AP}$ gate source/drain underdiffusion (MM9 parameter);
$L_{AP_n}$, $L_{AP_p}$ nMOS, pMOS gate source/drain underdiffusion (MM9 parameter);
$\lambda_{VAR}$ process bias on the channel length (MM9 parameter);
$m$ normalization constant;
$PHIB$ surface potential for strong inversion (MM9 parameter);
$t_{on}$ transition time of input voltage;
$V_{SBSX}$ voltage of transition between $KO$ and $K$ (MM9 parameter);
$V_{th}$ physical threshold voltage with no substrate or drain bias (MM9 parameter);
$V_{thn}$, $V_{thp}$ nMOS, pMOS physical threshold voltage with no substrate or drain bias (MM9 parameter);
$W$ channel width;
$W_n$, $W_p$ nMOS, pMOS Channel width;
$W_{OT}$ isolation reduction of channel width (MM9 parameter);
$W_{VAR}$ process bias on the channel width (MM9 parameter);

B) Intermediate Parameters

$\alpha$ velocity saturation index;
$\alpha_n$, $\alpha_p$ nMOS, pMOS velocity saturation index;
$\beta$ MOSFET conductivity;
$\beta_n$, $\beta_p$ nMOS, pMOS conductivity;
$\delta$ body effect coefficient;
$\delta_n$ nMOS body effect coefficient;
$\delta V$ overvoltage;
$\delta V_{max}$ maximum overvoltage;
$\delta V_TH$ overvoltage at $t = t_n$;
$\Delta Q_{out}$ charge stored at the output node flowing through the nMOS transistor when the output is discharged;
$\Delta V_{out}$ output voltage swing;
$\rho_{eff}$ effective carriers mobility;
$\rho_{s}$ interconnect resistance per square;
$C_a$ area capacitance;
$C_f$ fringing capacitance to underlying conductor for single line;
$C_M$, $C_H^M$, $C_L^M$ input to output coupling capacitance;
$C_{min}$ minimum output capacitance allowed by the technology;
$C_S$ total output capacitance when the input voltage is low;
$E_{sec}$ short-circuit energy;
$E_{sec}$, $E_{sec}$ short-circuit energy dissipated in a falling, rising input transition;
$E_{tr}$ transient energy;
transient energy dissipated in a falling, rising input transition;
\( I_{\text{tr}} \), \( I_{\text{pp}} \) nMOS, pMOS drain current;
\( I_{\text{DD}} \) drain current at \( V_{\text{DS}} = V_{\text{GS}} = V_{\text{DD}} \);
\( I_{\text{DD0}} \) drain current at \( V_{\text{DS}} = V_{\text{DD}} \) and \( V_{\text{GS}} = V_{\text{DD}} / 3 \);
\( I_{\text{DD0n}}, I_{\text{DD0p}} \) nMOS, pMOS drain current at \( V_{\text{DS}} = V_{\text{GS}} = V_{\text{DD}} \);
\( I_{\text{D0}} \) saturated drain current;
\( I_{\text{D0n}}, I_{\text{D0p}} \) nMOS, pMOS saturated drain current;
\( I_{\text{DS}} \) drain current;
\( I_{\text{L_{max}}}, I_{\text{C_{L_{max}}}} \) maximum short-circuit current;
\( I_{\text{C_{L_{max}}}}^{\max} \), \( I_{\text{C_{L_{max}}}}^{\min} \) maximum short-circuit current for heavily loaded buffers;
\( I_{\text{C_{L_{max}}}}^{\max} \), \( I_{\text{C_{L_{max}}}}^{\min} \) maximum short-circuit current for unloaded buffers;
\( I_{\text{SC}} \) short-circuit current;
\( L_{\text{eff}} \) effective channel length;
\( L_{\text{min}} \) minimum lithographic channel length allowed by the technology;
\( L_{\text{P_{max}}} \) effective channel length of pMOS;
\( P_{\text{total}} \) total power consumption;
\( Q_{\text{out}} \) charge stored at \( C_L \) and \( C_M \);
\( Q_{\text{C}} \) overshoot charge transferred (OCT);
\( Q_{\text{C}}^{\ell} \) overshoot charge transferred for a fast input transition \( (t_{\text{ov}} > t_{\text{n}} + T_{\text{sc}}) \);
\( Q_{\text{C}}^{p} \) overshoot charge transferred for a slow input transition \( (t_{\text{ov}} < t_{\text{n}} + T_{\text{sc}}) \);
\( Q_{\text{C}}^{\ell} \) overshoot charge fraction transferred during the interval \( t_{\text{n}} < t < t_{\text{ov}} \);
\( Q_{\text{C}}^{p} \) overshoot charge fraction transferred during the interval \( t < t_{\text{ov}} \);
\( Q_{\text{SC}} \) short-circuit charge transferred when \( C_M = 0 \);
\( Q_{\text{SC}}^{\ell} \) short-circuit charge transferred for a rising input transition;
\( R_{\text{n}} \) effective resistance of nMOS;
\( T_{\text{SC}}^{\ell} \) short-circuit current time interval when \( C_M = 0 \);
\( T_{\text{SC}}^{p} \) short-circuit current time interval;
\( t \) time; \( t_{\text{max}} \) time at which the short-circuit current is maximum;
\( t_{\text{max}}^{\ell} \) time at which the short-circuit current is maximum for unloaded buffers;
\( t_{\text{max}}^{p} \) time at which the short-circuit current is maximum for heavily loaded buffers;
\( t_{\text{max}}^{\ell} \) mathematical limit of \( t_{\text{max}}^{\ell} \) when \( C_L \rightarrow \infty \);
\( t_{\text{SC}}^{\ell} \) time at which the nMOS transistor starts to conduct;
\( t_{\text{ov}} \) overshoot time;
\( t_{\text{OV}}^{\ell} \) time at which the overshoot current is maximum;
\( t_{\text{OV}}^{p} \) time during which the pMOS device passes a negative current;
\( V_{\text{DD}} \) saturation voltage at \( V_{\text{GS}} = V_{\text{DD}} \);
\( V_{\text{DD0}} \) saturation voltage of pMOS at \( V_{\text{GS}} = V_{\text{DD}} \);
\( V_{\text{DDn}}, V_{\text{DDp}} \) saturation voltage; saturation voltage of pMOS;
\( V_{\text{DS}} \) supply voltage; drain voltage;
\( V_{\text{DSL}} \) smoothing function used by MM9 between limiting values \( V_{\text{DS}} \) and \( V_{\text{DD}} \);
\( V_{\text{GS}} \) gate voltage;
\( V_{\text{in}} \) input voltage;
\( V_{\text{out}} \) output voltage;
\( V_{\text{SC}} \) short-circuit input voltage swing for \( C_M = 0 \);
\( V_{\text{TH}}, V_{\text{TP}} \) nMOS, pMOS alpha-power law threshold voltage;
\( W_{\text{eff}}, W_{\text{r}} \) effective channel width; minimum lithographic channel width allowed by the technology;
\( W_{\text{ref}}, W_{\text{ref}} \) nMOS, pMOS effective channel width.

I. INTRODUCTION

POWER dissipation emerged as an important concern in circuit design during the last decade due to heating problems in high-density/high-performance circuits and to power savings required for portable applications. A growing fraction of the power consumed by very large scale integration (VLSI) circuits is due to the clock distribution network, signal repeaters, I/O drivers, and busses all based on inverters. Moreover, the analytical description of the power dissipated in a CMOS inverter is the most important step in the description of more complex gates [3].

It is well known that power dissipation in CMOS circuits has a dynamic and a static component. The dynamic dissipation is due to the charge/discharge of the gate output load and to the short-circuit current due to the direct supply-ground conducting path created during the transition [4]–[6]. Many high-level approaches consider only the first contribution to compute power in large circuits [7], [8], given the complexity and computing cost of considering the short-circuit component.

The static dissipation is mainly due to current leakage from transistors in the off state, although reverse biased diodes from the device diffusions and well regions also contribute to this current. The off-state current, referred to as subthreshold leakage, is becoming more important with technology scaling since the device threshold voltage \( V_{\text{TH}} \) is reduced to maintain circuit performance. This term is usually neglected since \( V_{\text{TH}} \approx (kT/q) \) \( (k \) being the Boltzmann constant, \( T \) the temperature in Kelvin, and \( q \) the electron charge) [9].

Several works model the short-circuit power consumption. Veendrick [4] obtained an expression for unloaded buffers although the short-circuit component has a strong dependence on the output capacitance. Hedenstierna et al. [5] calculated the short circuit current using the Shockley MOSFET model for loaded buffers. This model cannot be applied to short-channel devices since they exhibit a linear dependence of saturation current versus the gate voltage, rather than the traditional square-law model used for long-channel devices [6]. Sakurai et al. [6] derived a model for long and short channel devices using their \( \alpha \)-power law MOSFET model for unloaded buffers. This model, as the model in [4], is useful to estimate the short-circuit component upper bound but cannot be used for a detailed power dissipation description since the output load effect must be included. Turgis et al. [10] derived an expression
for the short-circuit dissipation assuming that carriers were always moving at the saturation velocity and took into account overshooting effects. This model is valid only for deep submicron technologies and introduces the concept of equivalent capacitance, thus allowing a direct and frequency-independent comparison of the various power components. Hamoui et al. [11] obtained an expression for the short-circuit power dissipation using a modified version of the \( n \)-th power law MOSFET model in [12]. This approach requires numerical computing in a three-step process to determine the time when the short-circuiting transistor changes its mode of operation. Nikolaidis et al. [13] obtained an expression of the short-circuit dissipation for buffers driving long interconnect lines using the \( \pi \)-model of an RC load and the \( \pi \)-power law model for submicron devices. They took into account the input–output capacitance effects while the short-circuit current contribution was neglected when computing the output waveform. Recently, Nose et al. [14] derived a closed-form expression for the short-circuit power dissipation of CMOS gates taking into account short-channel effects, but not overshooting effects of increasing importance in submicron CMOS circuits. The work concluded that the short-circuit power to total power ratio \( (P_{sc}/P_{total}) \) will not change with scaling if the ratio \( V_{TH}/V_{DD} \) is kept constant.

In this work, we propose and evaluate a model to compute accurately the power consumption of CMOS buffers accounting for the main effects in submicron technologies as the input–output coupling capacitance and carriers velocity saturation effects avoiding time-consuming numerical procedures. The energy is computed from a detailed description of the various charge transference mechanisms involved during transitions. The final expression is the sum of the power dissipated when charging/discharging the output capacitance plus the short-circuit component that is accurately computed describing the impact of overshooting effects. The modeling process is described in detail providing a deep insight into the energy mechanisms involved as well as their interrelationship. The model is compared to HSPICE simulations (level 50) and to other models previously published considering a large set of parameters for a 0.35- and 0.18-\textmu m technology, showing significant improvements over previous works.

The paper is organized as follows: in Section II we describe the energy components and the method used to derive each contribution. Section III presents the MOSFET models used, while Sections IV and V derive the short-circuit and transient energy components, respectively. Section VI presents the results and Section VII concludes the work.

II. ENERGY COMPONENTS

We derive the energy consumption of a CMOS buffer (Fig. 1) describing the charge transfer mechanisms in the circuit when an input ramp is applied. The dynamic behavior of the circuit in Fig. 1 is described by

\[
(C_L + C_M) \frac{dV_{out}}{dt} = I_p - I_n + C_M \frac{dV_{in}}{dt}
\]

where \( C_L \) is the output load that includes the gate capacitances driven by the buffer, the interconnect wiring capacitance, and the diffusion capacitances of the buffer. \( C_M \) is the input–output capacitance, \( V_{out} \) the output voltage, \( I_p \) and \( I_n \) the pMOS and nMOS transistors current, respectively, and \( V_{in} \) the input voltage.

From this picture it is clear that the overshoot and short-circuit current components are related and their relative contribution is determined by the input transition time. If the input voltage is below \( V_{DD} + V_{thp} \) (the input voltage value at which the pMOS
turns off; \( V_{thp} \) is the pMOS threshold voltage) at \( t = t_{ov} \), then there will be a short-circuit current period from this time until the pMOS is turned off. This case will be referred to as a slow input transition. In the case of a fast input transition the output voltage is still beyond \( V_{DD} \) when the input reaches \( V_{DD} + V_{thp} \). In this case, \( t_{ov} > t_{n} + T_{SC} \) (\( t_{n} \) being the time at which the nMOS transistor starts to conduct and \( T_{SC} \) is the short-circuit current time interval when there is no overshooting; see Fig. 2) and there is no short-circuit period. For a fixed input transition time, the greater the overshoot time (i.e., the greater the input-output coupling capacitance), the shorter is the duration of the short-circuit current contribution. The overshoot time depends on the relationship between the input–output capacitance (that drives the output voltage beyond \( V_{DD} \)) and the driving strength of the nMOS transistor (that pulls this output voltage down), while the short-circuit period termination is determined uniquely by the input transition time. This effect is illustrated in Fig. 3, obtained from simulation, showing a set of pMOS current curves for a fixed input rise time and different coupling capacitance values. The short-circuit current contribution tail end is independent of the time at which short-circuit starts (since it is determined by the time at which the input voltage is \( V_{DD} + V_{thp} \)), while its negative slope does not vary significantly in all cases because the pMOS transistor is saturated. Note that overshooting does not only impact short-circuit duration, but also decreases its maximum value.

We included the overshooting effects on the short-circuit expression through the overshoot time using a geometrical approach following the behavior reflected in Fig. 3. This process is detailed in Section IV.

The short-circuit energy (\( E_{SC} \)) will be computed from the pMOS current since this component can be clearly identified. The energy associated with the overshoot charge transferred will be neglected since this energy is proportional not only to the small amount of charge involved, but also on the voltage difference through which it flows, which is also very small when compared to the supply voltage.

The transient energy (\( E_{TR} \)) corresponds to the energy dissipated when discharging the output capacitance. The energy of a discharging constant capacitance is simply \( E = QV/2 \), where \( V \) is the voltage swing and \( Q \) the charge initially stored in the capacitance and discharged through the nMOS transistor (defined as dynamic charge). The voltage swing of the output node in the buffer circuit is not the supply voltage \( V_{DD} \) but a higher one, say \( V_{DD} + \delta V \), due to overshooting. The dynamic charge will be derived using charge conservation by computing the charge at the beginning of the transition minus the charge at its end. It is important to remark that the overshoot charge (that is the charge transferred from the output node to the supply rail) must be subtracted from this term since it flows through a voltage difference \( \delta V \) and not through \( V_{DD} + \delta V \) as occurs for the charge flowing from the output node to ground. Therefore, although the overshoot charge is small and its overall energy contribution can be neglected, this charge must be subtracted from the charge variation at the output node since only the dynamic charge is multiplied by a large voltage difference \( V_{DD} + \delta V \) when computing the transient energy.

The output node charge is stored in both the input–output capacitance \( C_{M} \) and the load capacitance \( C_{L} \). Since the transient energy is derived from the computation of the charge at the beginning and the end of the transition, we require an expression of the input–output capacitance (which is strongly dependent on the input–output voltage) only for these two static operating conditions. The value of \( C_{M} \) in the static input low state (\( C_{M}^{L} \)) considering the side-wall capacitance of both transistor drains and the gate to drain overlap capacitance of the pMOS transistor in the linear region is given by

\[
C_{M}^{L} = C_{ox} \left( \frac{W_{p} L_{peff}}{2} + L_{AP} W_{peff} + L_{AN} W_{reff} \right)
\]

with \( L_{AP} \) and \( L_{AN} \) being the gate source/drain underdiffusion for the pMOS and nMOS transistors, respectively. \( W_{peff} \) and \( W_{reff} \) are the nMOS and pMOS effective channel width, while \( L_{peff} \) is the effective channel length of the pMOS transistor. The effective channel length and width are [2]

\[
W_{eff} = W + W_{VAR} - 2W_{OT}
L_{eff} = L + L_{VAR} - 2L_{AD}
\]
where $W_{\text{VAR}}$ and $L_{\text{VAR}}$ are the process bias on the channel width and length, respectively, while $W_{\text{ISOL}}$ is the isolation reduction of channel width. For a static input high the capacitance $C_{M}^{H}$ can be obtained similarly.

III. MOSFET MODEL USED

Equation (1) cannot be solved in a closed form even when the simple Shockley MOSFET model is considered. Moreover, since carrier saturation effects become important with technology scaling, more advanced MOSFET models accounting for such effects must be considered.

The alpha-power law MOSFET model [16] is a simple short-channel MOSFET model expressed as

$$I_{DS} = \begin{cases} 0 & (V_{GS} \leq V_{TH}) \\ (2 - \frac{V_{DS}}{V_{DD}})^{\frac{V_{DS}}{V_{DD}}} & (V_{DS} < V_{D}'_{DO}) \\ I_{D}'_{DO} & (V_{DS} \geq V_{D}'_{DO}) \end{cases}$$

with

$$I_{D}'_{DO} = I_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^{\alpha}.$$  (7)

The saturation voltage $V_{D}'_{DO}$ is given by [16]

$$V_{D}'_{DO} = V_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^{\alpha}.$$  (8)

We took $n = 1$ instead of the traditional $\alpha/2$ dependence of the saturation voltage [6] to simplify Sakurai’s equations in the linear region. A comparison of (8) with the physically based saturation voltage for an nMOS transistor with dimensions $W_{n} = 5 \mu m$ and $L_{n} = 0.35 \mu m$ is plotted in Fig. 4.

The parameter $\alpha$ is the velocity saturation index that takes a value between two (long-channel devices) and one (short-channel). The two parameters $I_{D0}$ and $V_{D0}$ are the drain current and the saturation voltage for $V_{GS} = V_{DS} = V_{DD}$, while parameters $V_{TH}$ and $\alpha$ are fitting parameters [6].

$I_{D0}$ and $V_{D0}$ can be computed using any physically based MOSFET model. We used the MOSFET model 9 (MM9) to compute these parameters (see [1] and [2] for a detailed description of the MM9 structure). The MM9 is a short-channel model that is being included in many widely used general-purpose circuit simulators and thus is drawing a significant amount of attention. A relatively small number of model equations and parameters are used when compared with many other models [2]. MM9 basic equation is

$$I_{DS} = \beta \left( \frac{V_{GS} - V_{th}}{V_{DS}} - \frac{1}{2} \frac{V_{DS}^{2}}{V_{DD}^{2}} \right) \frac{1}{1 + \theta_{1} (V_{GS} - V_{th}) (1 + \theta_{3} V_{DS})}.$$  (9)

where $V_{DS1}$ is a smoothing function being equal to the drain to source voltage when the transistor is in the linear region and to the saturation voltage when it is saturated. $V_{GS}$ is the gate to source voltage and $V_{th}$ is the threshold voltage. $\beta = \mu_{0} C_{ox} (W_{eff}/L_{eff})$ with $C_{ox}$ being the gate oxide capacitance,

$\mu_{0}$ the carriers mobility, while $W_{eff}$ and $L_{eff}$ are the effective channel width and length, respectively. The parameters $\theta_{1}$ and $\theta_{3}$ are the gate field and drain bias mobility reduction coefficients and $\delta$ is given by

$$\delta = \frac{\lambda_{1}}{(PHIB)^{1/2}} \left[ \frac{(K + \frac{V_{SBX}}{V_{DD}})^{2}}{V_{SBX}^{2} + (\lambda_{0} (V_{GS} - V_{th}))^{2}} \right].$$  (10)

where $KO$ is the substrate sensitivity when depleting surface doping and $K$ is the substrate sensitivity when depleting bulk doping. $V_{SBX}$ is a voltage of transition between $KO$ and $K$. $PHIB$ is the surface potential for strong inversion while $\lambda_{1} = 0.3$ and $\lambda_{0} = 0.1$ are MM9 constants from [1] and [2]. A detailed description of parameter extraction for this model from the device I-V current characteristics can be found in [17].

The value of $I_{D0}$ in (7) is computed using (9) while $V_{D0}$ in (8) is obtained from [2]

$$V_{D0} = \frac{2(V_{DD} - V_{th})}{(1 + \delta) \left( 1 + \sqrt{1 + \frac{2\theta_{1}(V_{DD} - V_{th})}{(1+\delta)}} \right)}. $$  (11)

Alpha-power law model equations are mathematically simpler than physically based MOSFET models like MM9 with the disadvantage that the relationship between the empirical parameters $\alpha$, $V_{TH}$ and the foundry supplied process parameters is not developed directly. We use the following relationship [18] to relate $\alpha$ to physical parameters:

$$\alpha = 1 + \frac{\mu_{\text{eff}}}{\mu_{0}} \left. \frac{V_{GS}=V_{DS}=V_{DD}}{V_{GS}=V_{DS}=V_{DD}} \right. \cdot $$  (12)

The expression for $\mu_{\text{eff}}$ in the MOSFET Model 9 is [2]:

$$\mu_{\text{eff}} = \frac{\mu_{0}}{(1 + \theta_{1} (V_{GS} - V_{th}) (1 + \theta_{3} V_{DS})}. $$  (13)

To obtain a better fitting of the transistor characteristics we derive a value of $V_{TH}$ by adjusting this threshold voltage to fit the drain saturation current at $V_{GS} = V_{DD}/3$. We obtain this...
current value (denoted as $I_{DO3}$) from (9) and equate it to the predicted value of the alpha-power law

$$I_{DO3} = I_{D0} \left( \frac{V_{DD} - V_{TH}}{3} \right)^{\alpha}$$

(14)

with $\alpha$ given by (12). We use (14) to get $V_{TH}$ from the predicted $I_{DO3}$ value leading to

$$V_{TH} = \frac{\left( \frac{I_{D0}}{I_{DO3}} \right)^{1/\alpha} - V_{DD}}{1 - \left( \frac{I_{D0}}{I_{DO3}} \right)^{1/\alpha}}.$$  \hspace{1cm} (15)

The parameters $I_{D0}$, $I_{DO3}$, $\mu_0$, and $\mu_{eff}(V_{GS} = V_{DD})$ can be obtained using any physically based MOSFET model since $I_{D0} = I_{DS}(V_{GS} = V_{DD}, V_{DS} = V_{DD})$ and $I_{DO3} = I_{DS}(V_{GS} = V_{DD}/3, V_{DS} = V_{DD})$.

Equations (12) and (15) relate $\alpha$ and $V_{TH}$ analytically to physical parameters without the need for time-consuming numerical computations.

IV. SHORT-CIRCUIT COMPONENT

We derive the short-circuit charge in a first step, neglecting overshooting effects which are included in a second step through the overshoot time.

In the case of unloaded buffers ($C_L = 0$), the short-circuit current is equal to the nMOS saturation current for the interval $t_n < t < t_{max}$ and to the pMOS saturation current when $t_{max} < t < t_n + T_0^{sc}$, where $t_{max}$ is the time at which the short-circuit current is maximum. Therefore, using the alpha-power law MOSFET model, the short-circuit current for an unloaded buffer is

$$I_{sc} = \begin{cases} 
I_{D0} \left( \frac{V_{DD} - t - V_{TH}}{V_{DD} - V_{TH}} \right)^{\alpha_n} & (t_n < t \leq t_{max}) \\
I_{D0} \left( \frac{V_{DD} - V_{TH}}{t_{max} - V_{TH}} \right)^{\alpha_p} & (t_{max} < t < t_n + T_0^{sc})
\end{cases}$$

(16)

This current dependence is taken as a reference to obtain an expression for the short-circuit current for loaded buffers using the values of the maximum current ($I_{max}$) and the time at which it occurs ($t_{max}$) derived in Appendix A. $I_{max}$ and $t_{max}$ have been derived considering the limit of unloaded buffers (i.e., parameters $I_{max}^{C_L=0}$ and $t_{max}^{C_L=0}$) and heavily loaded buffers (referred to as $I_{max}^{C_L=\infty}$ and $t_{max}^{C_L=\infty}$).

As an analogy to (16), the short-circuit current is shown in (17) at the bottom of the page.

The short-circuit charge transferred (SCCT) is the integral of $I_{sc}$ from the time at which the nMOS is turned on, $t_n$, up to the time at which the pMOS turns off, $t_n + T_0^{sc}$ (see Fig. 2)

$$Q_0^{sc} = \int_{t_n}^{t_n + T_0^{sc}} I_{sc} dt.$$  \hspace{1cm} (18)

That leads to the following:

$$Q_0^{sc} = mI_{max} \left( \frac{t_{max} - t_n + T_0^{sc} - t_{max} + t_n}{\alpha_n + 1} + \frac{t_{max} - t_{max} + t_n}{\alpha_p + 1} \right).$$  \hspace{1cm} (19)

The value of $m$ is selected to set the value of (19) equal to the HSPICE simulation of $Q_0^{sc}$. This parameter is required to adjust the charge and accounts for deviations mainly due to differences between $V_{TH}$ (from the $\alpha$-power model) and the physical $V_{th}$, as well as other second-order effects not considered similar to channel length modulation.

This value has been found to be close to 1.5 and 1.7 for 0.35- and 0.18-µm technologies, respectively, for a large range of parameters values, including variation of the channel length (and therefore $\alpha$), the buffer symmetry (which has a strong effect on the short-circuit current shape), input rise/fall time, output capacitance, input–output capacitance, etc. The value of $m$ is the same for rising or falling input transitions and must be adjusted only once for the technology since its value remains constant for all the parameters considered in the model.

Fig. 5 compares (17) to HSPICE simulations for various values of the output capacitance (from $C_{min}$ to 50$C_{min}$). When $C_L$ is small, the maximum current time is close to $t_{max}^{C_L=0}$ while HSPICE results are close to the current shape predicted by (16). As the output capacitance increases, the time at which the short-circuit current is maximum also increases and is close to $t_{max}^{C_L=\infty}$ for large values of $C_L$. The areas under the curves described by (17) and HSPICE simulations (i.e., the charge) are nearly the same.

Overshooting effects are included through the overshoot time $t_{ov}$. Simulations in Fig. 3 showed that overshooting displaces the short-circuit current to the right, maintaining the current slopes invariant. This can be described analytically using a geometrical approach as shown in Fig. 6. The short-circuit current curve displacement is approximated with straight lines of constant slope. $Q_0^{sc}$ is the area of the triangle $ABC$, while the reduced SCCT due to overshooting ($Q_0^{sc}$) is the area within the triangle $DEC$. With this geometrical analogy, $Q_0^{sc}$ is derived from $Q_0^{sc}$ using $t_{ov}$ as

$$Q_0^{sc} = Q_0^{sc} \left( 1 - \frac{t_{ov} - t_n}{T_0^{sc}} \right)^2$$  \hspace{1cm} (20)

with $Q_0^{sc}$ given by (19). Equation (20) is an approximation obtained assuming that the short-circuit current is linear with time.

$$I_{sc} = \begin{cases} 
mI_{max} \left( \frac{t - V_{TH}}{V_{DD} - t - V_{TH}} \right)^{\alpha_n} & (t_n < t \leq t_{max}) \\
mI_{max} \left( \frac{V_{DD} - V_{TH}}{t_{max} - V_{TH}} \right)^{\alpha_p} & (t_{max} < t < t_n + T_0^{sc})
\end{cases}$$

(17)
and it does not vanish for $t_{ov} > t_n + T_{sc}$. A more accurate relation for any $t_{ov}$ is

$$Q_{sc}^r = Q_{sc}^0 e^{-(2(t_{ov} - t_n)/T_{sc})^2}. \tag{21}$$

The mean deviation of (21) with respect to (20) is less than 1% of $Q_{sc}^0$ and provides negligible values for $t_{ov} > t_n + T_{sc}$.

We plot a comparison of (21) to HSPICE in Fig. 7 showing SCCT versus $C_M$ for different values of $C_L$, with good accuracy. The mathematical derivation of $t_{ov}$ is given in Appendix B. Finally, the energy associated with SCCT for a rising input is computed as

$$E_{sc}^r = Q_{sc}^r V_{DD}. \tag{22}$$

V. TRANSIENT DISSIPATION

The transient energy is dissipated when the output node is discharged. Conceptually this contribution can be subtracted from the short-circuit energy given that only the short-circuit charge flows through the short-circuiting transistor (that is the pMOS (nMOS) transistor for a low to high (high to low) input transition). In the buffer model of Fig. 1 the charge at the output node is stored in both the output and coupling capacitances, leading to $Q_{ov} = (C_M + C_L)V_{out} - C_M V_{in}$.

As discussed in Section II, overshooting has two main effects: first, it raises the voltage from which the output is discharged beyond $V_{DD}$ (say $V_{DD} + \delta V$), and second, part of the charge initially stored at the output node is transferred to $V_{DD}$ through the pMOS (this charge was referred to as the overshoot charge $Q_{ov}$). The overshoot charge does not contribute to the transient energy in the same way as the charge that goes to ground since the overshoot charge flows through a smaller voltage drop $\delta V$ (see Fig. 2). Therefore, the energy expression associated with the discharge of the output node must contain only a fraction of the charge initially stored in the output capacitances. The energy dissipated by the nMOS transistor for a high-to-low output transition is given by $E_{t2} = (Q_{ov} + \Delta Q_{ov} V_{ov}/2$ (energy dissipated for a constant discharging capacitance). $\Delta Q_{ov}$ is the charge transferred from the output node to ground through the nMOS transistor and $\Delta V_{ov}$ is the voltage swing given by $V_{DD} + \delta V_{tov}$. Thus, the discharging energy is expressed as

$$E_{ov} = \left( Q_{ov} \left( t_n \right) - Q_{ov} \left( \infty \right) \right) \frac{V_{DD} + \delta V_{tov}}{2} \tag{23}$$

where $Q_{ov} \left( t_n \right)$ is the charge at the output node at time $t_n$ (when the nMOS transistor starts to conduct), $Q_{ov} \left( \infty \right)$ is the charge stored at the output node when the transition is finished, and $Q_{ov}$ is the overshoot charge transferred from the output node capacitances through the pMOS transistor (and therefore not dissipated in the nMOS transistor) during time that both transistors are conducting ($t > t_n$). The output voltage swing is $V_{DD} + \delta V_{tov}$.

We write $Q_{ov} \left( t_n \right) = (C_L + C_M) V_{DD} - Q_{ov}$, considering that the charge for $t = t_n$ at the output node is equal to the charge at that node at the beginning of the input transition ($Q_{ov} \left( t = 0 \right) = (C_L + C_M) V_{DD}$) minus the charge that has been transferred from the output to $V_{DD}$ through the pMOS transistor until the nMOS transistor starts to conduct (defined
as \( Q_{\text{ov}}(x) \). Since \( Q_{\text{ov}}(\infty) = -C_M V_D \), and the total overshoot charge transferred is \( Q_{\text{ov}} = Q_{\text{ov}}^S + Q_{\text{ov}}^F \), (23) takes the form

\[
E_{\text{ov}} = \frac{1}{2} \left( (C_L + C_M^I + C_M^H) V_D - Q_{\text{ov}}^F \right) (V_D + \delta V_{\text{max}}).
\]

(24)

If the parasitic input–output capacitance is neglected \((C_M \approx 0)\), the overshooting and overvoltage vanish \((Q_{\text{ov}}(C_M = 0) = 0\) and \(\delta V_{\text{max}}(C_M = 0) = 0\) and the well-known expression for the transient energy is obtained as

\[
E_{\text{ov}} = \frac{1}{2} C_L V_D^2.
\]

(25)

To compute the transient energy in (24) we derive an expression for the overshoot charge \(Q_{\text{ov}}\) and the overvoltage \(\delta V_{\text{max}}\).

A. Overshoot Charge Transfer

The overshoot charge transfer differs for fast and slow input transitions since in the first case the pMOS transistor is off when overshooting ceases and in the second case this device is still conducting. We compute the charge for the slow and fast input transitions \((Q_{\text{ov}}^S\) and \(Q_{\text{ov}}^F\), respectively) and then combine both equations in a unified expression.

1) Slow Input: In this case \(t_{\text{ov}} < t_n + T_{\text{sc}}^0\), i.e., the input \(V_{\text{in}}\) has not reached \(V_D + V_{TP}\) at \(t = t_{\text{ov}}\). The charge stored at the output node at time \(t = t_{\text{ov}}\) is given by \(Q_{\text{ov}}(t_{\text{ov}}) = (C_L + C_M^I) V_D - C_M^I V_{\text{in}}\) while at the beginning of the transition the charge stored at the output node is \(Q_{\text{ov}}(0) = (C_L + C_M^I) V_D\). The difference \(Q_{\text{ov}}(0) - Q_{\text{ov}}(t_{\text{ov}})\) is the overshoot charge plus the charge that passed through the nMOS until \(t = t_{\text{ov}}\). Therefore, the overshoot charge expression is

\[
Q_{\text{ov}}^S = C_M^I \left( V_{\text{TN}} + \frac{t_{\text{ov}} - t_n}{T_{\text{sc}}^0} V_{\text{sc}} \right) - \int_{t_{\text{ov}}}^{t_n} I_{\text{ov}} dt.
\]

(26)

For a slow input transition \(t_{\text{ov}}\) is close to \(t_n\), therefore the approximation \(C_M \approx C_M^I\) is still valid and the integral over the nMOS transistor current can be neglected. The term into the parenthesis of (26) corresponds to the input voltage swing that determines the amount of charge flowing through the pMOS transistor during overshooting and cannot go beyond \(V_D + V_{TP}\) since for these values the pMOS is off. The time \(t_{\text{ov}}\) can go beyond \(t_n + T_{\text{sc}}^0\) leading to an overestimation in the input voltage swing. Therefore, we correct the expression of \(t_{\text{ov}}\) in (26) with a time \(t'_{\text{ov}}\)

\[
Q_{\text{ov}}^S = C_M^I \left( V_{\text{TN}} + \frac{t'_{\text{ov}} - t_n}{T_{\text{sc}}^0} V_{\text{sc}} \right).
\]

(27)

Knowing that \(t'_{\text{ov}}\) must be equal to the overshoot time \(t_{\text{ov}}\) for small values and that saturates to \(t_n + T_{\text{sc}}^0\) for large ones, we use the following expression:

\[
t'_{\text{ov}} = t_n + \text{lzp}(t_{\text{ov}} - t_n, T_{\text{sc}}^0, 0.1 T_{\text{sc}}^0)
\]

(28)

where \(\text{lzp}\) is a function that saturates to \(T_{\text{sc}}^0\) when \(t_{\text{ov}} - t_n > T_{\text{sc}}^0\) and is equal to \(t_{\text{ov}} - t_n\) when \(t_{\text{ov}} < t_n + T_{\text{sc}}^0\) defined as

\[
\text{lzp}(x, x_0, \varepsilon) = x_0 + \text{hyp}(x_0 - x - \frac{\varepsilon}{x_0}, \varepsilon)
\]

(29)

with

\[
\text{hyp}(x, \varepsilon) = \frac{1}{2} \left( x + \sqrt{x^2 + 4\varepsilon^2} \right)
\]

(30)

where the parameter \(\varepsilon\) defines the transition between the two regions.

2) Fast Input: For a high-speed transition the pMOS is off before \(t_{\text{ov}}\) (i.e., \(t_{\text{ov}} > t_n + T_{\text{sc}}^0\)) and the overshoot charge is calculated from the current passed through the pMOS. We use (9) neglecting the velocity saturation effect of the gate voltage \((V_{GP} = 0)\) with neither of the quadratic or higher order terms of \(V_{DS}\) (the drain voltage of the pMOS during overshooting is small) leading to

\[
I_p \approx \beta_p (V_{GS} + V_{th_p}) V_{DS} + \alpha V_D^2.
\]

(31)

Given that \(V_{GS} + V_{th_p} = V_{sc}(1 - (t - t_n)/T_{sc}^0)\), we obtain

\[
Q_{\text{ov}}^F = \int_0^{T_{sc}^0 + t_n} I_p dt \approx \int_0^{T_{sc}^0 + t_n} \beta_p V_{sc} \left( 1 - \frac{t - t_n}{T_{sc}^0} \right) V_{DS} dt.
\]

(32)

At the beginning of the transition \(V_{DS}\) is computed from (1) neglecting the nMOS and the pMOS current \((I_p = I_n = 0)\), which are negligible with respect to the current injected through \(C_M\) that is proportional to \(\frac{dV_{an}}{dt}\) (which is large for fast inputs). So in this case (1) simplifies to

\[
dV_{an} = \frac{C_M}{C_L + C_M} \frac{dV_{an}}{dt}
\]

(33)

from (33), the pMOS drain voltage takes the form \(V_{DS} = V_{sc}C_M^I/(C_M^I T_{sc}^0)\), thus (32) leads to

\[
Q_{\text{ov}}^F = \frac{\beta_p (V_{sc} + V_{TN})^2 C_M^I T_{sc}^0}{6V_{sc} C_M^I}.
\]

(34)

3) Combining Expressions for \(Q_{\text{ov}}^S\) and \(Q_{\text{ov}}^F\): Equations (27) and (34) are unified into a single analytical expression for the overshoot charge that is equal to \(Q_{\text{ov}}^F\) for high-speed transitions and to \(Q_{\text{ov}}^S\) for slow inputs

\[
Q_{\text{ov}} = Q_{\text{ov}}^F + Q_{\text{ov}}^S.
\]

(35)

Fig. 8 shows the charge transferred through the pMOS transistor for a high-to-low output transition as a function of \(t_{\text{an}}\) for different nMOS channel widths. Squares are HSPICE simulations while solid lines correspond to the charge model developed \((Q_{\text{an}}^S - Q_{\text{an}}^F)\). The charge transferred for high-speed transitions is mostly due to overshooting (a total negative charge is obtained) while for slow-input transitions the short-circuit component is dominant (overall positive charge). The curves in Fig. 8 show a good fit to (35) and (21) to describe the charge through the pMOS.

B. Overvoltage Evaluation

We obtain the maximum overvoltage \(\delta V_{\text{max}}\) computing the output voltage at \(t = t_n\). We adjust this expression to meet the limit of an ideal step input using charge conservation.
The derivation of $V_{\text{out}}(t_n)$ requires solving

$$C_L \frac{dV_{\text{out}}(t)}{dt} = -I_p + C_M T_0 \frac{V_{\text{sc}}}{T_0}.$$ \hspace{1cm} (36)

Equation (36) is derived from (1) and is valid for $t < t_m$ where $I_m$ is neglected since the nMOS is off. To get an analytical solution of (36) and compute $V_{\text{out}}(t_n)$, we take the pMOS current expression with the approximations of small drain voltage and neglect gate saturation effects (31). We also take an average value of $V_{GS} + V_{thp}$ as $V_{sc}/2$ leading to

$$I_p \approx \frac{1}{2} \beta_p V_{sc} (V_{\text{out}}(t) - V_{DD}).$$ \hspace{1cm} (37)

We solve (36) using (37) and compute the overvoltage when the nMOS transistor starts to conduct ($\delta V_{TH} = V_{\text{out}}(t_n) - V_{DD}$) obtaining

$$\delta V_{TH} = 2 \frac{C_L}{\beta_p T_{sc}} \left(1 - e^{(-\beta_p V_{TN} T_0^2)/(\sqrt{C_L})}\right).$$ \hspace{1cm} (38)

Equation (38) provides a partial description of the maximum output voltage.

For an ideal step input, the maximum overvoltage is obtained from charge conservation as

$$\delta V_{\max} = V_{DD} \frac{C_M}{C_L}.$$ \hspace{1cm} (39)

This result is not provided by (38) when $T_{sc} = 0$. The maximum overvoltage is also determined by the discharging nMOS current and the capacitance being discharged ($\sqrt{C_L}$). To account for these effects we define an effective resistance for the nMOS transistor as $R_{n} = V_{DD0.1} / I_{DD0.1}$ and incorporate an additional term to (38) to estimate the maximum overvoltage beyond $t_n$

$$\delta V_{\max} = \delta V_{TH} + \left(C_M \frac{V_{DD} - V_{TN}}{C_L} e^{-(t_n - t_m)/(R_n C_L^2)}\right).$$ \hspace{1cm} (40)

Equation (40) is an empirical expression that provides an excellent agreement with HSPICE simulations as shown in Fig. 9. The plot presents the overvoltage versus the input rise time for different values of $C_L$ (from $C_{\min}$ to 50$C_{\min}$).

VI. RESULTS

We compare the model predictions versus HSPICE simulations for 0.35- and 0.18-$\mu$m technologies considering energy versus input transition time, output capacitance, channel length, supply voltage, temperature, and the pMOS to nMOS channel width ratio. For each plot we include results from the models in [10], [11], or [14], depending on which best fits the energy consumption for the parameter considered.

Fig. 10 plots the energy dissipated per cycle versus the input-to-output time ratio for the 0.18-$\mu$m technology. The output time is taken from HSPICE simulations as an average value of the rise and fall times at the buffer output ($t_f$), where $t_f$ and $t_r$ are proportional to the time from 0.9$V_{DD}$ at the output ($t_{90}$) to 0.1$V_{DD}$ at the output ($t_{10}$) (i.e., $t_{90} = t_{90} - t_{10}$). Three different design scenarios are considered: a minimum inverter driving a minimum sized inverter ($I_{\min} = W_p / W_n = 0.58 \mu$m/$0.28 \mu$m, $C_L = C_{\min}$) and two scaled inverters ($3I_{\min}$ and $5I_{\min}$) with output capacitances of $3C_{\min}$ and $5C_{\min}$, respectively. The input time ranges from 20 ps to 1 ns (0.5 to 3.5 in the plot) covering both fast
Fig. 11. Power dissipation versus input time to output time ratio for three different inverter sizes in 0.18-μm technology. Different \( W_p/W_n \) ratios are considered.

Fig. 12. Power dissipation versus output capacitance for three different input rise times for 0.18-μm process technology with \( W_n = 1 \mu m \) and \( W_p = 2 \mu m \). The model is compared to HSPICE simulations and a recently published model [14].

Fig. 13. Energy dissipation versus input-to-output time ratio for different \( W_p/W_n \) ratios for 0.35-μm technology. When \( t_{in}/t_{out} > 1 \) the short-circuit component is dominant while for \( t_{in}/t_{out} < 1 \) the overvoltage contribution increases the total power.

Fig. 14. Energy dissipation versus channel length for 0.35-μm technology. This graph shows the behavior of the proposed model when varying the velocity saturation index \( \alpha \) and the coupling capacitance.

and slow input transitions. Simulation results show that both models (the model proposed and the model in [10]) provide a good description of the energy for the whole design space considered.

Fig. 11 plots the power dissipation versus the input-to-output time ratio for five inverters, three with \( C_L = 5C_{\text{min}} \) and \( W_p/W_n \) ratios 3 \( \mu m/1.5 \mu m \), 2.25 \( \mu m/2.25 \mu m \) and 1.5 \( \mu m/3 \mu m \) and two with \( C_L = C_{\text{min}} \) and \( W_p/W_n \) ratios 2 \( \mu m/1 \mu m \) and 2 \( \mu m/2 \mu m \) for 0.18-μm technology. The input times range from 20 ps to 1 ns. The model presented reports better fitting with respect to previous works (we include results only from [14] as the model better describing the energy for this parameter) maintaining an accurate description of power for different values of \( W_p/W_n \).

Fig. 12 shows the power dissipation dependence with the output capacitance for three input rise time values (\( t_{in} = 20 \) ps, \( t_{in} = 0.5 \) ns, and \( t_{in} = 1 \) ns) for a 0.18-μm technology CMOS inverter with \( W_p/W_n = 2 \mu m/1 \mu m \). The model proposed and the model in [14] are compared to HSPICE simulations showing similar accuracies, thus verifying the accuracy of the proposed model to describe the short-circuit and transient components.

Fig. 13 plots the energy dissipated per one cycle period versus the input-to-output time ratio for 0.35-μm technology. Four different inverters are considered with different values of the \( W_p/W_n \) ratio and two values of the output capacitance (all the values specified in the graph). The input time ranges from 20 ps to 1.5 ns. Results show that the model presented describes correctly HSPICE simulations under all the parameter combinations considered representing an improvement over previous models.

Fig. 14 compares the energy versus channel length with HSPICE for different values of \( \alpha \) and a minimum channel width (\( W_n = 0.5 \mu m \)). This graph shows good accuracy of the model proposed for different velocity saturation index values \( \alpha \) since this parameter is strongly dependent on the channel length. As the device length decreases, the conductance of the transistors is reduced leading to lower values of SCCT; this is the case for \( t_{in} = 1 \) ns. The power contribution due to the coupling capacitance increases with channel length for shorter input transitions.
The energy variation with the power supply voltage is shown in Fig. 15 for a CMOS buffer with $W_p = 20 \mu m$, $W_n = 10 \mu m$, an output capacitance of $C_L = 50C_{\text{min}}$, and different input rise times (from 50 ps to 1 ns). HSPICE simulations are compared to the proposed model and the model in [11]. Energy models require good accuracy with supply voltage since it is one of the key parameters used to control power consumption. Simulations show that the model proposed represents a contribution in describing this dependence.

The energy versus power supply for 0.18-$\mu$m technology is shown in Fig. 16 for a CMOS buffer with $W_p/W_n = 10 \mu m/5 \mu m$, $C_L = 10C_{\text{min}}$, and $t_{in} = 50$ ps, 500 ps, and 1 ns. The model proposed provides a good description in all the range considered.

Fig. 17 plots the energy versus temperature for three different supply voltages. The MOSFET Model 9 temperature dependence [2] is included in the basic parameter set. The model accuracy is maintained over the whole period considered ($-50^\circ C$ to $150^\circ C$) showing that the low voltage regime is less sensitive to temperature variation. Fig. 18 plots the energy in one cycle versus the pMOS channel width for a fixed value of the nMOS channel width for 0.18-$\mu$m technology. Different values of the input times $t_{in}$ are considered (from 50 ps to 1 ns). Fig. 18 reflects the accuracy of the model to describe the time at which the maximum short-circuit current takes place as the maximum dc current location is dependent on the buffer symmetry.

Fig. 19 shows the energy dissipated by a CMOS buffer driving a long interconnect line. We considered a 1-$\mu$m width level 2 metal simulated using a π3 model [20] for 0.35-$\mu$m technology. For this technology the metal 2 to well capacitance per unit of area is $C_n = 11.7 \times 10^{-18} F/\mu m^2$, its fringing capacitance is $C_f = 0.036 F/\mu m$, and its square resistance is $R_s = 55$ m$\Omega$ sq. The total output capacitance (used in the model) was computed as $C_L = C_{\text{out}} + C_{\text{wire}} + C_{\text{in}}$ where $C_{\text{out}}$ is the output capacitance of the buffer, $C_{\text{wire}}$ is the parasitic interconnect capacitance, and $C_{\text{in}}$ is the input capacitance of the driven gate. We compared HSPICE simulations (squares) to the model developed and a specific model for long interconnect lines [13] for two different input times of 50 ps and 1 ns. The model derived in this work reports good fitting for interconnect lines lengths below 5 mm, leading to energy underestimation for higher wire lengths. This underestimation is
due to the increase of the line resistance with its length. When the line resistance becomes comparable to the buffer output resistance, part of the output capacitance is shielded from the buffer, thus decreasing the gate delay and increasing the short-circuit dissipation [21], [22]. We compare this result with the model developed by Nikolaidis et al. [13] (dashed line in Fig. 19). That model provides an accurate description for the delay of an inverter driving an RC line but uses a fixed value for the time at which maximum short-circuit takes place (set as $(t_{	ext{off}} + t_m + T_{\text{L}})/2$). Therefore, the dependence of this time with the output capacitance increase (Fig. 5) is not described, leading to a underestimation of the energy dissipated as shown in Fig. 19.

We calculated the mean deviation of the Energy versus input time, channel length, supply voltage, output capacitance, and pMOS channel width from HSPICE simulations for the models in [10], [11], and [14] and the model proposed for the 0.18- and 0.35-μm technologies considered. The percentage deviations are reported in Table I showing the improvement achieved by the model proposed with an overall accuracy within 5% of HSPICE simulations, thus representing an improvement over previous works. The model in [10] provides the better agreement with HSPICE simulations for the 0.18-μm technology rather than the 0.35-μm technology since it was developed for deep-submicron technologies.

Table II reports the percentage of energy due to coupling capacitance effects ($\%E_{\text{cc}}$) and short-circuit currents ($\%E_{\text{sc}}$) with respect to the total power. The ratio $V_{\text{TH}}/V_{\text{DD}}$ is kept constant for the two technologies considered and the input time is selected to be equal to the output time ($t_{\text{in}} = t_{\text{off}}$). The percentage of the short-circuit power is nearly the same for the two technologies (as predicted in [14]) while the influence of the coupling capacitance on the total power increases for the 0.18-μm technology.

Finally, we computed the power dissipation for $10^6$ transitions with HSPICE and the analytical models on a Pentium III processor-based computer. Simulation times are reported in Table III showing that the models in [5], [6], [10], [14], and the proposed model are up to three orders of magnitude faster than HSPICE, while the models in [11] and [13] are two orders of magnitude faster than the simulator. The numerical approaches used in [11] to obtain the time at which the maximum short-circuit current takes place (using the bisection method) and the complex equations in [13] to model the RC line at the output of the buffer increase computation considerably.

VII. CONCLUSION

We have developed an analytical expression for the energy dissipation in CMOS buffers by solving the dynamics of the circuit and considering a physically based $\alpha$-power law MOSFET model. The various energy components are discussed and modeled in detail highlighting the parameters involved and their interrelationship. Several nonlinear problems are solved with simple and accurate formulas to avoid time-consuming numerical procedures. Exhaustive comparisons to HSPICE results are provided to demonstrate the validity and accuracy of the model for both submicron (0.35 μm) and deep-submicron (0.18 μm) technologies. This model is expected to remain valid for scaled technologies as long as the subthreshold leakage component does not dominate active power, since the main submicron effects are accounted at the device level. It represents an improvement over previous models since results show an overall high accuracy in describing the energy for all the parameters considered (Figs. 10–19), while previous models maintain accuracy only for some parameters or technologies. The model proposed has an average error within 5% of HSPICE simulations for a wide range of parameter variation.
The model presented underestimates energy when the resistance of the line driven by the buffer is comparable to the effective resistance of the transistors in the buffer. Additional efforts are required to account for these resistive effects [22].

**APPENDIX A**

**COMPUTATION OF THE MAXIMUM SHORT-CIRCUIT CURRENT**

**Maximum Current for Unloaded Buffers**

When the output capacitance is small (i.e., when the short-circuit current has a greater impact [4]) the circuit behavior is close to the inverter dc operation in the sense that $I_D \approx I_n$. At the beginning of the transition, the pMOS transistor drives a current equal to the nMOS saturation current, while at the end of the transition the pMOS is saturated. In this particular case, the maximum current takes place when $I_{D_{00}} = I_{D_{0n}}$. Using (7) and (2) and using $V_{GS} = V_{in}$ for the nMOS and $V_{GS} = V_{DD} - V_{in}$ for the pMOS, $I_{D_{0p}} = I_{D_{0n}}$ leads to

$$I_{D_{0p}} = I_{D_{0n}} \left( \frac{V_{in}}{V_{DD} - V_{TN}} \right)^{\alpha_n} \left( \frac{V_{DD} + V_{TP}}{V_{SC}} \right)^{\alpha_p}$$

where $C_{max} = t_n + T_{SC} 2 - \frac{1}{2} (\alpha_n - \alpha_p - 1) (\alpha_n - \alpha_p - 1) + \alpha_p - \alpha_n - 1 + S_p F_p$.

Equation (A3) leads to the exact solution of (A1) for $1 \leq \alpha_p, \alpha_n \leq 2$ (which is the range for these parameters) as

$$I_{D_{0p}} = I_{D_{0n}} \left( \frac{V_{in}}{V_{DD} - V_{TN}} \right)^{\alpha_n} \left( \frac{V_{DD} + V_{TP}}{V_{SC}} \right)^{\alpha_p}$$

where

$$S_p = \frac{1 + \sqrt{1 + 4 F_p \alpha_p - \alpha_p - 1}}{1 + \sqrt{5 (\alpha_p - \alpha_n - 1) (\alpha_n - \alpha_p - 1)}}$$

Equation (A3) also gives an accurate description for the case $\alpha_n \neq \alpha_p$. A comparison of (A3) to HSPICE simulations is shown in Fig. 20 for various buffers with different values of the $(W_n/W_p)$ ratio for 0.35-μm technology. Once the maximum time is obtained we derive the maximum short-circuit current for unloaded buffers as $I_{D_{0p}} = I_{D_{0n}}$ from (7), using (2) and (A3)

$$I_{D_{0p}} = I_{D_{0n}} \left( \frac{V_{in} (t_n + T_{SC} 2 - \frac{1}{2} (\alpha_n - \alpha_p - 1) (\alpha_n - \alpha_p - 1) + \alpha_p - \alpha_n - 1 + S_p F_p)}{V_{DD} - V_{TN}} \right)^{\alpha_n}$$

**Computation of the Maximum Current for Heavily Loaded Buffers**

Equation (A3) is inaccurate for large $C_L$ values. In this case, the current through the pMOS transistor can be neglected and $C_M$ taken as $C_M \approx C_L^M$, while the input voltage is described with a linear ramp [see (2)].

Under these assumptions and neglecting overshooting effects (that will be included later), (1) becomes

$$V_{car}(t) = V_{DD} - V_D \left( t - t_n \right)^{\alpha_n+1}$$

where

$$V_D = \frac{I_{D_{0n}} T_{SC}^{\alpha_n} C_{MAS}^{\alpha_n+1}}{V_{DD} - V_{TN}^{\alpha_p}}$$

Equation (A7) is used for the linear expression of the pMOS current ($V_{DS} < V_{D_{0p}}$) in the alpha-power law MOSFET model (6) with

$$V_{DS} = V_D \left( \frac{t - t_n}{T_{SC}^{\alpha_n+1}} \right)$$

$$V_{GS} + V_{TP} = V_{SC} \left( 1 - \frac{t - t_n}{T_{SC}^{\alpha_p}} \right)$$

$$V_{D_{0p}} = \frac{V_{D_{0p}} V_{SC} \left( 1 - \frac{t - t_n}{T_{SC}^{\alpha_p}} \right)}{V_{DD} + V_{TP}}$$

that are obtained from (A7), (2), and (8), respectively, leading to

$$I_P = \left[ 2 - \frac{K_p (t - t_n)}{T_{SC}^{\alpha_n+1}} \right] - \frac{K_p I_{D_{0p}} (t - t_n)}{T_{SC}^{\alpha_p}}$$
The time at which the maximum current takes place for heavily loaded buffers $t_{\text{max}}^{C_L\rightarrow\infty}$ is obtained solving $\partial t I_p = 0$

$$t_{\text{max}}^{C_L\rightarrow\infty} = t_{\text{max}}^{C_L\rightarrow0} - T_{\text{sc}} A_p K_p + o\left(C_L^{-2}\right) \tag{A12}$$

where $t_{\text{max}}^{C_L\rightarrow\infty}$ and $A_p$ take the form

$$t_{\text{max}}^{C_L\rightarrow\infty} = t_n + T_{\text{sc}} \alpha_n + 1$$

$$A_p = \left(\frac{\alpha_n + 1}{\alpha_n + \alpha_p}\right)^{\alpha_p+2}, \frac{\alpha_p}{2(\alpha_p - 1)} \tag{A13}$$

Now we obtain the maximum short-circuit current for heavily loaded buffers computing (6) at $t = t_{\text{max}}^{C_L\rightarrow\infty}$ given by (A13). We take into account only the linear dependence of $I_p$ on $V_{GS}$ to simplify the expression. This approach is justified since the drain voltage swing during the input transition is not large for heavily loaded buffers. The result is

$$I_{\text{max}}^{C_L\rightarrow\infty} = 2K_p I_{\text{max}} \left(\frac{t_{\text{max}}^{C_L\rightarrow\infty} - t_n}{T_{\text{sc}}^0}\right)^{\alpha_n+1}$$

$$\cdot \left(1 - \frac{t_{\text{max}}^{C_L\rightarrow\infty} - t_n}{T_{\text{sc}}^0}\right)^{\alpha_p-1} \tag{A14}$$

Combining Both Expressions

Equations (A12) and (A14) diverge to $-\infty$ and $\infty$, respectively, for $C_L = 0$. We construct an expression for the maximum short-circuit current $I_{\text{max}}$ and time $t_{\text{max}}$ that leads to $I_{\text{max}}^{C_L\rightarrow\infty}$ and $t_{\text{max}}^{C_L\rightarrow\infty}$ for large values of $C_L$ and to $I_{\text{max}}^{C_L\rightarrow0}$ and $t_{\text{max}}^{C_L\rightarrow0}$ for $C_L = 0$.

The $I_{\text{max}}$ function must be a constant for small output loads (since $I_{\text{max}}^{C_L\rightarrow0}$ is load independent) while for large loads it must show the same dependence as $I_{\text{max}}^{C_L\rightarrow\infty}$, i.e., a function of the form

$$g(C_L) = \frac{A}{C_L} \tag{A15}$$

where $A$ is load-independent. An asymptotic function that follows the required limits is

$$f(C_L) = \frac{g(C_L)}{g(C_L) + g_0} \tag{A16}$$

where $g_0$ is the constant value of $I_{\text{max}}$ for $C_L = 0$. Therefore, the equivalent expression for $I_{\text{max}}$ in terms of $I_{\text{max}}^{C_L\rightarrow0}$ and $I_{\text{max}}^{C_L\rightarrow\infty}$ is

$$I_{\text{max}} = \frac{I_{\text{max}}^{C_L\rightarrow\infty} + I_{\text{max}}^{C_L\rightarrow0}}{I_{\text{max}}^{C_L\rightarrow\infty} + I_{\text{max}}^{C_L\rightarrow0}} \tag{A17}$$

The function for the time at which the short-circuit is maximum must lead to (A12) for heavily loaded buffers and to (A3) for unloaded buffers. Equation (A12) is a function of the form $h(C_L) = h_\infty - A/C_L$ (where both $h_\infty$ and $A$ are load-independent). We define the linear transformations $I$ and $T_{\text{sc}}$ (inversion and translation) as $I(f(x)) = -f(x)$ and $T_{\text{sc}}(f(x)) = t_n + f(x)$, and the linear transformation $L = \frac{1}{L_{\text{sc}}}$ (a translation followed by an inversion). With this transformation we obtain a function similar to (A15) from $h(C_L)$

$$g(C_L) = L h(C_L) \frac{A}{C_L} \tag{A18}$$

The same transformation $L_{\rightarrow\infty}$ is applied to (A3). We define $h_0 \equiv h_{\text{max}}$, $g_0 \equiv L_{\rightarrow\infty}(h_0) = h_\infty - h_0$ and then obtain a smoothing function $f(C_L)$ between $g(C_L)$ in (A18) and $g_0$ using (A16). Then we transform $f(C_L)$ with the inverse of
transformation $L_{h_{\infty}} \left( L_{-h_{\infty}}^{-1} \right) = T_{h_{\infty}} \cdot I$ to obtain a smoothing function between $h(C_L)$ and $h_0$ as

$$L_{h_{\infty}}^{-1} \left( f(C_L) \right) = h_{\infty} - \frac{A_{h_{\infty}}}{A_{h_0}} \left( h_{\infty} - h_0 \right).$$

(A19)

Given that $h_{\infty} = \frac{C_{l_{\text{max}}}}{C_{l_{\text{max}}}}$, $h_0 = \frac{C_{l_{\text{max}}}}{C_{l_{\text{max}}}}$ and $A/C_{l_{\text{max}}} = \frac{V_{DD}}{t_{\text{max}}} K_p$ for the maximum current time, (A19) leads to

$$t_{\text{max}} = \frac{C_{l_{\text{max}}} \left( 1 + \delta_0 \right)}{C_{l_{\text{max}}} + \frac{V_{DD}}{t_{\text{max}}} K_p}.$$  

(A20)

### APPENDIX B

OVERSHOOT TIME COMPUTATION

An accurate description of the overshoot time requires a detailed modeling of the device behavior; for this reason we use the MOSFET Model 9 (MM9) [1], [2] to calculate the overshoot time $t_{ov}$ used in (21) and (26).

The solution of (1) to get $t_{ov}$ is a nonlinear problem. To avoid numerical procedures we compute the time at which the overshoot current is at maximum and relate $t_{ov}$ to this value. Fig. 21 shows HSPICE simulations of $t_{ov}$ and $t_{ov}^{\text{max}}$ for a 0.18- and 0.35-$\mu$m process technology obtained from a minimum-sized inverter driving another minimum-sized inverter showing a linear relationship. The different simulations correspond to different input rise time values. This relationship can be expressed as

$$t_{ov} = t_n + \frac{3\left( t_{\text{max}}^{\text{ov}} - t_n \right)}{2}.$$

(B1)

We will compute $t_{ov}^{\text{max}}$ since its analytical derivation is simpler than the derivation of $t_{ov}$.

The maximum overshoot time is obtained from (1) neglecting the short-circuit current. Since at $t = t_{ov}^{\text{max}}$ the nMOS transistor is saturated and $dV_{\text{sat}}/dt = 0$, (1) is reduced to

$$C_M \frac{dV_{\text{sat}}}{dt} - I_{D0} = 0.$$

(B2)

Using the long-channel approximation of the saturation voltage $V_{\text{sat}} = (V_{GS} - V_{th})/(1 + \delta_0)$ [12], [19] and neglecting the product term $t_{\text{ov}}^{\text{max}}/\theta_{\text{ov}}$, in the denominator of (9) as a first approximation, the drain saturation current of the nMOS is

$$I_{D0} = \beta_n \frac{(V_{GS} - V_{th})^2}{2(1 + \delta_n) \left[ 1 + \left( \theta_{\text{ov}} + \theta_{\text{ov}} \right) \frac{(V_{GS} - V_{th})}{2} \right]}.$$

(B3)

Since $V_{GS} = \frac{V_{DD}}{t_{\text{max}}}$, we then have (B4), shown at the top of the page. The solution of (B4) is

$$t_{ov}^{\text{max}} = t_n + T_c + \sqrt{T_c^2 + \frac{2C_M \left( 1 + \delta_0 \right)}{\beta_n V_{DD}}}.$$

(B5)

with

$$T_c = \frac{C_M \left( 1 + \delta_0 \right)}{\beta_n}.$$

(B6)

Fig. 22 is a comparison of $t_{ov}$ with HSPICE for different values of $C_L$. It is shown that this time is almost independent of the load capacitance as predicted by (B5).

### ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their constructive comments.

### REFERENCES


**José Luis Rosselló** (M’02) received the M.S. and Ph.D. degrees in physics from the Balearic Islands University, Spain, in 1996 and 2002, respectively. He is a Teaching Assistant in the Physics Department, Balearic Islands University. His research interests include device and circuit modeling, CMOS IC testing, and low-power design.

Jaume Segura (M’94) received the M.S. and Ph.D. degrees from the Balearic Island University and the Polytechnic University of Catatonia in Spain, respectively.

Since 1993, he has been an Associate Professor in the Physics Department, Balearic Islands University (UIB), Palma de Mallorca, Spain. He was on leave from UIB in 1994 and 2001 as an Invited Researcher at Philips Semiconductors, USA, and Intel Corporation, respectively. His research interests include device and circuit modeling and VLSI design and testing.

Dr. Segura has served on the technical program committee of a number of conferences including the Design and Test in Europe (DATE), the International On-Line Testing Workshop (IOLTW), and the International workshop on current and Defect Based Testing (DBT). He is also member of the organizing committee of the VLSI Test Symposium (VTS) and served as General Co-Chair of the IOLTW in 2000.

José Luis Rosselló (M’02) received the M.S. and Ph.D. degrees in physics from the Balearic Islands University, Spain, in 1996 and 2002, respectively. He is a Teaching Assistant in the Physics Department, Balearic Islands University. His research interests include device and circuit modeling, CMOS IC testing, and low-power design.

Jaume Segura (M’94) received the M.S. and Ph.D. degrees from the Balearic Island University and the Polytechnic University of Catatonia in Spain, respectively.

Since 1993, he has been an Associate Professor in the Physics Department, Balearic Islands University (UIB), Palma de Mallorca, Spain. He was on leave from UIB in 1994 and 2001 as an Invited Researcher at Philips Semiconductors, USA, and Intel Corporation, respectively. His research interests include device and circuit modeling and VLSI design and testing.

Dr. Segura has served on the technical program committee of a number of conferences including the Design and Test in Europe (DATE), the International On-Line Testing Workshop (IOLTW), and the International workshop on current and Defect Based Testing (DBT). He is also member of the organizing committee of the VLSI Test Symposium (VTS) and served as General Co-Chair of the IOLTW in 2000.
Usefulness of antenna array model: With this antenna array model, array errors can be easily corrected by multiplying the transmit/receive data or weight vector by the inverse of the estimated error matrix. With error correction, all the elements in the array will have the same beam patterns, which is important for side-lobe reduction in the beamforming and downlink broadcasting channel which should be transmitted by a omni-directional beam pattern.

The antenna array model is very useful for a smart antenna system requiring array manifold for the AoA estimation, because the element patterns can be predicted from the array manifold measured at a small number of directions and so the measurement time will be significantly reduced in the mass production of the antenna array.

![Element beam patterns](image)

**Fig. 1. Element beam patterns**

a Before error correction  
b After error correction

**Experimental results:** Experiments using a uniform linear array with eight printed dipole elements were conducted to evaluate the model based antenna array calibration. The inter-element spacing is 5 cm and the aluminium ground plate is 40 cm wide and 24 cm high. No effort to reduce the edge scattering of the ground plate is made to see the effect of the modelling error. Fig. 1a shows the measured element patterns before error correction. Note that all the elements, in particular the edge elements, have unsymmetric patterns due to the edge scattering. The ripples in the patterns are mainly due to mutual coupling. The measured array manifold for angles equally spaced from $-60^\circ$ to $60^\circ$ with $10^\circ$ intervals was used to estimate the error matrix. Fig. 1b shows the element patterns after error correction. The correction was performed by multiplying the measured array manifold $B$ by the inverse of the estimated error matrix $M^{-1}$. Owing to modelling errors such as edge scattering, the element patterns are somewhat different from each other for angles with absolute values greater than $60^\circ$. In particular, the first and eighth elements have significantly different patterns, which is because edge scattering is more dominant for those elements. However, the pattern deviation is less than $0.5$ dB within one sector angle region ($-60^\circ$ to $60^\circ$) in which three sector based base station systems are interested.

Fig. 2 shows pattern prediction examples for the first and fourth antenna elements. The pattern prediction was performed by using the relationship in (1) in which the isolated element pattern $J(\theta)$ was assumed to be isotropic. When the isolated element pattern is not isotropic, it can be estimated by averaging the corrected element patterns, as shown in Fig. 1b. The prediction error for all elements in the array was less than $0.5$ dB within the sector angle region in the experiment.

![Element pattern prediction](image)

**Fig. 2. Element pattern prediction**

Conclusions: A model based antenna array calibration was experimentally evaluated to be applied to smart antenna systems. With this model based array calibration technique, all the elements can be made to have almost the same patterns. It was also shown that the element patterns can be predicted from the array response vectors measured at a small number of directions. The calibration error was less than $0.5$ dB within one sector angle region in the experiment.

**Acknowledgment:** The antenna array was built and measured by the ElectroScience Laboratory in the Ohio State University.

© IEE 2002 20 August 2001
Electronics Letters Online No: 20020585
DOI: 10.1049/el:20020585

Hyung Geun Park, Jae Ho Jung, Hyun Seo Oh and Mun Geon Kyeong (Smart Antenna Research Team, Radio & Broadcasting Technology Laboratory, Electronics and Telecommunications Research Institute (ETRI), 161 Gajeong-dong, Yuseong-Gu, Daejeon, 305–350, Korea)
E-mail: hgpark@etri.re.kr

**References**


**Simple and accurate propagation delay model for submicron CMOS gates based on charge analysis**

J.L. Rosselló and J. Segura

A simple method to evaluate the propagation delay of complex CMOS gates, based on inverter delay models and the $\alpha$-power law MOSFET model is presented. The method is based on a transistor collapsing technique developed for complex gates and takes into account short-channel effects, internal coupling capacitances and the body effect. The propagation delay of complex gates for a 0.18 µm technology is evaluated, showing excellent results.

**Introduction:** Several studies have focused on modelling CMOS gates delay [1–3]. Sakurai and Newton [1] developed a propagation delay model based on reducing each gate to a single equivalent inverter without considering the short-circuit current contribution. Hirata et al. [3] developed a propagation delay model for CMOS inverters considering short-circuit currents while complex gates were modelled using the collapsing technique developed in [1]. The relative
error of this method could be as high as 92% for a 2NAND gate for 0.8 μm devices as reported in [3]. In this Letter we present a simple method to compute the propagation delay of complex CMOS gates when one of the inputs changes and the remaining ones are static using inverter delay models. The method is applied to a previously published model for the inverter delay [3] and compared to HSPICE for a 0.18 μm technology showing good accuracy.

When \( V_{\text{DD}} \) makes a low to high transition the charge at node \( N-1 \) changes. Initially \( V_{DQ1}=0 \), \( V_{DQ2}=V_{\text{DD}}-V_{\text{Th}} \), and \( Q_{\text{DQ1}} = Q_{\text{DQ2}} = (V_{\text{DD}} - V_{\text{Th}}) \). \( V_{\text{Th}} \) is the threshold voltage of transistor \( T_{\text{Th}} \). \( Q_{\text{DQ1}} \) is constant as long as \( T_{\text{N}} \) is off (\( V_{\text{TSS}} < V_{\text{Th}} \)) and \( V_{\text{Th}} \) rises due to the coupling capacitance \( C_{\text{MSS,CMSS}} \). When \( V_{\text{DD}} \) is high, the transistor \( T_{\text{DD}} \) starts to conduct and the charge at node \( N-1 \) decreases, the voltage \( V_{\text{DQ1}} \) also starts to decrease and transistor \( T_{\text{DD}} \) is off until \( V_{\text{DQ1}} = V_{\text{DD}} - V_{\text{Th}} \). When transistor \( T_{\text{DD}} \) is turned on the chain starts to conduct and we consider that the input \( V_{\text{TSS}} \) reaches its dynamic threshold voltage \( V_{\text{Th,N-1}} \). The charge stored at node \( N-1 \) when the chain starts to conduct is \( Q_{\text{DQ1}} = Q_{\text{DQ2}} = (V_{\text{DD}} - V_{\text{Th}}) \). During the interval \( V_{\text{Th}} < V_{\text{TSS}} < V_{\text{DD}}-V_{\text{Th}} \), the total charge \( Q_{\text{DQ1}} = Q_{\text{DQ2}} \) is transferred through transistor \( T_{\text{DD}} \). Using the saturation expression of the n-p-n power law MOSFET model [1] we compute the charge transferred through transistor \( T_{\text{DD}} \) as:

\[
t_{\text{DD}} = \frac{I_{\text{DD}}}{V_{\text{DD}} - V_{\text{Th}} - x_{\text{N-1}}} \int_{I_{\text{DD}}}^{V_{\text{DD}}} \frac{1}{V_{\text{DD}} - V_{\text{Th}} - x_{\text{N-1}}} \text{d}t = \frac{Q_{\text{DQ1}}}{V_{\text{DD}} - V_{\text{Th}} - x_{\text{N-1}}} = \frac{C_{\text{MSS,CMSS}} V_{\text{DD}}}{V_{\text{DD}} - V_{\text{Th}} - x_{\text{N-1}}} \]

where \( x_{\text{N-1}} \) is the time at which \( T_{\text{DD}} \) starts to conduct and \( t_{\text{DD}} \) is the time when the chain starts to conduct. To solve (3) we assume a linear variation of voltage \( T_{\text{DD}} \) with an input rise time \( t_{\text{DD}} \), i.e. \( T_{\text{DD}} = V_{\text{DD}}/t_{\text{DD}} \). The dynamic threshold voltage is obtained by solving (3):

\[
V_{\text{Th,N-1}} = V_{\text{Th}} + \frac{V_{\text{DD}}}{x_{\text{N-1}}} \left( 1 + \sqrt{1 + 2x_{\text{N-1}}} \right)
\]

where the velocity saturation index of \( x_{\text{N-1}} \) is taken as \( x_{\text{N-1}} = 1 \), which is a reasonable approximation for submicron devices. The parameter \( x_{\text{N-1}} \) is given by:

\[
x_{\text{N-1}} = \frac{I_{\text{DD}}}{V_{\text{DD}} - V_{\text{Th}} - x_{\text{N-1}}}
\]

The dynamic threshold voltage of the chain when varying \( V_{I} (i \neq N) \) is computed similarly as \( V_{\text{DD}} = V_{\text{Th}} + \frac{V_{\text{DD}}}{t_{\text{DD}}} \left( 1 + \sqrt{1 + 2t_{\text{DD}}} \right) \), with \( t_{\text{DD}} \) given by:

\[
x_{\text{DD}} = \frac{I_{\text{DD}}}{V_{\text{DD}} - V_{\text{Th}} - x_{\text{N-1}}}
\]

The parameter \( t_{\text{DD}} \) is obtained by collapsing the chain of \( N-1 \) transistors \( T_{\text{N}}, T_{\text{DD}} \) that are discharging the node \( N-1 \) into a single equivalent transistor with a maximum saturation current of \( t_{\text{DD}} \) (see [2] for a detailed analysis of the collapsing technique).

When \( V_{I} \) switches, the dynamic threshold voltage is taken as the threshold voltage of transistor \( T_{\text{DD}} (V_{\text{DD}} - V_{\text{Th}}) \). Figure 1 shows HSPICE simulations of the dynamic threshold voltage for a four transistor chain (symbols) with respect to each input voltage \( V_{I} \) making a low to high transition with an input rise time \( t_{\text{DD}} \). The accuracy obtained is sufficient for the purpose of this work.

**Modelling effective capacitance**: The output response of a CMOS gate is a function of the position of the switching transistor in the chain and its relative switching speed with respect to the output response. For an ideal input step all the capacitances (the output capacitance to \( V_{\text{DD}} \) and internal capacitances to \( V_{\text{DD}} \) and \( V_{\text{Th}} \)) are discharged at a rate given by the maximum current of the chain \( f_{\text{DD}} \). The total output capacitance when \( V_{I} \) switches \( (C_{\text{DD}}) \) can be expressed as \( C_{\text{DD}} = C_{\text{DD}} + \sum_{i=1}^{N-1} (C_{i} + C_{\text{DD}}) \left( V_{\text{DD}} - V_{\text{Th}} \right) / V_{\text{DD}} \), where \( C_{\text{DD}} \) includes all the output capacitance terms and \( C_{\text{DD}} \) is the gate capacitance of transistor \( T_{\text{DD}} \). For slow input transitions the internal capacitances \( C_{i} \) and \( C_{\text{DD}} \) are discharged faster (through a smaller number of transistors) than the output capacitance \( C_{\text{DD}} \). In this case, for simplicity, we express the total capacitance discharged through the chain as \( C_{\text{DD}} = C_{\text{DD}} \). The transition between fast and slow input transitions is expressed as a function of the ratio \( K = (Q_{\text{DD}} t_{\text{DD}}) / (C_{\text{DD}} V_{\text{DD}}) \) (the relation between the maximum charge that can be discharged through the chain during a time interval \( t_{\text{DD}} \) and the charge stored at the output capacitance). The input is considered fast when \( K < 1 \), otherwise the input is considered slow. An accurate expression for the 'effective' output capacitance discharged through the chain as a function of \( K \) is given by:

\[
C_{\text{DD}} = C_{\text{DD}} + \sum_{i=1}^{N-1} \frac{V_{\text{DD}} - V_{\text{Th}}}{V_{\text{DD}}} (C_{i} + C_{\text{DD}}) e^{-K}
\]
Results: We performed HSPICE simulations of the propagation delay for several 0.18 μm CMOS technology gates varying each input. We used parameters $f_i^{(2)}$, $V_{PD}$, and $n_i^{(2)}$ in the propagation delay model of [3] for CMOS inverters, with $n_i^{(2)}$ extracted for the best fit of the propagation delay. When the upper transistor switches we assume $n_i^{(2)} = n_{to}$ (where $n_{to}$ is the velocity saturation index of $T_{PD}$). Fig. 2 shows the results for a three input NAND gate with $n_i^{(2)} = 0.8$ and $n_i^{(3)} = 0.66$, showing good agreement with respect to HSPICE simulations. Results are also compared with the collapsing technique in [3], showing that the proposed method provides better description with respect to HSPICE. Note that the results from [3] are identical to the proposed model only when input $V_5$ switches with $V_{PD} = V_{PP}$, and $C_{pq} = C_{Gm}$. 

![Propagagation delay of 3-NAND CMOS gate against input rise time](image)

Fig. 2 Propagation delay of 3-NAND CMOS gate against input rise time

- this work
- gate delay model from [3]
- HSPICE (varying $V_{PD}$)
- HSPICE (varying $V_{PP}$)
- HSPICE (varying $V_{IP}$)

Conclusion: We have presented an accurate method to evaluate the propagation delay of complex CMOS gates when only one input changes. The method uses an accurate description of the dynamic threshold voltage for a chain of transistors that is collapsed into a single device. We then use the propagation delay model of CMOS inverters developed in [3]. HSPICE simulations (level 50) for a 0.18 μm technology show that this approach represents an improvement over existing models, providing very good accuracy.

Acknowledgment: This work has been supported by the Spanish Government under project CICYT-TIC98-0284.

© IEE 2002 24 January 2002
Electronics Letters Online No: 20020548
DOI: 10.1049/el:20020548
J.L. Roßello and J. Segura (Physics Department, Balearic Islands University, Palma de Mallorca 07071, Spain)
E-mail: j.rossello@uib.es

References


Test structure for crosstalk characterisation

F. Picot, P. Coll and D. Auvergne

Based on the characterisation of an industrial driver library in terms of crosstalk-induced noise possibility, a specific test structure to measure crosstalk signal on interconnect lines is presented. An original implementation is proposed for direct amplitude and pulse with measurement of the crosstalk-induced parasitic signal. A validation is given with HSPICE simulation of the extracted layout of the structure implemented in a 0.25 μm process.

Introduction: With the increasing level of integration in VLSI the problem of signal propagation in interconnect lines is of great concern in designing high performance ICs. The robustness to crosstalk has been particularly widely investigated [1]. Crosstalk glitches appear on metal layers when the cross-coupling capacitance between metal lines has a greater value than a threshold. The determination of this threshold is of fundamental importance to guarantee the signal integrity.

Many studies have been conducted to model [2, 3] metal lines and crosstalk effects but few are concerned with measurement systems. The main difficulties in directly measuring parasitic crosstalk effects are the very high swiftness of the induced signal and its very low energy. Several methods [4] can be used for measuring crosstalk effect but no technique has been set up to directly obtain both the amplitude and the timing characteristics of the parasitic signal. Where systems give both informations, they are based on sampling techniques [5], which are difficult to tune.

In this Letter we present a self-calibrated structure for evaluating the voltage amplitude and the timing characteristics of the crosstalk-induced parasitic signal. The novelty of this structure is to provide the facility to accurately measure short duration glitch width.

Description of measurement technique: The structure is composed of two parts: a threshold-based amplitude detection of overshoots and a time duration measurement of these overshoots. The test structure is composed of two aggressor lines surrounding the victim line. The input of each line is controlled by a tri-state buffer allowing the fixing of different drive strengths for imposing the state of the line. The worst-case configuration defined in the ATMEL (0.25 μm) process is considered: the victim driver imposes a low level at the input of the line and the aggressor drivers generate a positive pulse.

The output of the victim line is fed to the amplitude and pulse width measurement system.

Amplitude measurement: In Fig. 1 we describe the voltage amplitude measurement of the crosstalk signal induced on the victim line. It constitutes an elementary analogue-to-digital flash converter. The array of inverters connected to the output of the victim line has been designed with different threshold voltages. As shown in Fig. 1, the filtering action of this array allows observation of only a stable level at the output of the inverter with a threshold voltage value immediately superior to the amplitude value of the crosstalk signal.

![Crosstalk peak amplitude detection](image)

Fig. 1 Crosstalk peak amplitude detection

Pulse width measurement: Considering that the pulses at the inverter output have the width of the parasitic signal at the threshold voltage,
An Analytical Charge-Based Compact Delay Model for Submicrometer CMOS Inverters
José Luis Rosselló and Jaume Segura

Abstract—We develop an accurate analytical expression for the propagation delay of submicrometer CMOS inverters that takes into account the short-circuit current, the input–output coupling capacitance, and the carrier velocity saturation effects. The propagation delay is computed by the time required for a charge of the input terminal to propagate through the device and charge the output capacitance. The model is compared with HSPICE level 50 simulations and other previously published models for a 0.18-μm and a 0.35-μm process technologies show significant improvements over previous models.

Index Terms—Analytical model, circuit modeling, delay estimation, submicrometer MOSFETs.

I. INTRODUCTION

Timing analysis is one of the most critical topics in very large-scale integration (VLSI) design. The nonlinear behavior of CMOS gates requires numerical calculations for accurate timing analysis at the expenses of large computation times. Moreover, the impact of design parameters such as fan-in, fan-out, or transistor sizes on the propagation delay are difficult to understand and optimize using numerical procedures.

The dynamic behavior of submicrometer CMOS inverters depends on several nonlinear effects like the velocity saturation of carriers due to the high electric fields in submicrometer technologies, the short-circuit current appearing when both pMOS and nMOS transistors conduct simultaneously [1], and the additional effect of the input–output coupling capacitance [2].

Several methods have been proposed to derive the delay of CMOS inverters [2]–[8] as a first step to describe more complex gates [9], [10]. Cocchini et al. [3] obtained a piecewise expression for the propagation delay based on the Berkeley short-channel IGFET model (BSIM) MOSFET model [11]. The input–output coupling capacitance was neglected. Jeppson in [2], and Bisdounis et al. presented a model for the output response of CMOS inverters using a quadratic current-voltage dependence for MOSFET devices which is not longer valid for submicrometer technologies. Daga and Auvergne [5] obtained an empirical expression for the propagation delay taking into account both overshooting and short-circuit currents.

Hirata et al. [6] derived a propagation delay model with numerical procedures based on the α-th-power-law MOSFET model [12] considering both short-circuit and overshooting currents. The model provides an accurate description of the propagation delay but the numerical procedures used in their analysis increase the computation time considerably. Bisdounis et al. [7] developed a piecewise solution with seven operation regions for the transient response of a CMOS inverter based on the α-power-law MOSFET model [13] including both overshooting and short-circuit currents. Recently, Kabbani et al. obtained, in [8], a transition time model considering a quadratic relationship between the saturation current and the gate–source voltage (assumption only valid for micronic devices) without considering the effect of the input–output coupling capacitance. In [9], Sakurai and Newton obtained a simple expression for the propagation delay of CMOS gates based on their α-th-power-law MOSFET model neglecting both short-circuit and overshooting currents.

In this paper, we propose an analytical model to accurately compute the propagation delay of a CMOS inverter accounting for the main effects of submicrometer technologies like the input–output coupling capacitance, carriers velocity saturation effects and short-circuit currents. The model is based on an accurate physically-based α-th-power-law MOSFET model [14] and on a previous power dissipation model for CMOS inverters [15]. The model can be applied to compute the propagation delay of CMOS inverters and represents a valuable approach for the evaluation of delay in complex gates as these can be collapsed to a single equivalent inverter for delay evaluation [10]. Comparisons with previously published models and HSPICE simulations using the Philips MOSFET Model 9 (MM9) for a 0.18-μm and 0.35-μm process technologies show significant improvements in terms of accuracy.

This paper is organized as follows. In Section II, we describe briefly the switching characteristics of CMOS inverters and introduce the MOSFET device model used in this work. In Section III, we derive an analytical charge-based expression for the switching response of a single pMOS/nMOS transistor charging/discharging a capacitor, that are generalized to CMOS inverters in Section IV by including overshooting and short-circuit effects. Section V presents an analytical model for the output transition time based on the delay model developed. The model proposed is compared to HSPICE simulations and other previously published models for a 0.35-μm and a 0.18-μm process technology in Section VI. Finally, in Section VII, we conclude the paper.

Manuscript received May 25, 2003; revised August 31, 2003. This work was supported in part by the Spanish Ministry of Science and Technology, in part by the Regional European Development Funds (FEDER) under EU Project TIC2002-01238, and in part by Intel Laboratories-CRL. This paper was recommended by Associate Editor A. I. Karsilayan.

The authors are with the Physics Department, Balearic Islands University, 07071 Palma de Mallorca, Spain (e-mail: j.rossello@uib.es; jaume.segura@uib.es).

Digital Object Identifier 10.1109/TCSI.2004.830692
Fig. 1. CMOS inverter model.

II. ANALYSIS OF CMOS INVERTER SWITCHING CHARACTERISTICS

The dynamic behavior of the CMOS inverter in Fig. 1 is described by

\[(C_L + C_M) \frac{dV_{\text{out}}}{dt} = I_p - I_n + C_M \frac{dV_{\text{in}}}{dt}\]  \hspace{1cm} (1)

where \(C_L\) is the output capacitance that is composed by the drain capacitances of both nMOS and pMOS transistors, the output node wiring capacitance, and the input capacitance of the gates connected to the inverter output. \(V_{\text{out}}\) and \(V_{\text{in}}\) are the output and input voltages, respectively, while \(I_p\) and \(I_n\) are the pMOS and nMOS currents, respectively. \(C_M\) is the input-to-output coupling capacitance, which is strongly voltage dependent. The static value of \(C_M\) when the input is low (\(C_{M0}\)) is computed considering the overlap capacitances of both transistor drains and the gate-to-drain capacitance of the pMOS transistor that operates in the linear region [16]

\[C_{M0} = C_M \left( \frac{W_{\text{eff}}}{L_{\text{eff}}} + L_{Dp} \frac{W_{\text{eff}}}{L_{\text{eff}}} + L_{Dn} \frac{W_{\text{eff}}}{L_{\text{eff}}} \right) \]  \hspace{1cm} (2)

with \(W_{\text{eff}}\) and \(L_{\text{eff}}\) being the effective channel width of pMOS and nMOS, respectively, \(L_{Dp}\) is the effective channel length of pMOS, while \(L_{Dn}\) and \(L_{Dp}\) are the gate–drain underdiffusion for the nMOS and pMOS transistors, respectively. For a static input high, the capacitance \(C_{M0}\) is obtained similarly.

Fig. 2 illustrates the input and output voltages evolution of the inverter along with the current through the nMOS and pMOS transistors for a low-to-high input transition. The current through the pMOS transistor (\(I_p\) in Fig. 2) has two components clearly distinguished by the sign of the current. The negative pMOS current is due to a partial discharge of the output capacitance from the output node toward the supply rail and appears when the input–output capacitance drives the output voltage beyond the supply value (\(V_{\text{DD}}\)) at the beginning of the transition [2]. This effect is known as overshoot and the time during which the output voltage is beyond the supply value is defined as the overshoot time, \(t_{\text{ov}}\). When the nMOS device starts to conduct, it pulls the output voltage down. Once the output voltage goes below \(V_{\text{DD}}\), the pMOS current is positive corresponding to the short-circuit component due to the simultaneous conduction of both devices.

The propagation delay (defined as \(t_{\text{pu}}\)) for a high-to-low output transition is typically defined as the time interval from the 50% \(V_{\text{DD}}\) voltage input to the 50% \(V_{\text{DD}}\) voltage output. The dependence of the propagation delay with design parameters is nonlinear and difficult to model given that (1) cannot be solved in a closed form even using the simple Shockley MOSFET model [17]. Moreover, carrier saturation effects become important with technology scaling, thus requiring more complex MOSFET models accounting for such effects.

The \(n\)th-power-law MOSFET model [12] is a widely used short-channel drain-current model, and will be used in this work to derive the propagation delay and the output transition time of CMOS inverters. The drain current is given by

\[I_D = \begin{cases} 
0, & \text{if } V_{\text{GS}} \leq V_{\text{TH}} \\
\frac{(2 - V_{\text{DS}}/V_{\text{DD}})^n - V_{\text{DS}}}{V_{\text{DD}} - V_{\text{TH}}} I_D^{0}, & \text{if } V_{\text{DS}} < V_{\text{DD}} \\
I_D^{0}, & \text{if } V_{\text{DS}} \geq V_{\text{DD}}
\end{cases}\]  \hspace{1cm} (3)

with

\[I_D^{0} = I_D^{0} \left( \frac{V_{\text{GS}} - V_{\text{TH}}}{V_{\text{DD}} - V_{\text{TH}}} \right)^n (1 + \lambda(V_{\text{DS}} - V_{\text{DD}})) \]  \hspace{1cm} (4)

where \(V_{\text{GS}}, V_{\text{DD}}, \) and \(V_{\text{DD}}^{0}\) are the gate, supply, and saturation voltages, respectively, and \(I_D^{0}\) is the drain current at \(V_{\text{GS}} = V_{\text{DS}} = V_{\text{DD}}\). The parameter \(n\) is the velocity saturation index that ranges between 2 (long-channel devices) and 1 (short-channel) [12], and \(\lambda\) describes the channel length modulation. The saturation voltage \(V_{\text{DD}}^{0}\) is given by

\[V_{\text{DD}}^{0} = V_{\text{DD}} \left( \frac{V_{\text{GS}} - V_{\text{TH}}}{V_{\text{DD}} - V_{\text{TH}}} \right)^m \]. \hspace{1cm} (5)

The parameter \(V_{\text{DD}}^{0}\) is the saturation voltage at \(V_{\text{GS}} = V_{\text{DD}}\), while \(m\) and \(V_{\text{TH}}\) are empirical parameters [12]. These equations are mathematically simpler than physically-based MOSFET models such as BSIM3v3 or MM9 with the disadvantage that, in the original model developed by Sakurai and Newton, the relationship between the empirical parameters and the process parameters supplied by manufacturers is not provided. Therefore, the variation of \(n\)th-power-law model predictions with key parameters like the supply voltage are not taken into account in the original formulation performed by Sakurai and Newton, where each parameter must be recomputed if the supply voltage or any device dimension change. In this paper, we use the physical formulation for the \(n\)th-power-law MOSFET model proposed in [14] and used in [15]. This physical formulation provides an analytical relationship between the \(n\)th-power-law parameters and the foundry-provided MOSFET parameters.

III. CHARGE-BASED PROPAGATION DELAY OF SINGLE nMOS DISCHARGING TRANSISTOR

The analysis of discharging a capacitor through a nMOS transistor is developed in Appendix A (for a pMOS charging a capacitor the analysis is equivalent). Four cases are considered depending on the input voltage state (rising or static high) and the nMOS region of operation (ohmic or saturation). In this section we use the output voltage expressions obtained in Appendix A to evaluate a charge-based expression for the propagation delay. We state initially some definitions that will be used extensively during the development of the model.
Fig. 2. Output voltage and current evolution in the nMOS and pMOS transistors for a rising input linear voltage.

In the context of this work, a fast-output transition refers to the cases where the nMOS transistor enters in its linear region while the input is changing. Otherwise (if the nMOS is saturated while $V_{\text{in}} < V_{\text{DD}}$), the output transition is considered to be slow.

We also define the discharging time ($t_d$) up to a given voltage $V_f$ as the time interval from the beginning of the input transition until the instant at which the output is discharged at $V_{\text{out}} = V_f$ (an arbitrary voltage value between 0 and $V_{\text{DD}}$).

Finally, the propagation delay ($t_{\text{plh}}$) is defined as the time interval from $V_{\text{in}} = V_{\text{DD}}/2$ to $V_{\text{out}} = V_{\text{DD}}/2$ (See Fig. 2). If $t_{\text{in}}$ is defined as the input transition time, the propagation delay is related to $t_d$ for $V_f = V_{\text{DD}}/2$ as

$$t_{\text{plh}} = t_d \left( \frac{V_{\text{DD}}}{2} \right) - \frac{t_{\text{in}}}{2}.$$  \hfill (6)

In this section, we first obtain an analytical expression of the discharging time $t_d$ for slow outputs, and then, we correct this expression to account also for fast outputs. The discharging time will be expressed in terms of the charge transferred through the nMOS transistor (defined as $Q_f$ and equal to $C_L (V_{\text{DD}} - V_f)$). This charge is expressed in terms of four components

$$Q_f = Q_{\text{sat}}^r + Q_{\text{lin}}^r + Q_{\text{sat}}^H + Q_{\text{lin}}^H,$$  \hfill (7)

where $Q_{\text{sat}}^r$ and $Q_{\text{lin}}^r$ are the charges transferred while the input is rising and the nMOS is in saturation and linear region, respectively. $Q_{\text{sat}}^H$ and $Q_{\text{lin}}^H$ are defined similarly but when the input is static high (i.e., $V_{\text{in}} = V_{\text{DD}}$). The discharging time expression will be given in terms of these four charge components.

### A. Slow-Output Transitions

For a slow-output transition, the nMOS transistor is saturated during the whole input transition and the output voltage evolution is given by (A6) (see Appendix A). Once the input is high, the output voltage is described by (A2) and (A4) since the nMOS is initially saturated and enters the linear region at the end of the output transition. We express the time needed to discharge the output from $V_{\text{DD}}$ until $V_{\text{out}} = V_f$ (discharging time) as a function of the charge transferred during each region ($Q_{\text{sat}}^r$, $Q_{\text{sat}}^H$, and $Q_{\text{lin}}^H$) using (A2), (A6), and (A4)

$$t_d = t_n + \frac{Q_{\text{sat}}^r}{I_{D_{on}}} \left( \frac{n_n + 1}{n_n} \right)^{1/(n_n+1)} - \frac{t_{\text{in}}}{n_n} + \frac{Q_{\text{lin}}^H}{2I_{D_{on}}} \ln \left( 1 + \frac{2Q_{\text{lin}}^H}{Q_{\text{lin}}^H} \right)$$  \hfill (8)

where $Q_{D_{on}} \equiv C_L V_{D_{on}}$, and $n_n$ is the velocity saturation index of nMOS (in general, a subindex $n(p)$ denotes a parameter for the nMOS(pMOS)

To obtain $Q_{\text{sat}}^r$, $Q_{\text{sat}}^H$, and $Q_{\text{lin}}^H$, we define the saturation charge $Q_{\text{sat}}$, and the linear charge $Q_{\text{lin}}$ as the total charge transferred when the nMOS transistor is in the saturation and in the linear region respectively. From these definitions, it follows that

$$Q_{\text{sat}} = \min \{ Q_f, Q_{\text{sat}}^\text{max} \}$$

$$Q_{\text{lin}} = Q_f - Q_{\text{sat}}$$  \hfill (9)

where $Q_{\text{sat}}^\text{max}$ is the maximum charge that can be transferred through the nMOS transistor operating in the saturation region, given by

$$Q_{\text{sat}}^\text{max} = C_L (V_{\text{DD}} - V_{D_{on}})$$

where $C_L$ is the output capacitance. We obtain $Q_{\text{sat}}^r$ from $Q_{\text{sat}}$ as

$$Q_{\text{sat}}^r = \min \{ Q_{\text{sat}}, Q_{\text{sat}}^\text{max} \}$$  \hfill (10)

where $Q_{\text{sat}}$ is given by (9) and $Q_{\text{sat}}^\text{max}$ is the maximum charge that can be transferred through the saturated nMOS transistor during the input transition. The value of this maximum charge...
is obtained integrating the nMOS current during the rising input transition

\[ Q_{\text{nsat}}^r = \int_0^{t_{\text{nsat}}} (I_{\text{Do}} - I_{\text{Dn}}) dt = I_{\text{Do}} n_{\text{nsat}} - t_n, \]

Therefore, part of the saturation charge \( Q_{\text{nsat}}^r \) is transferred while the input is rising \( Q_{\text{nsat}}^r \), and the other part is transferred while the input voltage remains high (defined as \( Q_{\text{nsat}}^h \))

\[ Q_{\text{nsat}}^h = Q_{\text{nsat}} - Q_{\text{nsat}}^r. \]

Finally, for the slow-output ramp case we have that \( Q_{\text{nsat}}^h = Q_{\text{lin}} \), since by definition, the input is high when the device enters the linear region.

**B. Inclusion of Fast-Output Transition Range**

Equation (8) is valid only if the nMOS transistor is saturated during the whole rising input period (for this case, the term \( Q_{\text{lin}}^r \) is zero). For fast-output transitions, the case in which the nMOS enters in its linear region while the input is rising has no analytical solution (see Appendix A), and an approximation is required for an analytical description of the discharging time.

The delay components of \( t_d \) in (8) show mathematical analogies among the terms corresponding to the period during which the input is rising (the two first terms), and those corresponding to the input being static high (the last two). The time during which the nMOS is saturated and the input is high is given by the third term in (8), (that is, \( Q_{\text{nsat}}^h / I_{\text{Dn}} \)) while the time during which the nMOS is saturated but the input is rising also contains the same charge to current ratio \( Q_{\text{nsat}}^h / I_{\text{Dn}} \) but corrected with an expression that depends on the transition time and the saturation velocity

\[ t_{\text{nsat}}^r = \left[ \frac{Q_{\text{nsat}}^r}{I_{\text{Dn}}} (n_{\text{nsat}} + 1) (t_n - t_{\text{nsat}}) \right]^{1/(1+n_{\text{nsat}})}. \]

When the output transition is fast, the linear charge is nonzero when the input is rising (i.e., \( Q_{\text{lin}}^r \neq 0 \), and must appear in the delay expression of \( t_d \). We include the term \( Q_{\text{lin}}^r \) in \( t_d \) assuming that the charge component due to the static and dynamic input periods are of the same form

\[ t_d = t_n + \left\{ \frac{Q_{\text{lin}}^r}{I_{\text{Dn}}} + \frac{Q_{\text{Do}}^n}{I_{\text{Do}}} \ln \left( 1 + \frac{2Q_{\text{lin}}^r}{Q_{\text{Do}}^n - Q_{\text{lin}}^r} \right) \right\}^{1/(1+n_{\text{lin}})} (t_n - t_{\text{lin}})^{n_{\text{lin}}/(n_{\text{lin}}+1)} + \frac{Q_{\text{lin}}^h}{I_{\text{Do}}} + \frac{Q_{\text{Do}}^n}{I_{\text{Do}}} \ln \left( 1 + \frac{2Q_{\text{lin}}^h}{Q_{\text{Do}}^n - Q_{\text{lin}}^h} \right). \]

From (14), we must evaluate the parameters \( Q_{\text{lin}}^r \) and \( Q_{\text{lin}}^h \). Similar to the previous analysis of the saturation charge, the linear charge transferred during the rising input transition \( Q_{\text{lin}}^r \) is bounded by a maximum value \( Q_{\text{lin}}^{\text{max}} \) that is the maximum charge that can be transferred through the nMOS transistor operating in the linear region while the input is rising

\[ Q_{\text{lin}}^r = \min (Q_{\text{lin}}, Q_{\text{lin}}^{\text{max}}). \]

The value of \( Q_{\text{lin}}^{\text{max}} \) is obtained by computing the discharging time (14) at the end of the input transition (that is, when \( t_d = t_{\text{lin}} \)). For this special case, we can set that \( Q_{\text{lin}}^r = Q_{\text{lin}}^{\text{max}} \)

\[ t_d = t_n + \left[ \frac{Q_{\text{lin}}^r}{I_{\text{Dn}}} + \frac{Q_{\text{Do}}^n}{I_{\text{Do}}} \ln \left( 1 + \frac{2Q_{\text{lin}}^{\text{max}}}{Q_{\text{Do}}^n - Q_{\text{lin}}^{\text{max}}} \right) \right]^{1/(1+n_{\text{lin}})} (t_n - t_{\text{lin}})^{n_{\text{lin}}/(n_{\text{lin}}+1)} = t_{\text{lin}}. \]

After some algebra, (16) leads to

\[ Q_{\text{lin}}^r + \frac{Q_{\text{Do}}^n}{I_{\text{Do}}} \ln \left( 1 + \frac{2Q_{\text{lin}}^{\text{max}}}{Q_{\text{Do}}^n - Q_{\text{lin}}^{\text{max}}} \right) = t_{\text{lin}} - t_n \]

In (17), we use \( Q_{\text{lin}}^r = Q_{\text{lin}}^{\text{max}} \) because the input transition is finished \( (t_d = t_{\text{lin}}) \) and \( Q_{\text{lin}}^{\text{max}} \) achieves its maximum value. The value of \( Q_{\text{lin}}^{\text{max}} \) is obtained solving (17)

\[ Q_{\text{lin}}^{\text{max}} = Q_{\text{Do}}^n \tanh \left( \frac{Q_{\text{lin}}^{\text{max}} - Q_{\text{lin}}^r}{Q_{\text{Do}}^n} \right) \]

where \( Q_{\text{lin}}^{\text{max}} \) is given by (11). Finally, \( Q_{\text{lin}}^h \) is obtained from \( Q_{\text{lin}} \)

\[ Q_{\text{lin}}^h = Q_{\text{lin}} - Q_{\text{lin}}^r. \]

In Fig. 3, we plot HSPICE simulations of the output response of a single discharging nMOS transistor for a 0.35-μm technology showing very good accuracy. Both slow- and fast-output transitions are simulated for various output capacitance values (from \( C_{\text{lin}} \) to \( 31C_{\text{lin}} \)), where \( C_{\text{lin}} \) is the input capacitance of a minimum sized inverter. (14), the propagation delay is obtained applying (6).

**IV. PROPAGATION DELAY FOR CMOS INVERTER**

In this section, we consider a whole inverter (and not only a charging/discharging transistor) and include the effect of the short-circuit (due to the pMOS/nMOS transistor during the discharging/charging process) and overshooting currents (due to the input–output coupling capacitance \( C_{\text{M}} \)). Their contribution is included as additional charges to be transferred through the charging/discharging transistor. We develop only the discharging case as the charging case is equivalent.

**A. General Equation**

The charge transferred from the output capacitances (both \( C_{\text{M}} \) and \( C_{\text{L}} \) in Fig. 1) through the nMOS transistor \( Q_f \) is computed from the difference between the charge initially stored at the output node \( (C_{\text{M}} + C_{\text{L}})V_{\text{DD}} \) and the charge remaining in this node when the output voltage reaches the desired value \( V_f \). Assuming that \( V_{\text{in}}(V_{\text{out}} = V_f) = V_{\text{DD}} \), we have

\[ Q_f = C_{\text{L}} V_{\text{DD}} - (C_{\text{M}} V_f - C_{\text{H}} V_{\text{DD}}) \]

where \( C_{\text{L}} = C_{\text{L}} + C_{\text{M}} \) and \( C_{\text{H}} = C_{\text{L}} + C_{\text{M}} \) are the sum of the output capacitances when the input is low and high, respectively. Initially, \( V_{\text{in}} = 0 \) and \( C_{\text{M}} = C_{\text{L}} \), while, when \( V_{\text{out}} = V_f \), we
assume that $C_M = C_M^H$. From (20), the maximum saturation charge $Q_{\text{sat}}^{\text{max}}$ is

$$Q_{\text{sat}}^{\text{max}} = C_S^H V_{DD} - (C_S^H V_{D0h} - C_M^H V_{DD})$$ \hspace{1cm} (21)

For the evaluation of the propagation delay, we are interested in the charge transferred only from the output capacitances $C_L$ and $C_M$ through the nMOS transistor. Therefore, we use the same terms defined in the previous section for charges $Q_{\text{sat}}^r$, $Q_{\text{lin}}^r$, $Q_{\text{sat}}^H$, and $Q_{\text{lin}}^H$. It is important to remark that, by definition, these charges come from the capacitances connected to the output node and therefore do not include short-circuit contribution. Therefore, $Q_{\text{sat}}^{r \text{max}}$, the maximum value of saturation charge during the rising input, is obtained from (11) by subtracting the short-circuit charge transferred (SCCT)

$$Q_{\text{sat}}^{\text{max}} = I D_{0h} t_{\text{in}} - t_n - Q_{\text{sc}}$$ \hspace{1cm} (22)

where $Q_{\text{sc}}$ is the SCCT for a rising input transition. The expression of this component is taken from [15].

The charge transferred through the nMOS transistor comes from the output capacitances ($Q_{\text{sat}}^r$) and the power supply through the pMOS transistor (SCCT). The propagation delay is computed considering both contributions by including an additional charge (a fraction of the SCCT) to the charge transferred from the output capacitances. Therefore, from (6) and (14), the propagation delay expression is given by

$$t_{\text{pin}} = t_n + \left\{ \frac{Q_{\text{sat}}^r + Q_{\text{sc}}}{I D_{0h}} + \frac{Q_{\text{Sat}}^h + Q_{\text{Sat}}^H}{2 I D_{0h}} \left( 1 + \frac{2 Q_{\text{Sat}}^h}{Q_{\text{Sat}}^h - Q_{\text{Sat}}^H} \right) \right\}^{1/(1 + \rho_n)} + (t_{\text{in}} - t_n)^{\rho_n/(\rho_n + 1)} + \frac{Q_{\text{Sat}}^h}{I D_{0h}} + \frac{Q_{\text{Sat}}^H}{2 I D_{0h}} \left( 1 + \frac{2 Q_{\text{Sat}}^H}{Q_{\text{Sat}}^H - Q_{\text{Sat}}^H} \right) - \frac{t_{\text{in}}}{2}$$ \hspace{1cm} (23)

where $q_{\text{sc}}^r$ is a fraction of the SCCT ($Q_{\text{sc}}^r$) and is defined as the SCCT during the time interval such that $V_{\text{out}} > V_{DD}/2$. This short-circuit charge is added to the saturation charge $Q_{\text{sat}}^r$ for simplicity. The expression used for the evaluation of $q_{\text{sc}}^r$ is computed in the Appendix B while parameters $Q_{\text{Sat}}^r$, $Q_{\text{Sat}}^H$, and $Q_{\text{Sat}}^H$ are computed at $V_f = V_{DD}/2$.

V. OUTPUT TRANSITION TIME COMPUTATION ($t_{\text{out}}$)

The effective output transition time $t_{\text{out}}$ can be related to a percentage of the $V_{\text{out}}$ derivative at the half $V_{DD}$ point. Using this property, Sakurai et al. approximate $t_{\text{out}}$ from the 70% of the $V_{\text{out}}$ derivative [9]. From SPICE simulations, we observed that the output voltage shape is different depending on the relative switching speed between the input and the output voltages. For fast-output transitions (see Fig. 4), the output voltage evolution is close to the dc voltage characteristics of the inverter and the output transition slope is strongly dependent of $V_{\text{in}}$ (since the nMOS conductance mainly depends on $V_{\text{in}}$). For slow-output transitions (see Fig. 5), the output voltage slope is nearly constant since the nMOS gate voltage is at $V_{DD}$ almost from the beginning of the output transition. Therefore, the percentage of the $V_{\text{Sat}}$ derivative at $V_{DD}/2$ that must be used to compute $t_{\text{out}}$ depends on the relative switching speed of $V_{\text{in}}$ and $V_{\text{out}}$, varying from 90% to 40% for a slow or fast-output transitions respectively. We use an empirical expression for the output transition time $t_{\text{out}}$ that takes into account this effect

$$t_{\text{out}} = \frac{V_{DD}}{0.4 + 0.5 \min \left( 1, \frac{t_{\text{Sat}}^H}{t_{\text{in}}} \right) } \left| \frac{\frac{d V_{\text{Sat}}}{d t}}{V_{DD}/2} \right|$$ \hspace{1cm} (24)

where the $T_{\text{Sat}}/T_{\text{in}}$ ratio provides the information of the relative switching speed between the two nodes. This ratio is close to 0 or greater than 1 for fast and slow output transitions respectively. The exact definition of $T_{\text{Sat}}$ and $T_{\text{in}}$ is given in Appendix B. The expression for $t_{\text{out}}$ used in (24) shows an excellent agreement with respect to HSPICE simulations for both
slow and fast-output transitions. In the analysis for the evaluation of the $V_{\text{out}}$ derivative at $V_{\text{DD}}/2$ the short-circuit and overshooting currents are neglected for simplicity.

The derivative $dV_{\text{out}}/dt$ is obtained from the discharging time $t_d$ as

$$
\frac{dV_{\text{out}}}{dt} = \left( \frac{dt_d}{dQ_f} \right)^{-1}
$$

where $dQ_f/dV_f = -C_L$. The derivative $dt_d/dQ_f$ must be computed for $V_{\text{out}} = V_{\text{DD}}/2$. At this time point, the transistor can be in the linear or saturation region and the input can be rising or already at the high value. Therefore, $dt_d/dQ_f = dt_d/dQ_{f_{\text{sat}, \text{lin}}}$ depending on the transistor state at $V_{\text{out}} = V_{\text{DD}}/2$. For the special case in which the input is high, the discharging time derivative is

$$
\frac{dt_d}{dQ_f} = \frac{Q_{D0_n}^2 - (Q_{H_{\text{lin}}}^2)^2}{Q_{D0_n}^2 - (Q_{H_{\text{lin}}}^2)^2} I_{D0_n}.
$$

(25)

Note that when $Q_{H_{\text{lin}}}^2 (V_{\text{DD}}/2) = 0$, we obtain the case in which the nMOS is saturated [that is, $1/I_{D0_n}$, see (14)], and when $Q_{H_{\text{lin}}}^2 (V_{\text{DD}}/2) \neq 0$ the device is in the linear region when $V_{\text{out}} = V_{\text{DD}}/2$ and (25) leads to $dt_d/dQ_{H_{\text{lin}}}^2$.
If the input voltage is rising when the output is at \( V_{DD}/2 \), the discharging time at \( V_{DD}/2 \) can be simplified to
\[
t_d = t_n + f(Q_{\text{in}}^{\text{lin}}, n_n + 1) \left( t_{\text{in}} - t_n \right)^{n_n/(1+n_n)}
\]
where \( f(Q_{\text{in}}^{\text{lin}}, n) = \frac{Q_{\text{in}}^{\text{lin}}}{Q_{\text{in}}^{\text{lin}} + Q_{\text{in}}^{\text{lin}}} \left( 1 + \frac{2Q_{\text{in}}^{\text{lin}}}{Q_{\text{in}}^{\text{lin}} - Q_{\text{in}}^{\text{lin}}} \right) \).
\[
\frac{df(Q_{\text{in}}^{\text{lin}}, n)}{dQ_{\text{in}}} = \frac{t_{\text{in}} - t_n}{f(Q_{\text{in}}^{\text{lin}}, n + 1)} n_n/(1+n_n) \frac{df(Q_{\text{in}}^{\text{lin}}, n)}{dQ_{\text{in}}}. \]

Therefore, \( dt_d/dQ_{\text{in}} \) is given by
\[
\frac{dt_d}{dQ_{\text{in}}} = \frac{Q_{\text{Dn}}^2}{Q_{\text{in}}^{\text{lin}} - Q_{\text{in}}^{\text{lin}}} I_{Dn}. \]

Since \( df(Q_{\text{in}}^{\text{lin}}, n) \) is similar to (25), then
\[
\frac{dt_d}{dQ_{\text{in}}} = \frac{Q_{\text{Dn}}^2}{Q_{\text{in}}^{\text{lin}} - Q_{\text{in}}^{\text{lin}}} I_{Dn}. \]

We use the following expression for \( dt_d/dQ_{\text{in}} \) that leads to (25) and (29) for each case:
\[
\frac{dt_d}{dQ_{\text{in}}} = \frac{Q_{\text{Dn}}^2}{Q_{\text{in}}^{\text{lin}} - Q_{\text{in}}^{\text{lin}}} I_{Dn}. \]

Note that the power term in (30) is equal to 1 when \( t_d(V_{DD}/2) > t_{\text{in}} \) and (25) is recovered. Finally, the derivative of \( V_{\text{out}} \) at \( V_{DD}/2 \) is
\[
\frac{dV_{\text{out}}}{dt} \bigg|_{V_{DD}/2} = -\frac{Q_{\text{Dn}}^2 - Q_{\text{in}}^{\text{lin}}}{C_L Q_{\text{Dn}}} \frac{I_{Dn}}{2} \left[ \frac{Q_{\text{in}}^{\text{lin}}}{Q_{\text{in}}^{\text{lin}}} + \left( \frac{2Q_{\text{in}}^{\text{lin}}}{Q_{\text{in}}^{\text{lin}} - Q_{\text{in}}^{\text{lin}}} \right) \right] \left( t_{\text{in}} - t_n \right)^{n_n/(1+n_n)} (n_n + 1). \]

This expression is valid for both fast- and slow-output transitions.

**VI. RESULTS**

We plotted model results versus HSPICE level 50 simulations for a 0.35-\( \mu \)m and for a 0.18-\( \mu \)m technology. Results show that the propagation delay versus the input transition time, the supply voltage, and the \( W_n/W_p \) ratio.

In Figs. 4 and 5, we show the output waveform of a CMOS inverter driven by a slow and a fast input transition, respectively (with parameters \( C_L = 5 \, \text{fF}, t_{\text{in}} = 600 \, \text{ps} \) and \( C_L = 500 \, \text{fF}, t_{\text{in}} = 50 \, \text{ps}, \) respectively). Both figures show that the model developed for the propagation delay and the output time approximate the output response of the inverter with very good accuracy.

In Fig. 6, we plot the propagation delay \( t_{\text{prop}} \) versus the input time \( t_{\text{in}} \) for different values of the \( W_n/W_p \) ratio for a 0.35-\( \mu \)m technology. HSPICE simulations (dots) are compared to the model proposed in a previous model [3]. Since the short-circuit current is not taken into account in the model in [3], the propagation delay is underestimated. The model in [3] proposes a piecewise solution of the propagation delay, i.e., depending on the input transition (fast or slow input transitions) it uses an approximated or an exact expression. The approximated propagation delay is used when the nMOS transistors change from the saturation to the linear region and the input is rising [3, eq. (11)]. Otherwise, an exact expression for the output response is used. This leads to a discontinuity in the propagation delay when changing between regions (see Fig. 6).

Fig. 7 plots the propagation delay versus the input rise time for a 0.18-\( \mu \)m technology. When the \( W_n/W_p \) ratio is small the propagation delay decreases for increasing input rise times. The model in this work (solid lines) provides an excellent approximation to HSPICE simulations (dots) while previously published models [6], [7], that account for both overshooting and short-circuit currents, lead to underestimations.
Fig. 7. Propagation delay versus input rise time for different values of the configuration ratio for a 0.18-μm technology.

Fig. 8. Propagation delay versus supply voltage for a 0.18-μm technology.

Fig. 8 is a plot of HSPICE simulations (dots) and model predictions of the propagation delay versus the supply voltage. The model proposed in this paper provides a much better fit of the delay than the models in [6] and [7], especially for the low-voltage regime. The model in [7] shows discontinuities because it uses different expressions (Taylor series expansions) for the output response depending on the operation region of transistors.

In Fig. 9, we show the propagation delay variation with temperature for three different supply voltage values. Different trends can be appreciated for each supply voltage value when increasing temperature. The propagation delay increases for $V_{DD} = 1.8\, V$, while for $V_{DD} = 0.9\, V$, the propagation delay decreases. This effect is described properly by the proposed model since we are using the physically-based $\eta$th-power-law MOSFET model developed in [14] that relates the $\eta$th-power-law parameters to physical parameters (that are temperature-dependent) as the carrier mobility. For a more detailed description of this model, the reader is referred to [14], [15].

Fig. 10 plots the output switching response of a CMOS inverter for different output loads ($C_L = C_{\text{min}}$, $C_L = 10C_{\text{min}}$ and $C_L = 20C_{\text{min}}$) for a 0.18-μm technology. It is shown that the model provides an excellent description of the gate delay, as the deviation of the output voltage with respect to the simulations at $V_{\text{out}} = V_{DD}/2$ is very small.

Table I compares the model with HSPICE simulations of the propagation delay for a 60-CMOS inverters chain with
Fig. 9. Propagation delay versus temperature. Different delay trends versus temperature can be appreciated at different supply voltages.

Fig. 10. Comparison between HSPICE simulations of inverter output switching characteristics and model predictions.

### TABLE I

<table>
<thead>
<tr>
<th>$V_{DD}(V)$</th>
<th>Model (ns)</th>
<th>HSPICE (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>3.69</td>
<td>3.14</td>
</tr>
<tr>
<td>1.3</td>
<td>3.07</td>
<td>2.73</td>
</tr>
<tr>
<td>1.4</td>
<td>2.63</td>
<td>2.42</td>
</tr>
<tr>
<td>1.5</td>
<td>2.30</td>
<td>2.17</td>
</tr>
<tr>
<td>1.6</td>
<td>2.05</td>
<td>1.99</td>
</tr>
<tr>
<td>1.7</td>
<td>1.86</td>
<td>1.84</td>
</tr>
<tr>
<td>1.8</td>
<td>1.70</td>
<td>1.74</td>
</tr>
</tbody>
</table>

$W_p/W_n = 10 \mu m/5 \mu m$ and $L = 0.18 \mu m$ for different power supply values. A 1-GHz input signal drives the first inverter input. This experiment is used to check both the output transition time and the propagation delay model developed since the propagation delay of each stage is dependent on its input rise time $t_{in}$, that is, the output transition time $t_{out}$ of the previous stage. The proposed model provides very good accuracy with respect to HSPICE.

In Table II, we show the mean error of the model proposed and the models in [6] and [7] with respect to HSPICE simulations performed in Figs. 6–8. As can be appreciated, the model proposed provides a better accuracy than previously published models for scaled technologies. Table III compares the computation time used to perform HSPICE calculations of the propagation delay of one single inverter with different analytical models and HSPICE simulations using a Pentium-III 500-MHz pro-
TABLE II
MEAN ERROR FOR PROPOSED MODEL AND PREVIOUSLY PUBLISHED MODELS FOR 0.35-µm AND 0.18-µm TECHNOLOGIES

<table>
<thead>
<tr>
<th>Model</th>
<th>0.35µm technology</th>
<th>0.18µm technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>[7]</td>
<td>9%</td>
<td>5.5%</td>
</tr>
<tr>
<td>[6]</td>
<td>7%</td>
<td>14%</td>
</tr>
</tbody>
</table>

TABLE III
COMPUTATION TIME OF 10⁶ TRANSITIONS WITH PENTIUM-III PROCESSOR@500 MHz

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10 Calc.</td>
<td>3 Days 3h</td>
<td>5‘50”</td>
<td>10‘06”</td>
<td>52”</td>
</tr>
</tbody>
</table>

An accurate analytical expression to compute the propagation delay and the output transition time of CMOS inverters based on charge analysis has been presented. The main effects present in current submicrometer CMOS technologies like the input–output coupling capacitance, carriers velocity saturation effects and short-circuit currents are taken into account in a single analytical expression. The model is compared to HSPICE simulations (using the MM9 model) and to other previously published works for a 0.18-µm and a 0.35-µm process technology reporting a high degree of accuracy. It provides an analytical relationship of the delay to design parameters (device dimensions), input transition time, supply voltage, and temperature. The model is an accurate tool to compute the propagation delay and the input rise/fall time. It provides up to 15% and 100% of improvement in computing time with respect to the models presented in [7] and [6], respectively, and a ×1000 factor with respect to HSPICE simulations.

APPENDIX A
DERIVATION OF OUTPUT RESPONSE OF SINGLE nMOS DISCHARGING TRANSISTOR

The differential equation to be solved is derived from (1)

\[ C_L \frac{dV_{out}}{dt} = -I_n. \]  (A1)

The output voltage evolution depends on the nMOS current expression \( I_n \) and the initial condition at \( V_{out} \). Therefore, there will be different solutions for \( V_{out} \) depending on the input voltage evolution and the operation region of the transistor (linear or saturation). We derive the analytical solution for each case.

CASE 1: nMOS Saturated and \( V_{in} = V_{DD} \): For this case, the nMOS current expression is fixed at \( I_{D0n} \) (the where the subindex \( n \) is referring to the nMOS value of \( I_{D0} \)). Assuming that \( V_{out}(t_0) = V_0 \), then, the solution of (A1) leads to

\[ V_{out}(t) = V_0 - \frac{I_{D0n}}{C_L} (t - t_0). \]  (A2)

CASE 2: nMOS in the Linear Region and \( V_{in} \rightleftharpoons V_{DD} \): This case is more complex since the nMOS current depends on the output voltage. Using the linear expression for the nMOS current (3), then

\[ C_L \frac{dV_{out}}{dt} = - \left( 2 - \frac{V_{out}}{V_{DD}} \right) V_{out} \frac{I_{D0n}}{V_{DD} h}. \]  (A3)

We solve (A3) with the initial condition \( V_{out}(t = t_0) = V_0 \) obtaining

\[ V_{out}(t) = \frac{2V_0}{1 + e^{2(I_{D0n}/C_L V_{DD} h) \lambda(t-t_0)}}. \]  (A4)

CASE 3: nMOS Saturated and the Input is Rising: We model a rising input transition as a linear function of the form

\[ V_{in} = V_{DD} \frac{t-t_0}{t_{in}} \]  (A5)

where \( t_{in} \) is the rise input time. At the beginning of the transition, the nMOS is off and \( V_{out} = V_{DD} \). When the input voltage goes beyond the nMOS threshold voltage (i.e., \( t > t_{in} \)), the nMOS starts to conduct in the saturation region. The initial condition is given by \( V_{out}(t_{in}) = V_{DD} \). We solve (A1) using (4) and (A5) and neglecting channel-length modulation as a first approach

\[ V_{out} = V_{DD} - \frac{I_{D0n}}{C_L} \left( t_{in} - t_{in} \right) \frac{t_{in} - t - t_{in}}{n_{sat} + 1} \]  (A6)

where \( n_{sat} \) is the velocity saturation index of nMOS.

CASE 4: nMOS in the Linear Region and the Input is Rising: For this case, the differential equation (A1) has no analytical solution.

APPENDIX B
DERIVATION OF SHORT-CIRCUIT CHARGE TRANSFERRED AT \( V_{out} = V_{DD}/2 \)

The propagation delay, defined as the time interval between the time points at which the input and output voltages are at \( V_{DD}/2 \), depends on the SCCT during this interval. This charge is defined as \( q_{sc} \) (for a rising input transition, for a falling input transition we use the notation \( q_{dc} \)) and is a fraction of the total SCCT during the transition. We obtain an empirical expression of \( q_{sc} \) as a function of the total SCCT \( Q_{sc} \) developed in [15]. This empirical relationship is developed considering both fast and slow-output transitions.

For slow-output transitions, the input is high before \( V_{out} = V_{DD/2} \) and \( q_{sc} \) can be considered to be equal to \( Q_{sc}/2 \). For fast-output transitions the short-circuit current is maximum when \( V_{out} = V_{DD/2} \) and the SCCT at this time point \( (q_{dc}) \) can be approximated to the half value of the total SCCT \( (Q_{sc}/2) \).

Therefore, since \( q_{dc} \) is dependent on the relative switching speed of the input with respect to the output, we express \( q_{dc} \) as a function of the ratio \( T_{out}/T_{in} \), where parameters \( T_{out} \) and \( T_{in} \) depends on the output and the input transition time and are defined as \( T_{out} = t_d (V_{DD}/2) - t_{in} \) and \( T_{in} = t_{in} - t_{in} \). If \( T_{out}/T_{in} \ll 1 \) the input is considered slow and the SCCT is taken as \( q_{dc} = Q_{sc}/2 \). If \( T_{out}/T_{in} \gg 1 \) we consider that
the input is fast and the expression for the short-circuit charge at \( V_{DD}/2 \) is 
\[ q_{sc}^r = Q_{sc}^r. \]

The parameter \( T_{out} \) is derived from (14) (considering the nMOS transistor in saturation, neglecting the short-circuit charge and assuming that \( V_{in} < V_{DD} \) for simplicity)
\[ T_{out} ≡ t_d \left( \frac{V_{DD}}{2} - t_n \right) \]
\[ \simeq \left[ \frac{C_L V_{DD}}{2I_{D0n}} \left( n_n + 1 \right) \right]^{1/(1+n_n)} \left( t_n - t_n \right)^{n_n/(n_n+1)}. \]

(B1)

Therefore, the relative switching speed \( T_{out}/T_{in} \) is
\[ \frac{T_{out}}{T_{in}} = \left( \frac{C_L V_{DD}}{2I_{D0n}} \left( n_n + 1 \right) \right)^{1/(1+n_n)}. \]

(B2)

An empirical expression is used for \( q_{sc}^r \) as a function of \( T_{out}/T_{in} \). A linear relationship of \( q_{sc}^r \) with respect to the ratio \( T_{out}/T_{in} \) is used for an output response with a value of \( T_{out} \) lower than a threshold time \( bT_{in} \). Otherwise, we take \( q_{sc}^r \) constant and equal to \( Q_{sc}^r \).
\[ \begin{align*}
\begin{cases}
q_{sc}^r = \frac{C_L \left( \frac{V_{DD}}{2} + bT_{in} \right)}{2b} & T_{out} < bT_{in} \\
q_{sc}^r = Q_{sc}^r & T_{out} > bT_{in}
\end{cases}
\end{align*}
\]

(B3)

Parameter \( b \) is taken such that \( t_n + bT_{in} \) is the time at which the pMOS transistor enters in the off state (that is, when \( T_{out} = bT_{in} \), the short-circuit current ceases and we can consider that \( q_{sc}^r = Q_{sc}^r \)). From this restriction, we find parameter \( b \) to be
\[ b = \left( \frac{V_{DD} - V_{THn} + V_{THp}}{V_{DD} - V_{THn}} \right). \]

(B4)

The empirical relationship between \( q_{sc}^r \) and \( Q_{sc}^r \) expressed in (B3) is found to be with a good agreement with HSPICE simulations over a wide variation of inverter configurations and transistor sizes.

The total SCCT is obtained from the model in [15]
\[ Q_{sc}^r = I_{max} \left( \frac{t_{max} - t_n + T_{sc}^r - t_{max} + t_n}{n_n + 1} \right) e^{-\left(2(t_{ov} - t_n)/T_{sc}^r\right)^2} \]

(B5)

where \( t_{ov} \) is the overshoot time (see Fig. 2), \( t_{max} \) is the time at which the short-circuit current is maximum, \( T_{sc}^r \) is the time during which the short-circuit takes place and \( I_{max} \) is the maximum short-circuit current.

The input–output coupling capacitance is taken into account through parameter \( t_{sc} \) in the exponential term while the maximum short-circuit current \( I_{max} \) takes into account both slow and fast-output transitions. The detailed description of all parameters involved are extensively explained in [15].

REFERENCES


An Analytical Model of the Charge Driven by CMOS Buffers

J. L. Rosselló, E. Isern, M. Roca, E. García and J. Segura
Physics Dept., Univ. Illes Balears
Ctra. Valldemossa, km. 7.5, 07071 Palma de Mallorca

Abstract

After showing that it is sufficient to calculate the charge driven by a gate to know its energy consumption, a high efficient simple analytical model to calculate the charge driven by CMOS buffers is presented. The model takes into account the input gate signal delay, transistor sizing and technological parameters as well as the output load capacitance of the buffer. Results of the model are compared with exact numerical calculations of accured MOSFET models showing a good accordance.

1 Introduction

The continuous demand of high reliable products in the microelectronic industry and the increasing miniaturization of integrated circuits feature sizes, have resulted in an enormous interest in the problem of the current and the power estimation. Although current dissipation has been traditionally of interest in CMOS circuit design [1], it has been and is one of the key issues in the development of low power circuits since they gained interest [2].

Several models have been used to calculate the power dissipation of CMOS integrated circuits all of them taking into account a number of factors such as load capacitance, transistor sizing and switching speed. As it is well known there are two primary mechanisms by which power is dissipated during the switching of a CMOS gate: (1) the short-circuit current, and (2) the charge and discharge of the internal nodes capacitance. The energy associated to the short-circuit power dissipation per transition, $E_{sc}$, can be expressed as

$$E_{sc} = V_{DD} \int_{t_{se}} i(t) \, dt$$

(1)

t_{se} being the time during which both P and N transistor nets are simultaneously conducting, and $i(t)$ the current trough both nets. Equation (1) can be integrated leading to $E_{se} = V_{DD}Q_{se}$, $Q_{se}$ being the charge driven into the logic gate from the power supply due to the short-circuit current. The second power dissipation mechanism is inherently a charge transfer.

Therefore the energy dissipated by a CMOS gate can be determined by calculating the charge involved during the input/output transitions.

In this work we present a new model to accurately calculate the charge driven by a CMOS buffer taking into account both the dynamic and the short-circuit current. Traditional approaches used to calculate such a currents are based on the Shockley model [1, 3] or the alpha-power model [4, 5, 6]. These models have in common the consideration of different device operation zones and different equations for each zone. These models (specially the alpha power model) describe the characteristics of MOSFET satisfactorily except in the points on the border between regions. In those points and their neighborhood (specially in the transition from the ohmic zone to saturation) the difference between the model and the experimental results is appreciable. In the continuous model used in this paper as a reference [7] the accordance with measurements is better because all the different transistor operation zones are treated continuously and offer a better adjust.

Other advantages of a continuous model are that
2 Derivation of the model

We start from the continuous expression of the drain current in a MOSFET taken from [7]

\[
I_{ds} = \frac{W \mu \left( n_s - n_d \right) + \frac{1}{2} \kappa (n_s^2 - n_d^2) }{\alpha L} + \frac{W \lambda V_{ds} \kappa n_s}{L} \tag{2}
\]

the second term in (2) stands for the channel modulation length, while the saturation velocity in the channel as well as the effect of the gate voltage on the mobility are not taken into account (i.e. the value of \( \mu \) is constant). Therefore the model must be used for long channel devices. The values of \( n_s \) and \( n_d \) are calculated from

\[
n_c = \frac{\alpha V_t \log(1 + \frac{1}{2} e^{\frac{V_{ds} - V_{th}}{\delta V_c}})}{\kappa} \tag{3}
\]

where the sub-index \( c \) in (3) must be substituted by \( s \) or \( d \) in (2). The constant \( \kappa \) is \( q/C_{ox} \), while the expression for \( V_{th} \) is

\[
V_{th} = V_{th0} + \gamma \left( \sqrt{2\phi + V_{th}} - \sqrt{2\phi} \right) - K_{th} (2\phi + V_{th}) - \sigma V_{ds} \tag{4}
\]

where \( \alpha \) has the following expression

\[
\alpha = 1 + \frac{\gamma g}{2 \sqrt{2\phi + V_{th}}}, \tag{5}
\]

with \( g \)

\[
g = 1 - \frac{1}{p_1 + p_2 \sqrt{2\phi + V_{th}}} \tag{6}
\]

The values used for the constants in equations (2-6) can be found in [7].

We start from this model to obtain a simplified analytical expression for the charge driven by CMOS buffers during input-output transitions. We consider the inverter shown in Fig. 1 with an input signal being a linear ramp between 0 and \( V_{DD} \) (a zero-one or a one-zero transition) with a duration of \( T_{in} \) (input delay). The transient behavior of the circuit is given by the output capacitor \( C_L \). The differential equation describing the circuit is

\[
C_L \frac{dV_{out}}{dt} = i_p - i_n \tag{7}
\]

Therefore, the charge driven by the inverter will depend on the input delay, the value of the output capacitor, the transistor sizes and technological parameters. By integrating (7) for an input zero to one transition we have

\[
Q_n - Q_p = V_{DD} C_L \tag{8}
\]

which shows that it is sufficient to calculate the value of \( Q_n \) or \( Q_p \) to know both values.

To investigate the dependence of the charge driven by the buffer with the parameters considered we proceed to solve numerically equation (7) where \( i_p \) and \( i_n \) are given by (2) and analyze the variation of the charge with the considered parameters.

For a zero-one input transition we can calculate the charge driven by the p-MOS transistor (i.e. the short-circuit current). The charge driven by the n-MOS transistor will be the contribution of the short-circuit current plus the charge stored in the output capacitor (for a one-zero transition the procedure would be analogous).

From the drain current expression of equation (2) we can express both the current for a n-MOS of p-MOS transistor in the form:

\[
i_{p,n} = \frac{W_{p,n}}{I_{p,n}} f_{p,n}(V_{in}, V_{out}, k) \tag{9}
\]

Figure 2: Variation of the charge in the p-MOS transistor vs. its width (a) and length (b).
Figure 3: Variation of the charge in the p-MOS transistor vs. the width (a) and length (b) of the n-MOS transistor.

Figure 4: Comparison of the charge dependence in the p-MOS transistor vs. the input delay for the proposed analytical model and the exact numerical calculation.

In those cases where $i_n >> i_p$ (large output capacitors, and fast input switching) from eq. (7) the time derivative of $V_{out}$ will not vary significantly with the dimensions of the p-MOS device. Therefore $f_p(V_{in}, V_{out}, k)$ will be almost constant with respect to $W_p$ and $L_p$, and the integral of such a function will be proportional to these values as stated in eq. (9). Thus, by integrating (9), the charge into the p-MOS device will be of the form:

$$Q_p = \frac{W_p}{L_p} F_p(T_{in}, W_n, L_n, k)$$

(10)

Given that equation (7) cannot be solved in closed form we derived an empirical dependence of the charge driven by the p-MOS transistor vs. the n-MOS size and the remaining parameters. The final expression is of the form

$$Q_p = \frac{W_p}{L_p C_L} \sqrt{\frac{W_n}{L_n}} (a T_{in} + b T_{in}^2)$$

(11)

The constants $a$ and $b$ depend on the technology being used. We adjusted such a parameters as a function of the voltage power supply, the n-MOS and p-MOS threshold voltages and the gate oxide capacitance per unit of area. The expressions for such a constants are:

$$b = \zeta (V_{DD} - |V_{th_o}| - V') C'$$

$$a = \xi (V_{DD} - |V_{th_o}| - V) (C_{ox} + C') b$$

(12)

(13)

where $\zeta$, $V'$, $V''$, $\xi$ and $C'$ are fitting parameters.

For a one-zero input transition the charge through the n-MOS transistor should be used having the same form than (11) changing the appropriate indexes.

3 Results

To compare the quality of the derived model we plotted the dependence of the charge through the p-MOS transistor vs. the parameters considered (i.e.
the transistor sizes, output capacitor, and input delay) and compare their dependence with the numerical solution of equation (7), using the model of equation (2). Fig. 2 reports the dependence of the charge in the p-MOS transistor vs. the dimensions of such device. Fig. 2.a clearly shows that the dependence of the charge with \( W_p \) is linear, while its dependence on \( L_p \) is inversely proportional.

The dependence of the charge with the n-MOS transistor size is shown in Fig. 3. The curves show a very good agreement between the proposed model and the exact numerical calculation.

Fig. 4 reports the variation of the charge in the device vs. the input delay. It can be also appreciated that the fitting of the model is very accurate.

Fig. 5 compares the dependence of the charge in the device vs. the buffer output capacitor. Again the agreement of the model with the exact calculation is quite good.

Finally Fig. 6.a reports a comparison of the dependence of parameter \( b \) with \( V_{PD} \) calculated using the expression of eq. (12) and the values derived from numerical calculations. As it would be expected, such a parameter is zero until the power supply voltage goes beyond \( |V_{th}| + V_{tn} \). The figure shows that the fitting of expression (12) is very good. In Fig. 6.b we report the variation of the ratio \( a/b \) which is shown to be linearly dependent with the power supply voltage, as expressed in eq. (13).

The comparison of the \( a \) and \( b \) dependence vs. the remaining parameters \( (V_{thi}, V_{thp} \) and \( C_{ox}' \)) between eqs. (12) and (13) and numerical calculations (not shown) is also very accurate.

4 Conclusions

In this work we have derived a semi-empirical model that allows the analytical calculation of the charge driven by a CMOS buffer and therefore its power consumption. The model takes into account the delay of the input signal, the output buffer capacitance and device geometry. Therefore both the short-circuit and charge/discharge currents are taken into account. The goodness of the model has been compared with the exact numerical solution of the equations describing the dynamic of the system and considering an explicit and high accurate single-piece MOSFET model.

References


Transient Current Testing Based on Current (Charge) Integration

I. de Paúl, R. Picos, J.L. Rosselló, M. Roca, E. Isern, J. Segura and C. F. Hawkins*
Physics Department, Balearic Islands University, 07071 Palma de Mallorca, SPAIN
* EECE Dept., The University of New Mexico and Sandia National Labs., Albuquerque, NM 57131, USA

Abstract

We evaluated a technique that uses power supply charge as the test observable. Charge was computed from the measured supply transient current waveform. Data shown that this method is efficient to detect those defects that prevent current elevation (mainly “hard” opens) and therefore represents a valid extension of $I_{DDQ}$

1 Introduction

Current testing (or $I_{DDQ}$ testing) is a widely used method to verify digital CMOS ICs. The main advantage of $I_{DDQ}$ is the high observability that allows a high defect coverage with a relatively small set of test vectors. Opens can leave internal nodes at intermediate voltages, and cause quiescent current elevation that is easily detected with this technique. Despite this, $I_{DDQ}$ has limitations in detecting those defects that prevent current elevation (mainly certain opens).

Several techniques are under development to overcome these limitations and extend the use of current monitoring. Most of the techniques point toward dynamic current testing that monitors the power consumption not only during the quiescent periods, but also during the transients.

Different transient techniques have been proposed. Beasley et al. [1] analyzed the $I_{DD}$ signature when pulsing both $V_{DD}$ and GND levels from an intermediate value to their nominal values. Maki et al. [2] analyzed the change in the transient current peaks to detect defects in the circuit, while Plusquelic [3] and Vinnakota [4] considered both time and frequency domains while monitoring the signal. Cole et al. [5] measured the transient $V_{DD}$ signature to detect defects.

Our work used the power supply charge driven into the circuit during a single transition or a set of transitions as the observable for test. This parameter is directly related to the transient current. The principles of this technique were given in [6].

We present experimental data from a test chip where several defects can be activated. The transient current was measured and the integrated value of the charge computed. This work analyzed the sensitivity of the method to defect location, circuit topology, and input stimuli. The next section describes the experimental methods, Section 3 discusses results, while Section 4 summarizes the work.

2 Experimental Methods

The experiments used an 8-bit shift register in which several mask defects were designed. We evaluated the charge based test technique to detect hard opens (i.e., those that prevent current elevation). Transmission gates were connected to selected nodes of the circuit to emulate the presence of open defects, and were controlled with a 5 to 32 decoder to reduce circuit pin count. Each decoder input, combination activated a unique open defect by turning off a given transmission gate. When a transmission gate is off (i.e., it does not propagate signal), we say that its corresponding open defect is activated. The code 00000 does not activate any defect and was used to obtain a reference measure from the fault-free circuit. Fig 1(a) is a schematic of the 8-bit shift register with the transmission gates labeled for each defect. Fig 1(b) shows a detailed design of the single flip-flops composing the shift register.

The circuit was designed with ES2 n-well dual metal 1.0 μm technology. Separate PADS were used for the 8-bit shift register core $V_{DD}$ supply, the decoder and the input/output PADS. Measurements were taken at the $V_{DD}$ circuit pin to avoid ground noise from input/output PADS and decoder switching. Current was measured by sensing the voltage drop at a very-low inductive MP930 Caddok 300 Ω resistor with a Tektronix P6247 1 GHz bandwidth differential probe connected to a TDS640 Tektronix scope with a 2 Gs/s sampling rate. A dedicated board with differ-
ent supply planes for core and PADS, and two ground planes connected with a low frequency bridge to avoid loop currents [7] was designed to minimize switching (high frequency) noise. Measurements were compared to HSPICE simulations from the extracted layout to evaluate the noise contribution. 50 waveform currents were measured and averaged for each test vector. The charge was computed by numerical integration of the current waveform.

The appropriate test vectors for each defect were generated following the procedure in [6]. Since this technique monitors the transient current, the test vectors were calculated to induce a given transition rather than setting a static input. For this reason we refer to State Transition Test Vectors (STTV) instead of static test vectors. STTVs were chosen to induce a current path in the fault free circuit that is missing in the defective one because of the defect. This depends on the defect site and its effect on the circuit. The mask generated defects shown in Fig 1(a) can be put into three classes:

1. Opens preventing data propagation at the output of one flip-flop (transmission gates labeled as $D_i$).
2. Opens preventing clock propagation from the defect site to the end of the register (transmission gates labeled as $K_i$).
3. Opens preventing clock signal to a single latch without affecting clock propagation to other cells of the register (transmission gates labeled as $H_i$ or $I_i$)

The charge difference between the faulty and the fault-free circuit can be maximized by shifting a STTV into the chain that:

1. Does not change the state of the flip-flops located between the chain input and the defect site, and
2. Induces a state transition to all the flip-flops between the defect site and the output.

This maximizes the charge difference between the fault, free and faulty circuit. Following this rule we derived one STTV for each defect (Fig. 2). The STTV derived for defect $d$ ($STTV_d$) was used to measure the transient current at the rising and falling edge of the clock transition with data settled (Fig. 2). The STTV that complements all bits of each $STTV_d$ (referred to as $STTV_d'$) was also used following rules 1 and 2 described before.

We measured the transient current through the circuit power supply. A difference between $i_{dd}(t)$ and
Figure 3: Transient current waveforms for circuit of Fig.1 measured at the rising and falling clock edges when no defect activated.

$\text{IDDQ}$ is that, for a given vector, the current signature differs if it is measured at the power supply or at the ground rail, because of the charge contribution coming from the logic gate output capacitor loads. Additionally different charge values will be obtained when sensing the current at the rising or falling edge of the clock. This is shown in Fig. 3 where two waveform currents measured at the rising and falling edges of the clock for the same STTV are reported.

3 Results and Discussion

Figure 4(a) shows eight current waveforms measured for $\text{STTV}_K$ (Fig 2) when $K_2, \ldots, K_8$ are activated and for the fault free configuration. Current contributions are due to switching of the flip-flop clock inverters (Fig. 1.b) since $\text{STTV}_K$ does not, change any register data values. The largest current peak was obtained for the fault free configuration because all flip-flops are driven by the clock. A smaller current, peak was obtained when defect $K_8$ was activated. This defect isolates the last flip-flop (FF8 in Fig. 1(a)) from the clock signal, so that the corresponding inverters do not change, and therefore do not contribute to the transient current. Defect $K_7$ isolates flip-flops FF7 and FF8 from the clock further reducing circuit activity that leads to a smaller current peak (Fig. 4). Less activity is observed when the active defect is closer to the input clock signal which is consistent with the successively smaller current waveforms.

Figure 4(b) shows the charge variation with time for each current waveform in Fig. 4(a). The charge corresponds to the area produced by each current waveform. The stable charge value as an observable for test provides an efficient way of comparing current peaks, as current waveform shapes can vary significantly even for the same STTV (Fig 3).

The values of the charge used to distinguish good and bad circuits were estimated taking into account the charge resolution of the measurement system and fabrication process parameter variations.

The resolution measurement is given by the instrumentation as

$$\Delta Q_{\text{meas}} = \frac{t_{av} \Delta V_{\text{res}} + V_{av} \Delta t_{\text{res}}}{R_{\text{sens}}}$$

where $t_{av}$ is the average width of the current peak, $V_{av}$ is the average voltage drop at the resistor, $A_{t_{res}}$ is the minimum time resolution of the oscilloscope, and $\Delta V_{\text{res}}$ its minimum vertical resolution. $R_{\text{sens}}$ is the resistor used for the measurement. In our case $t_{av} = 40 \text{ ns}, V_{av} = 150 \text{ mV}, A_{t_{res}} = 500 \text{ ps}, \Delta V_{\text{res}} = 5 \text{ mV}$, and $R_{\text{sens}} = 300 \text{ } \Omega\pm1\%$, giving a resolution of 916.6 fC.
The second factor to account for is current variation with process deviations. We estimated this variation by simulation using corner parameters supplied by the manufacturer. We obtained $\Delta Q_{par} = 0.83 \text{ pC}$, therefore we choose the charge resolution $\Delta Q_{res} = 1 \text{ pC}$.

If $Q_{no \, def}^{STTV_i}$ is the charge obtained for $STTV_i$ when no defect is activated, and $Q_{def \, j}^{STTV_i}$, the charge obtained when defect $j$ is activated, we consider such defect is detectable if

$$|Q_{def \, j}^{STTV_i} - Q_{no \, def}^{STTV_i}| > \Delta Q_{res} \quad (2)$$

Tables 1, 2, and 3 report the charge values measured for defect classes a, b and c respectively, and the fault free circuit. When a defect, is activated its corresponding STTV in Fig. 2 and $STTV_i$ is used. In those cases where (2) did not hold, values are shown in bold indicating that the defect was not detected.

For class-a defects we used a different STTV for each $D_i$. Results showed that defects $D_8$, $D_7$ and $D_8$ were not detected by their corresponding STTV at the rising clock edge. However $D_5$ and $D_7$ can be detected by the same STTV if current is measured at the falling clock edge. Detection of $D_8$ can only be achieved by measuring the transient current at the falling edge of $STTV_{D_8}$ ($AQ = 1.6 \text{ pC}$). Defect $D_8$ was initially expected to be hard to detect because the activity difference between the fault free and the faulty circuit is due only to a single flip-flop. A measurable charge difference in this case required minimizing circuit activity from FF1 to FF7, while maximizing activity of FF8. The results are consistent because:

1. Since the transient current was measured at $V_{DD}$, more charge occurs when the output load capacitance is charged, i.e when the output changes from zero to one. Therefore we must know which is the largest output capacitance in the cell and induce such a transition at that node.

2. The larger charge transference involved in FF8 will occur when the slave changes from zero to one, as its output capacitance is much larger than the output capacitance of the master cell. The output of FF8 changes from zero to one at the falling clock edge of $STTV_{D_8}$ and the defect is only detected in this case.

All class-b defects were tested with the same STTV. Results show that all defects can be detected at the rising or falling clock transition while using $STTV_K$ or its complement. Charge differences are large for defects close to the clock input and decrease as the defect is close to the output as expected.

Class-c defects were initially thought to be hard to detect, as they isolate the clock signal to a single latch of the chain without affecting the others. Results show that $H_i$ defects are easier to detect than $I_i$ defects, since most of them can be detected at the rising and the falling clock edge. Ha was interesting since it
Table 3: Charge values (in pC) for class c defects.

<table>
<thead>
<tr>
<th>Active defect</th>
<th>$Q_{no, def}$</th>
<th>$Q_{def}$</th>
<th>$Q_{no, def}$</th>
<th>$Q_{def}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$STIV_{H1}$</td>
<td>13.2</td>
<td>10.3</td>
<td>19.8</td>
<td>11.5</td>
</tr>
<tr>
<td>$H_1$</td>
<td>13.2</td>
<td>10.3</td>
<td>19.8</td>
<td>11.5</td>
</tr>
<tr>
<td>$H_2$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$H_3$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$H_4$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$H_5$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$H_6$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$H_7$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$H_8$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$I_1$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$I_2$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$I_3$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$I_4$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$I_5$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$I_6$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$I_7$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
<tr>
<td>$I_8$</td>
<td>15.4</td>
<td>10.2</td>
<td>16.8</td>
<td>10.5</td>
</tr>
</tbody>
</table>

comparatively drives more charge than the rest of defects. The transient waveform revealed that this open causes current elevation. Therefore the quiescent current contributes a higher background to the charge. This defect can be only detected at the rising clock edge. All but one $I_6$ defects ($I_2$) are not covered if the charge is measured at the rising clock edge, but they are detected if charge is measured at the falling clock edge. The result is consistent because these cells change their output value at the falling clock edge.

4 Conclusions

The results show that the charge computed from the transient current supply waveform is a valid observable for test. The method is an extension of $IDPQ$ that allows detection of hard opens, considered as those that do not cause current elevation. The experiments were done in a 8-bit register test chip with 30 mask induced defects and reported a 100% coverage by using this method. Results also show that the heuristic used to derive input stimuli optimized for this method is valid and predicts good coverage.

Acknowledgments

This work has been supported by the Spanish Government Project TIC98-0284.

References


An Analytical Model of CMOS Buffers Power Consumption

J. L. Rosselló, E. Isern, M. Roca and J. Segura

Physics Dept., Univ. Illes Balears
Ctra. Valldemossa, km. 7.5, 07071 Palma de Mallorca
Phone: +34 971 173 211 Fax: +34 971 173 426 email: dfsjrs4@ps.uib.es

Abstract

An analytical model of CMOS buffers consumption is presented. It considers the input slope, transistor sizing and technological parameters as well as the buffer output load capacitance. Excellent agreement with HSPICE level 6 simulations and its simplicity allows its application to the analysis of large circuits.

1. Introduction

The continuous demand of high reliable products in the microelectronic industry and the increasing miniaturization of integrated circuits feature sizes, motivated an increase interest in power estimation. As it is well known there are two primary mechanisms by which power is dissipated during the switching of a CMOS gate: (1) the short-circuit current, and (2) the charge and discharge of the internal nodes capacitance. Many techniques have been proposed to calculate power consumption by simulation [1]-[3] or using closed form models [4]-[8]. Most models consider a linear input transition, some of them consider an output load and calculate the dependence of consumption with device geometry.

Veendrick [4] obtained an expression for the short-circuit power dissipation using the Schichman-Hodges MOSFET model for unloaded buffers. Equal transconductance, threshold voltage for both MOSFET and a symmetric short-circuit current shape were considered. The model is only applicable to long-channel devices and symmetric-devices buffers. Veendrick’s work concludes that short-circuit power dissipation is less than 20% of total power dissipation when input and output transition time are equal.

Hedenstierna et al. [7] calculated the short circuit current using the Schichman-Hodges model (long channel devices) for a buffer with an output capacitor C_L and non symmetrical MOSFET devices. The short-circuit current was neglected respect to the current from the external capacitor. Such an approximation overestimates the short circuit when its contribution is similar to the current from the output capacitor.

Sakurai et al. derived the short-circuit dissipation for long and short channel devices using the Alpha-Power Law model [5], while [9] derived the same formula from a different MOSFET model. Unloaded gates, linear input transition and a symmetric current waveform around \( T_{in} \) (\( T_{in} \) being the input transition time) were assumed. The short-circuit dissipation was a function of supply voltage, MOSFET threshold voltage, input transition time, and four empirical parameters from the Alpha-Power Law model. This empirical parameters have non physical meaning and must be extracted from experimental data.

The dynamic short-circuit dissipation is difficult to describe analytically because of the complexity of the non-linear differential equations that describe the system. A closed form expression without reducing significantly accuracy requires simplifications, specially when short-circuit contribution is comparable to the dynamic transient component.

We present an analytical model for short-circuit dissipation of CMOS buffers given that an analytical description is preferable instead of numeric calculations to understand the power dependence with technological parameters and geometry. It also provides faster calculations specially for large circuits.

2. Derivation of the model

The following assumptions are used:

1. The CMOS logic inverter transient response is due to an output capacitor (C_L) that accounts for all the gates driven by the buffer.
2. The inverter input is linear with an input transition time $t_r$.

Using the MOSFET model in [10] the current on each transistor is:

\[ I_{ds} = \begin{cases} 0, & \text{if } V_{gs} < V_{th} \\ \beta \left( V_{gs} - V_{th} - \alpha \frac{V_{ds}}{2} \right), & \text{if } V_{gs} < \frac{V_{th}}{\alpha} \\ \beta \frac{V_{ds}^2}{2\alpha}, & \text{if } V_{ds} > \frac{V_{th}}{\alpha} \end{cases} \quad (1) \]

Where $V_{gs} = V_{gs} - V_{th}$, $\beta = C_{ox} B \frac{W_{eff}}{L_{eff}}$, $C_{ox}$ is the gate oxide capacitor per unit area, $\mu$ the carrier mobility, $W_{eff}$ the effective channel width and $L_{eff}$ the effective channel length. $V_{gs}$, $V_{ds}$ and $V_{th}$ are the gate, drain and threshold voltage of transistor respectively and $\alpha$ is described by:

\[ \alpha = 1 + \frac{g \gamma}{2 \sqrt{\phi}} \quad (2) \]

Where $\gamma$ is the body effect factor, $\phi$ is the surface inversion potential and, $g$ is:

\[ g = 1 - \frac{1}{1.744 + 0.8364 \phi} \quad (3) \]

The effective channel length and width can be expressed as $L_{eff} = L + X_L - 2L_D$, $W_{eff} = W + X_W - 2W_D$, where $L$ and $W$ are the geometrical channel length and width, $X_L$ and $X_W$ the length and width bias that account for masking and etching effects, $L_D$ is lateral diffusion into the channel from source and drain, and $W_D$ is the lateral diffusion into channel width from bulk.

The equation describing a CMOS logic inverter with an output capacitor $(C_L)$ is:

\[ C_L \frac{dV_{out}}{dt} = i_p - i_n \quad (4) \]

Were $V_{out}$ is the output voltage, $i_p$ and $i_n$ are the currents across the p-MOS and the n-MOS transistors as described in (1).

Integrating (4) for a zero to one input transition we obtain:

\[ C_L V_{DD} = Q^{0-1}_{n} - Q^{0-1}_{p} \quad (5) \]

Where $Q^{0-1}_{n}$ and $Q^{0-1}_{p}$ are the charge driven by both p-MOS and n-MOS transistors respectively for a zero to one input transition. The n-MOS transistor drives charge from the output capacitor $(C_L V_{DD})$ and from the p-MOS transistor (short-circuit charge). Integration of (4) for a one to zero input transition leads to:

\[ C_L V_{DD} = Q^{1-0}_{n} - Q^{1-0}_{p} \quad (6) \]

In this case, part of the charge driven by the p-MOS transistor goes to the capacitor output, and the rest goes through the n-MOS transistor (short-circuit charge).

The energy dissipated by a CMOS buffer can be expressed as:

\[ E = \int V_{DD} i_p \, dt \quad (7) \]

For one cycle period the total energy dissipated is:

\[ E = \int_{T} V_{DD} i_p \, dt = V_{DD} Q^{0-1}_{p} + V_{DD} Q^{1-0}_{n} \quad (8) \]

Substituting (6) in (8), the total power consumption of a CMOS buffer is:

\[ E = C_L V_{DD}^2 + V_{DD} Q^{0-1}_{p} + V_{DD} Q^{1-0}_{n} \quad (9) \]

The terms $V_{DD} Q^{0-1}_{p}$ and $V_{DD} Q^{1-0}_{n}$ are due to the short circuit consumption and have a complex dependence on several parameters as the input rise (or fall) time, geometrical and technological parameters of CMOS transistors and the output load capacitor. We will derive an expression of the short circuit consumption for a zero to one input transition. The case of a one to zero input transition, would be equivalent.

The input transition is linear, with an input rise time $t_r$:

\[ V_{in} = \begin{cases} 0, & \text{if } t < 0 \\ \frac{V_{DD}}{t_r} t, & \text{if } 0 < t < t_r \\ V_{DD}, & \text{if } t > t_r \end{cases} \quad (10) \]

Calculation of $Q^{0-1}_{p}$ requires an expression for the output voltage that can be obtained solving (4). Before the input transition takes place, the output voltage is at $V_{DD}$ and there is no consumption because the n-MOS transistor is off. We calculate the circuit charge transference by integrating $i_p$ over the input transition time.

\[ Q^{0-1}_{p} = \int_{0}^{t_r} i_p(V_{in}(t), V_{out}(t)) \, dt \quad (11) \]

At the beginning of the input transition the p-MOS device is in the linear region while the n-MOS is saturated. Assuming that the current through the n-MOS transistor is much larger than the current through the p-MOS and substituting (1) in (4) we have:

\[ C_L \frac{dV_{out}}{dt} = -\frac{\beta_n}{2\alpha_n} \left[ \frac{V_{DD}}{t_r} - V_{thn} \right]^2 \quad (12) \]

Setting $t_n$ and $t_p$ as the time at which the n-MOS transistor starts to conduct and the p-MOS transistor finishes conduction respectively we have:
\[ t_n = t_r \frac{V_{thn}}{I_{DD}} \]
\[ t_p = t_r (1 - \frac{V_{thp}}{V_{DD}}) \]

Since when the n-MOS starts to conduct the output voltage is high, then the initial condition for (12) is \( V_{out}(t_n) = V_{DD} \). Then (12) leads to:
\[ V_{out}(t) = V_{DD} - \frac{\beta_n V_{DD}^2 (t - t_n)^3}{6a_s C_L I^2} \]

Expanding (14) near the mid point between \( t_n \) and \( t_p \) to the first order (14) leads to:
\[ V_{out} = \frac{\beta_n V_{DD}^2}{8C_L t} + \frac{\beta_n (V_r + 3V_{th}) V_{DD}^2}{2a_s C_L V_{DD}} + V_{DD} \]

Where \( V_r = V_{DD} - V_{th} \).

Substituting \( V_{out} \) into p-MOS current expression (1) requires determining the p-MOS operation zone at the maximum current. Assuming that the maximum current takes place near the mid point between \( t_n \) and \( t_p \), such a maximum can only occur when the p-MOS device is in the linear region. While the p-MOS is in saturation no maximum can occur because \( V_{gs} \) decreases which implies an always decreasing \( I_p \) current. Therefore, the p-MOS saturates always after the maximum current takes place.

Substitution of (15) in (1) leads to an expression of the form:
\[ I_p(t) = I_{p\max} + G(t - t_{max})^2 \]

Which corresponds to an inverted parabola with its maximum being the maximum current.

Writing (15) as \( V_{out} = At + B \), then:
\[ I_{p\max} = \frac{\beta_n (V_r + V_{DD} - 2V_{th}) V_{DD}^2}{2a_s C_L V_{DD}} \]
\[ t_{max} = \frac{V_{DD} A + V_{DD} A - V_{DD} A}{2a_s C_L V_{DD}} \]
\[ G = \frac{\beta_n V_{DD}^2}{8a_s C_L I^2} \]

Finally, the short circuit current transference is:
\[ Q_{sc} = \int_{t_n}^{t_p} I_1 dt + \int_{t_n}^{t_p} I_p^m dt + \int_{t_p}^{t_{max}} I_2 dt \]

Substituting (17), (18) and (20) in (21) we obtain an expression for \( Q_{sc} \) as:
\[ Q_{sc} = \frac{4}{9} \alpha_n \frac{\rho}{I} \frac{24 \alpha_n + \alpha_p \frac{t_n}{t_r}}{8 \alpha_n + \alpha_p \frac{t_n}{t_r}} \frac{t_r}{t_p} \]
\[ Q_r = \frac{\beta_n t_r V_D^3}{6a_n V_{DD}} \]
\[ \tau_r = \frac{C_m V_{PD}}{\beta_n V_D^2} \] (23)

Equation (22) was derived assuming \( I_p \ll I_n \), and is therefore inaccurate when the short circuit charge is comparable to the charge from the output capacitor \( (C_L V_{DD}) \). When \( Q_{sc} = C_L V_{DD} \), the current from the p-MOS transistor can not be neglected when solving (4). We introduce a correction in (22) to obtain a valid expression for all cases as follows:

In a CMOS inverter, if \( \beta_n \) is kept constant and \( \beta_p \) increases, then the short-circuit current should saturate to a value determined by the n-MOS device because of its finite conductance. In this situation, a limiting value for the charge would be:

\[ Q_{max}^p = \int_{t_n}^{t_f} I_n^{sat}(t) dt = \frac{\beta_n t_r V_D^3}{6a_n V_{DD}} \] (24)

Since (22) is linear with \( \beta_p \), it never saturates to a maximum. Therefore (22) is corrected to include a saturation value determined by the n-MOS conductance:

\[ Q_{sc}^{0=1} = \frac{Q_{max}^p Q_{sc}^p}{Q_{max}^p + Q_{sc}^p} \] (25)

For a one to zero input transition a similar expression can be derived:

\[ Q_{sc}^{1=0} = \frac{Q_{max}^n Q_{sc}^n}{Q_{max}^n + Q_{sc}^n} \] (26)

where \( Q_{sc}^p \) and \( Q_{sc}^n \) are:

\[ Q_{sc}^p = \frac{\beta_p t_f V_D^3}{6a_p V_{DD}} \] (27)

\[ Q_{sc}^n = \frac{4}{9a_n Q_f} \left( \frac{24\alpha p + \alpha_n \frac{t_f}{t_j}}{(8\alpha p + \alpha_n \frac{t_f}{t_j})(32\alpha p + \alpha_n \frac{t_f}{t_j})} \right) \frac{t_f}{t_f} \] (28)

Being \( Q_f = \frac{\beta_p t_f V_D^3}{6a_p V_{DD}} \) and \( t_f = \frac{C_m V_{PD}}{\beta_p V_D^2} \)

Once the short-circuit charge is known, (9) gives the energy dissipated in one cycle as:

\[ E = C_L V_{DD}^2 + V_{DD} \left( \frac{Q_{max}^n Q_{sc}}{Q_{max}^n + Q_{sc}} + \frac{Q_{max}^p Q_{sc}}{Q_{max}^p + Q_{sc}} \right) \] (29)

3. Results

We compared the energy dissipated by a CMOS inverter in one clock cycle from equation (29) with HSPICE simulations and other models previously published in the literature. Since many works only consider the energy due to node capacitance, neglecting short-circuit contributions [11]-[14], we also include values of energy computed as \( E = C_L V_{PD}^2 \) to evaluate the goodness of such an approximation.

HSPICE simulations were carried out using level 6 parameters supplied by ES2 for its digital 0.7 \( \mu \)m dual metal n-Well technology. We also include the results from Veendrick [4], Sakurai [5] and Hedenstierna [7].

Energy is plotted versus output capacitor load, p-MOS transistors length and width, power supply and input rise time in Figs (2-6).

Fig (2) shows the energy dependence with the output load for all the models mentioned before. Capacitor values are given in units of the minimum output load \( (C_{ref}) \) computed as the output parasitic capacitor for a minimum size unloaded buffer. Inverter size is \( L_n = L_p = 0.8 \mu m \), while \( W_p = 20 \mu m \) and \( W_n = 10 \mu m \). Supply voltage is 5V while \( t_r = t_f = 0.5 \)ns. Fig (2) shows that the short-circuit current is not negligible with respect to the output capacitor dissipation \( (C_L V_{PD}^2) \). The model from Veendrick clearly overestimates dissipation. The other models ac-
accurately described current dissipation while the model presented shows the best fitting.

The plot of Fig (3) shows energy dissipated vs. p-MOS channel length varying from 2\,\mu\text{m} to 10\,\mu\text{m} (output load is equal to $C_{\text{ref}}$ and same values than in previous plot are given for p and n-MOS transistor width and rise time). Veendrick, Hedenstierna and Sakurai models overestimate energy dissipation and both Sakurai and Hedenstierna do not describe the energy dependence with $L_p$, reporting an almost flat curve. The model proposed in this work better describes energy dependence.

Fig 4 plots the energy versus p-MOS channel width. The model fits well with HSPICE level 6 calculations. Veendrick’s model overestimates power consumption as in previous cases, while Sakurai et al. model describes accurately the energy for wide transistors although the slope of the function is not fitted. Hedenstierna’s model overestimates dissipation for narrow transistors while underestimates dissipation for wide transistors. The model developed in this work shows the better fitting to HSPICE simulations although a better fitting is obtained for narrow transistors.

The reason of this is that for wide transistors the non linear part of equation (4) is greater than for narrow transistors.

Fig 5 is a plot of the energy dissipated vs. the power supply from 2\,V to 5\,V. The short-circuit power consumption goes to zero when the power supply goes below $(V_{th_p} + |V_{th_n}|)$. When $V_{DD}$ is below this voltage, the power consumption is due to dynamic transient dissipation ($C_LV_{DD}^2$), since both n-MOS and p-MOS devices never conduct simultaneously. As in the previous cases the model developed in this work better describes the energy dependence. Other model with an accurate description of dissipation is Hedenstierna’s model.

Finally, Fig 6 shows the energy dependence with input slope. The transient power dissipation does not change with $t_r$; since it corresponds to the charge stored in the output capacitor. The better approach to HSPICE level 6 is achieved by the model described in this work. The other models report inadequately fitting and a worst divergence. As the rise time increases, the power consumption increases because the time during which both transistor are conducting is larger. In the previous graphs, the total power consumption is greater than transient consumption traditionally used to compute power dissipation as in [11]-[14], that neglect short-circuit dissipation. The dependence of total power consumption with rise (or fall) time is of high interest for power minimization (see [15]).

4. Conclusions

An accurate model of short-circuit dissipation of CMOS buffers have been developed. Other models present in the literature leads to worst approximations to the consumption (in the case of the model of Sakurai et al., dissipation is a function of several empirical models like velocity saturation index ($\alpha$) or the threshold voltage used...
(V_TH) that have to be obtained from the transistor behavior). As has been expressed before, numerical values of power consumption are fitted well and also slope of the model as a function of several technological and geometrical parameters is a good approximation to real slope.

Simplicity of the model enables it use in power minimization and also allows a faster estimation of power consumption on CMOS IC's than HSPICE simulations. Other models (except for the Veendrick expression) have a complex function of technological parameters and have a worst approach to dissipation computed by HSPICE simulations.

As future work, model can be extended to multiple-input CMOS gates and to the submicronic range, including the overshoot current, very significant in submicronic CMOS.

References


A Physical Modelation of the Alpha-power Law MOSFET Model

José Luis Rosselló and Jaume Segura
Physics Dept. Balearic Islands University
Ctra. Valldemossa, km 7.5, 07071 Palma de Mallorca
Phone: +34 971 172 506 Fax: +34 971 173 426
e-mail:{dfsjs4,dfsjsf4}@ps.uib.es

Abstract
A simple physically-based expression of the alpha-power law MOSFET model is presented. A formulation for the empirical parameters $V_{TH}$ and $\alpha$ is given providing a physical meaning to the model and reducing parameter extraction efforts. The obtained expression is compared to HSPICE simulations (level 50) for a $0.35\mu m$ technology with a $2\%$ mean error in the drain saturation current.

1. Introduction
Shockley’s MOSFET model [1] is the simplest form to express the I-V characteristics of a MOSFET transistor. However, this traditional square-law model ignores the carriers’ velocity saturation effects which becomes prominent in actual short-channel devices. Due to its mathematical simplicity, the alpha-power law MOSFET model [2] is a widely used drain current model to analytically derive gate parameters like delay, output slew time or power dissipation [3]-[8]. The model takes into account the velocity saturation effect, which becomes dominant in short-channel devices, giving an excellent intuitive understanding of the relationship between the drain saturation current and the gate voltage. However, the empirical nature of the parameters included in the model (that must be extracted from measurements or SPICE simulations) cannot provide a real physical relationship between model parameters and carriers’ saturation effects therefore requiring additional computational effort for parameter extraction. Recently a physical alpha-power law MOSFET model has been presented [9]. A physical description is obtained at the expenses of a more complicated relationship between the drain saturation current and the gate-voltage. For this model, the $\alpha$ coefficient takes different values (from 3/2 for short-channel to 2 for long-channel) than in model developed by Sakurai et al. (that goes from 1 for short-channel to 2 for long-channel). The complexity of the expressions obtained in [9] prevent the application of model parameters to many handy formulas based on the traditional alpha-power law model.

In this work we present a physical description of the alpha-power law MOSFET model by providing a simple relationship between the model parameters and physical parameters. This model and the traditional alpha-power law are compared to the MOSFET model 9 developed by Philips laboratories [10, 11] for a $0.35\mu m$ technology showing a mean error of $2\%$.

2. Alpha-power law MOSFET models
The drain current for the alpha-power law MOSFET model [2] is expressed as:

$$I_D = \begin{cases} 
0 & (V_{GS} \leq V_{TH}) \\
\frac{V_{DS}^{\alpha}}{\sqrt{1 - V_{DS}^{\alpha}}} & (V_{DS} < V_{D0}) \\
I_{D0} & (V_{DS} \geq V_{D0})
\end{cases} \quad (1)$$

with

$$I_{D0} = I_{D0} \left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}}\right)^{\alpha} \quad (2)$$

$$V_{D0} = V_{D0} \left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}}\right)^{\beta}$$

where $V_{GS}, V_{DD},$ and $V_{TH}$ are the gate voltage, supply voltage and the threshold voltage respectively. $V_{D0}$ is the drain saturation voltage for $V_{GS} = V_{PD}$ and $I_{D0}$ is the drain current at $V_{GS} = V_{DS} = V_{DD}$. 
The parameter $\alpha$ is the velocity saturation index that takes a value between 2 (long-channel devices) and 1 (short-channel). The two parameters $I_D0$ and $V_{DD}$ are extracted from the I-V data, while $V_{TH}$ is obtained by solving:

$$f(V_{TH}) = \log \left( \frac{I_D}{I_D0} \right) \log \left( \frac{V_G - V_{TH}}{V_{G2} - V_{TH}} \right)$$

$$- \log \left( \frac{I_D}{I_D0} \right) \log \left( \frac{V_G - V_{TH}}{V_{G2} - V_{TH}} \right) = 0$$  \hspace{1cm} (3)$$

where $(V_{G1}, I_{D1})$, $(V_{G2}, I_{D2})$ and $(V_{G3}, I_{D3})$ are three points from the $V_G - I_D$ plot. Once $V_{TH}$ is determined, $\alpha$ is obtained as:

$$\alpha = \frac{\log \left( \frac{I_{D1}}{I_{D2}} \right)}{\log \left( \frac{(V_{G1} - V_{TH})}{(V_{G2} - V_{TH})} \right)}$$  \hspace{1cm} (4)$$

This model is entirely empirical providing a handy formula that can be easily used to solve systems of more than one transistor like inverters or complex gates as reported in many works [3]-[8].

Equation (3) provides higher threshold voltage than other physically-based expressions. As an example we show a comparison between the threshold voltage given by eq. (3) and the threshold voltage expression from the MOSFET model 9 in Fig. 1 for a 0.35$\mu$m technology. Expression (3) overestimates the threshold voltage with a maximum error of 14%. A lack of physical meaning (like the threshold voltage roll-up present) can be clearly appreciated in Fig. 1.

The physically-based alpha-power law MOSFET model [9] provides a physical meaning of the $\alpha$-power relationship between the drain current and the gate voltage. Such model modifies the saturation current expression $I_{DS0}$ by using:

$$I_{DS0} = \frac{\beta V_{DO}}{1 + \theta (V_G - V_{TH})} \left( \frac{V_G - V_{TH}}{V_{DD} - V_{TH}} \right)^\alpha \hspace{1cm} (5)$$

with

$$V_{DSAT} = E_C L \left\{ \frac{1}{1 + \frac{2}{E_C L} \left( \frac{V_G - V_{TH}}{\eta} \right)} - 1 \right\}$$

For this model $\alpha$ has the expression:

$$\alpha = \frac{\log \left( \frac{2V_{DSAT}V_{DD} - \eta (V_G - V_{TH})}{V_{DD} - V_{TH}} \right)}{\log(2)}$$  \hspace{1cm} (6)$$

where $V_{DSAT} = \frac{V_{DSAT}}{V_{DD} - V_{TH}}$. Using eq. (6) $\alpha$ is 3/2 for short-channel and 2 for long-channel.

The difference between eq (6) and the traditional alpha-power law is due to the relationship used between $I_{DS0}$ and $V_G - V_{TH}$ which is not as simple as the $\alpha$-power expression (5). Therefore, such model cannot be applied to many works based on the traditional alpha-power law model.

3. Model derivation

The $\alpha$ expression developed in this work is derived from the drain saturation current of MOS model 9 [10, 11], expressed as:
\[ I_{ds} = \beta \left( V_{gs} V_{dsat} - \frac{2 V_{dsat}^2}{2} \right) \]  (7)

Where \( V_{gs} = V_{g} - V_{th} \) (\( V_{gs} \) is the gate to source voltage and \( V_{th} \) is the threshold voltage), \( \beta = \mu_{eff} C_{ox} \frac{W}{L_{eff}} \), with \( C_{ox} \) being the gate oxide capacitor, while \( W_{eff} \) and \( L_{eff} \) are the effective channel width and length respectively. \( \mu_{eff} \) is the effective carriers mobility expressed as:

\[ \mu_{eff} = \frac{\mu_0}{1 + \theta_1 V_{g} (1 + \theta_2 V_{dsat})} \]  (8)

\( \theta_1 \) and \( \theta_2 \) are the gate field and drain bias mobility reduction coefficients respectively. \( V_{dsat} \) is the saturation voltage given by:

\[ V_{dsat} = \frac{2 V_{sat}}{\eta \left( 1 + \frac{2 \eta_{sat}}{1 + \eta_{sat}} \right)} \]  (9)

with \( \eta \) expressed as:

\[ \eta = 1 + \frac{\lambda_4}{(PHIB)^2} \left[ K + \frac{(K0 - K) VSBX^2}{VSBX^2 + (\lambda_3 V_{g})^2} \right] \]  (10)

where \( \lambda_4 = 0.3 \), \( \lambda_3 = 0.1 \), \( KO \) and \( K \) are the substrate sensitivity when depleting surface and bulk doping respectively, \( PHIB \) is the surface potential for strong inversion and \( VSBX \) is a voltage of transition between \( KO \) and \( K \). As a value of \( \eta \) we choose \( \eta = \left( V_{g} = (V_{DD} - V_{TH}) \right)/2 \). We can easily obtain \( I_{D0} \) and \( V_{DD} \) from equation (7) and (9) by using \( V_{DD} = V_{dsat} (V_{gs} = V_{DD}) \) and \( I_{D0} = I_{ds} (V_{gs} = V_{DD}) \).

As was shown in the previous section, the value of the threshold voltage is obtained by solving (3) numerically. In this work we consider a physically-based expression for the threshold voltage:

\[ V_{th} = V_{TH0} - \gamma_0 V_{DD} \]  (11)

where \( \gamma_0 \) accounts for the DIBL effect and \( V_{TH0} \) is taken as:

\[ V_{TH0} = V_{TO} + S_L \left( \frac{1}{W_{ref}} - \frac{1}{W_{ref}} \right) + S_{L2} \left( \frac{1}{W_{ref}} - \frac{1}{W_{ref}} \right) + S_W \left( \frac{1}{W_{ref}} - \frac{1}{W_{ref}} \right) \]  (12)

where \( V_{TO} \) is the threshold voltage with no substrate or drain bias, while \( S_L \) and \( S_{L2} \) account for the length dependence of the threshold voltage, and \( S_W \) is the width dependence parameter of the threshold voltage. Finally, \( L_{ref} \) and \( W_{ref} \) are a reference channel length and with for a given technology. All the parameters considered are included in the MOS-FET model 9 process parameters set and account for several short-channel effects as the drain-induced barrier lowering or the threshold voltage roll-up.

A simple and useful form to express the \( \alpha \) exponent in terms of velocity saturation coefficients is:

\[ \alpha = 1 + \frac{\mu_{eff}}{\mu_0} \left( \frac{V_{gs} - V_{DD}}{V_{dsat}} \right) \]  (13)

where \( \mu_{eff} \) is the effective mobility coefficient given by (8). Eq. (13) is very effective in describing the relationship between the saturation current and the gate voltage. Eq. (13) also bounds the \( \alpha \) values.
like the original Alpha-power model since for long-channel transistors, \( a = 2 \) (since \( \mu_C = \mu_0 \)) while for short-channel this expression tends to 1. Fig. 2 shows a comparison between equation (13) and (4) for various values of channel length showing that there is a closed relationship between the two formulations.

4. Model results

We compared the alpha-power law MOSFET model for the empirical and the physically based parameters presented here with accurate MOSFET model 9 calculations for a 0.35\( \mu \)m process technology.

In Figs. 3 and 4 we plot the saturation current of n-MOS and p-MOS transistors with \( L_{eff} = 0.35\mu \)m and various \( W_{eff} \) values (from 1\( \mu \)m to 6\( \mu \)m). As can be seen, both expressions fit data with high accuracy (there is a 2.7\% and 1.3\% mean error with respect \( I_D \) for the model present here and the traditional alpha-power law model respectively).

Figs. 5 and 6 plot the saturation current of n-MOS and p-MOS transistors with a channel length of 3.0\( \mu \)m and various values of \( W_{eff} \) (from 1\( \mu \)m to 6\( \mu \)m). Both expressions fits data with nearly the same accuracy (1.8\% and 1\% mean error for the model present here and the traditional alpha-power law model).

Identical plots have been performed for 0.5\( \mu \)m, 0.6\( \mu \)m, 0.8\( \mu \)m, 1.0\( \mu \)m, and 1.5\( \mu \)m of channel length for both p-MOS and n-MOS devices showing a similar accuracy. Figs. 7 and 8 shows such results by plotting HSPICE predictions vs. models calculations. Mean and maximum errors (calculated respect \( I_D \) as \( 100 \times \frac{I_{max}}{I_{DD}} \)) are shown in Table 1.

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Errors} & \text{Max.} & \text{Mean.} & \text{Max.}[2] & \text{Mean.}[2] \\
\hline
0.35\mu m & 3.1 & 2.7 & 2.5 & 1.3 \\
0.5\mu m & 2.1 & 2.6 & 1.7 & 1.2 \\
0.6\mu m & 2.6 & 2 & 2 & 1.1 \\
0.8\mu m & 2.6 & 1.9 & 1.9 & 1.0 \\
1.0\mu m & 2.5 & 1.9 & 1.8 & 0.9 \\
1.5\mu m & 2.4 & 1.8 & 1.8 & 0.9 \\
3.0\mu m & 2.1 & 1.8 & 1.9 & 0.8 \\
\hline
\end{array}
\]

Many works have been done based on such model obtaining useful formulas to predict delay, slew time or power dissipation in CMOS gates. Based on his MOSFET model, Sakurai et al. developed a simple formula for calculating the delay of a CMOS inverter [2]. This formula is given by:

\[
t_{pHL}, t_{pHL} = \frac{C_L V_{DD}}{2I_{DD}} + \left( \frac{v_T + \alpha}{1 + \alpha} - \frac{1}{2} \right) t_r \quad (14)
\]

where \( t_r \) is the input rise/fall time and \( v_T = V_{TH}/V_{DD} \). We can observe the difference of using in (14) the traditional parameters for \( V_{TH} \) and \( \alpha \) or equations (11) and (13). Fig. 9 shows the variation of the delay time of a CMOS inverter with input rise and fall time. As can be appreciated, both formulations of the alpha-power law model leads to the
same order of accuracy (within a 3% of mean error). Fig. 10 shows delay variation with channel length for a fixed n-MOS and p-MOS conductance \( W_n/L_n = 0.5/0.35 \) and \( W_p/L_p = 1.0/0.35 \).

5. Conclusions

A new formulation of the Alpha-power law MOSFET model have been presented. Model takes into account physical parameters to describe the carriers’ saturation effects in a simple way. Simplicity and accuracy of the preliminary model developed by Sakurai et al. have been maintained. The formulation can be included in many handy formulas developed from the alpha-power law MOSFET model [3]-[8] without loose of accuracy and allowing a direct relationship with process parameters.

Acknowledgments

This work has been supported by the Spanish "Comisión Interministerial de Ciencia y Tecnología" under the project CICYT-TIC98-0284.

References


Figure 7 - Comparison between HSPICE lev. 50 and sakurai's prediction for the n-MOS and p-MOS currents for a set of devices between 0.35μm and 3.0μm of channel length.

Figure 8 - Comparison between HSPICE lev. 50 and model prediction for the n-MOS and p-MOS currents for a set of devices between 0.35μm and 3.0μm of channel length.
Figure 9 – Time delay variation with input rise/fall time for a CMOS inverter

Figure 10 – Time delay variation with channel length for a CMOS inverter


A simple power consumption model of CMOS buffers driving RC interconnect lines

J.L. Rossello and Jaume Segura
Physics Dept., Balearic Islands University, Cra. Valdemossa km 7.5
07071 Palma de Mallorca, Spain
e-mail: {j.rossello,jaume.segura}@uib.es

Abstract. We present a simple and accurate model to compute the power dissipated in sub-micron CMOS buffers driving RC interconnect lines. The expression obtained accounts for the main effects in current sub-micron CMOS technologies as carrier velocity saturation effects, input-output coupling capacitor, output load, input slew time, device sizes and interconnect resistance. Results are compared to HSPICE simulations (level 50) and other previously published works for a 0.18µm and a 0.35µm technology showing significant improvements.

1 Introduction

As IC technology fabrication processes scale down, new physical effects must be considered when analyzing and modeling CMOS circuits. Velocity saturation due to high internal electric fields, input-output coupling capacitor effects due to the narrower gate oxide thickness, or on-chip interconnect resistance (that does not scale down with feature size) are some of the effects that must be considered to obtain accurate models.

A large fraction of the power dissipated in today VLSI ICs is due to the I/O drivers and busses and the clock distribution network, which are based on inverter gates. Hence, the analytical description of power dissipated in a CMOS inverter is of increasing importance for CMOS ICs power estimation. It is well known that power dissipation in CMOS circuits has a dynamic and a static component. The dynamic dissipation (defined as transient energy) is due to the charge/discharge of the gate output load, and to the short-circuit current due to the supply-ground conducting path created during the transition [1].

Several works have been focussed on modeling the short-circuit power consumption of CMOS buffers. Veenckrick [1] and Sakurai et al. [2] obtained analytical models for unloaded buffers using a long-channel and a short-channel MOSFET model respectively. Nose et al. [3] derived a model for submicron CMOS buffers driving a single capacitor. This model does not take into account both the input-output coupling capacitor (IOCC) effects nor the line resistance at the buffer output. Turgis et al. [4] derived an expression for the short-circuit power taking into account the IOCC and neglecting the resistance at the buffer output. Nikolaidis et. al [5], obtained an expression of the short-circuit dissipation for buffers driving long interconnect lines using the α–power law MOSFET model. The analysis was based on the π-model of an RC load and developed for sub-micron devices. This model takes into account the IOCC but neglects the short-circuit current contribution when computing the output waveform.

In this work we present an accurate and simple model to compute the power consumption of CMOS buffers accounting for the main effects in submicron CMOS technologies as the IOCC and the line interconnect resistance, of increasing importance in current submicron ICs. Closed-form expressions for power estimation are obtained
avoiding time-consuming numerical procedures. The model is compared to HSPICE simulations (level 50) and to other previously published models for a 0.18\(\mu\)m and a 0.35\(\mu\)m technologies showing a significant improvement in terms of accuracy. The rest of this work is organized as follows: in section 2 we obtain an expression for the transient energy. Section 3 derives the short-circuit energy component and Section 4 presents the results while Section 5 concludes the work.

2 Transient dissipation

The transient energy \(E_{tr}\) is defined as the energy dissipated when the output capacitor is charged/discharged. In a CMOS buffer the charge at the output node \(Q_{out}\) is stored in both the output and the coupling capacitor and can be expressed as:

\[
Q_{out} (V_{in}, V_{out}) = (C_M + C_L) V_{out} - C_M V_{in}
\]

where \(V_{out}\) is the output voltage, \(V_{in}\) the input voltage, \(C_M\) is the IOCC, and \(C_L\) is the total capacitance at the buffer output.

The value of \(C_M\) when the input is in the static low state (defined as \(C_M^L\)) is computed considering the side-wall capacitance of both transistor drains, the gate to drain capacitance of the pMOS transistor in the linear region, and is given by:

\[
C_M^L = C_{ox} \left( \frac{W_{p_{eff}} L_{p_{eff}}}{2} + L_{AP_p} W_{p_{eff}} + L_{AP_n} W_{n_{eff}} \right)
\]

with \(L_{AP_p}\) and \(L_{AP_n}\) being the gate drain underdiffusion for the pMOS and nMOS transistors respectively, \(W_{n_{eff}}\) and \(W_{p_{eff}}\) the nMOS and pMOS effective channel width, \(L_{p_{eff}}\) the effective channel length of the pMOS transistor, and \(C_{ox}\) the gate oxide capacitance. For a static input high the capacitance \(C_M^H\) can be obtained similarly.

The energy dissipated at the nMOS transistor for a high to low output transition is given by \(E_{tr} = QV_{DD}/2\), where \(Q\) is the charge transferred from the output node to ground through the nMOS transistor, and \(V_{DD}\) is the voltage swing. Thus, transient energy is expressed as:

\[
E_{tr} = \left[ Q_{out} (V_{in} = 0, V_{out} = V_{DD}) - Q_{out} (V_{in} = V_{DD}, V_{out} = 0) \right] \frac{V_{DD}}{2}
\]

where \(Q_{out}(0, V_{DD})\) is the charge stored at the output node at the beginning of the transition and \(Q_{out}(V_{DD}, 0)\) is the charge stored at the output node when the transition is finished. Using eq.(1) we obtain:

\[
E_{tr} = \frac{1}{2} (C_L + C_M^L + C_M^H) V_{DD}^2
\]

The same expression is obtained when a low to high output transition is considered.

3 Short-circuit power model

We compute the short-circuit energy dissipated in a CMOS buffer (Fig. 1) for a low to high input transition (for a high to low input transition the model is equivalent). In our analysis we initially neglect the resistance of the interconnect line \(R\) and the input-output coupling capacitor \(C_M\), that will be included later.
The short-circuit component is obtained by combining the solution obtained for the limit of a fast input transition (or a large output capacitor) and the one obtained for the limit of a slow input transition (or a small output capacitor). Finally we include the effects of the input-output capacitor and the resistance of the interconnect.

3.1 Short-circuit model for heavily-loaded buffers

The dynamic behavior of the circuit in Fig. 1 is described by:

\[(C_L + C_M) \frac{dV_{\text{out}}}{dt} = I_p - I_n + C_M \frac{dV_{\text{in}}}{dt}\]  \hspace{1cm} (5)

where \(I_p\) and \(I_n\) are the pMOS and nMOS currents respectively.

Using a simple short-channel MOSFET model from [6], and neglecting channel-length modulation effects, the drain current expression is given by:

\[I_{DS} = \begin{cases} 
0 & (V_{GS} \leq V_{TH}) \\
(2 - \frac{V_{GS}}{V_{D}}) \frac{V_{DS}}{V_{D}} I'_{D0} & (V_{DS} < V'_{D0}) \\
I'_{D0} & (V_{DS} \geq V'_{D0})
\end{cases}\]  \hspace{1cm} (6)

with:

\[I'_{D0} = I_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^n\]  \hspace{1cm} (7)

The saturation voltage \(V'_{D0}\) is given by [6]:

\[V'_{D0} = V_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^m\]  \hspace{1cm} (8)

The parameter \(n\) is the velocity saturation index that takes a value between 2 (long-channel devices) and 1 (short-channel). Parameters \(I_{D0}\) and \(V_{D0}\) are the drain current and saturation voltage for \(V_{GS} = V_{DS} = V_{DD}\), while parameters \(n, m\) and \(V_{TH}\) are fitting parameters.

The input voltage is described with a linear ramp as:

\[V_{\text{in}}(t) = V_{TN} + \frac{t - t_n}{t_{sec}} V_{SC}\]  \hspace{1cm} (9)

Fig. 1. CMOS buffer model
where $V_{SC} = V_{DD} - V_{TN} - |V_{TP}|$, $t_{sc} = \frac{V_{SC}}{V_{DD}} t_{in}$ and $t_n = \frac{V_{TP}}{V_{DD}} t_{in}$. $V_{TN}$ and $V_{TP}$ are the nMOS and pMOS threshold voltages, and $t_{in}$ is the input rise time.

At the beginning of the transition, the nMOS transistor is saturated. For heavily-loaded buffers the short-circuit and overshooting effects can be neglected and eq.(5) can be simplified to:

$$C_S \frac{dV_{out}}{dt} = -I'_{DO_n}$$

with $I'_{DO_n}$ given by eq. (7) and $C_S = C_L + C^L_M$. We can solve eq. (10) with the initial condition $V_{out}(t_n) = V_{DD}$ obtaining:

$$V_{out}(t) = V_{DD} - V_D \left(1 - \frac{t}{t_n}\right)^{n+1}$$

where $t_n$ is the time at which the nMOS transistor starts to conduct and $V_D$ is given by:

$$V_D = \frac{I_{DO_n} t_{sc}}{C_S} \left(\frac{V_{SC}}{V_{DD} - V_{TN}}\right)^{n+1}$$

Eq. (11) is used in the linear expression of the pMOS current (see eq. (6)). For the pMOS transistor we express the drain, gate and saturation voltages as:

$$V_{DS} = V_D \left(1 - \frac{t}{t_n}\right)^{n+1}$$

$$V_{GS} = |V_{TP}| + V_{SC} \left(1 - \frac{t - t_n}{t_{sc}}\right)$$

$$V_{iD_p} = \frac{V_{DO_p} V_{SC}}{(V_{DD} - |V_{TP}|)} \left(1 - \frac{t - t_n}{t_{sc}}\right)$$

where a value of $m = 1$ is considered at the saturation voltage expression.

Since we are now considering heavily-loaded buffers, then the drain-source voltage at the pMOS transistor is small because the output capacitor is large. Thus, we take only a first order expression for this current in terms of $V_{DS}$ from eq.(6).

$$I_p = 2 \frac{V_D}{V_{DO_p}} I_{D_p} \left(\frac{t - t_n}{t_{sc}}\right)^{n+1} \left(1 - \frac{t - t_n}{t_{sc}}\right)^{n_p-1}$$

where $I_{D_p}$ has the form:

$$I_{D_p} = I_{DO_p} \left(\frac{V_{SC}}{V_{DD} - |V_{TP}|}\right)^{n_p-1}$$

The time at which the maximum current takes place for heavily loaded buffers $t_{C_L \rightarrow \infty}^{max}$ is obtained solving $\partial_t I_p = 0$.

$$t_{C_L \rightarrow \infty}^{max} = t_n + t_{sc} \frac{n_{n+1}}{n_n + n_p}$$

Then, the maximum current for heavily-loaded buffers is expressed as:

$$I_{C_L \rightarrow \infty}^{max} = 2 \frac{V_D}{V_{DO_p}} I_{D_p} \left(\frac{n_{n+1}}{n_n + n_p}\right)^{n+1} \left(\frac{n_p - 1}{n_n + n_p}\right)^{n_p-1}$$
3.2 Short-circuit model for unloaded buffers

When the output capacitance is small (i.e. when the short-circuit current has the greater impact [1]) the circuit behavior is close to the inverter DC operation since $I_p \approx I_n$. At the beginning of the transition, the pMOS transistor drives a current equal to the nMOS saturation current, while at the end of the transition the pMOS is saturated. In this particular case, the maximum current takes place when $I_{D0_n} = I_{D0_p}$. Given that $V_{GS} = V_{in}$ for the nMOS and $V_{GS} = V_{DD} - V_{in}$ for the pMOS, the time at which the short-circuit is maximum (defined as $I_{\text{max}}^{CL=0}$) can be obtained solving:

$$I_{D0_n} \left( \frac{V_{DD} - V_{TN}}{V_{DD} - V_{TN}} \right)^{n_n} = I_{D0_p} \left( \frac{V_{DD} - V_{DD} - |V_{TP}|}{V_{DD} - |V_{TP}|} \right)^{n_p}$$

(18)

where $I_{D0_n}$ and $I_{D0_p}$ are the parameters $I_{D0}$ of the nMOS and the pMOS transistor respectively. We obtained a good analytical approximation to the solution of eq. (18) as:

$$i_{CL=0}^{\text{max}} = t_n + t_{in} \left( 1 - \frac{V_{TN}}{V_{DD}} - \frac{|V_{TP}|}{V_{DD}} \right) \left( 1 + F_p \right)$$

(19)

where $F_p$ is given by:

$$F_p = \frac{I_{D0_n}}{I_{D0_p}} \left( \frac{V_{SC}}{V_{DD} - V_{TN}} \right)^{n_n} \left( \frac{V_{DD} - |V_{TP}|}{V_{SC}} \right)^{n_p}$$

(20)

Then, the maximum short-circuit current for unloaded buffers ($I_{\text{max}}^{CL=0}$) is easily evaluated from the saturation expression of the nMOS or the pMOS transistor (eq. (18)).

3.3 Combining both expressions

We construct an expression for the maximum short-circuit current $I_{\text{max}}$ that leads to $I_{CL=0}^{\text{max}}$ for large values of $C_L$, and to $I_{CL=0}^{\text{max}}$ for $C_L = 0$. $I_{CL=0}^{\text{sum}}$ is a function of the form $A/C_L^k$ and therefore divergent when $C_L = 0$. To prevent such a divergence we use the following expression for the maximum short-circuit current.

$$I_{\text{max}} = \frac{I_{CL=0}^{\text{sum}}}{I_{CL=0}^{\text{max}} I_{CL=0}^{\text{max}}} + I_{CL=0}^{\text{max}}$$

(21)

The short-circuit current is initiated when the nMOS transistor starts to conduct (at $t_n = t_{in} \frac{V_{TX}}{V_{DD}}$) and ceases when the pMOS is off (at time $t_p = t_{in} \left( 1 - \frac{|V_{TP}|}{V_{DD}} \right)$). The short-circuit charge transferred (SCCT) is obtained approximating the short-circuit current shape as a triangle with maximum value $I_{\text{max}}$ (triangle ABC in Fig. 2). Then, the area under the triangle will be proportional to the SCCT:

$$Q_{SC}^0 = K I_{\text{max}} \frac{V_{TX}}{V_{DD}} \frac{|V_{TP}|}{V_{DD}}$$

(22)
where $\kappa$ is a fitting parameter accounting for the deviation of the short-circuit current shape from an ideal triangle. The parameter $\kappa$ is close to 1 for deep-submicron technologies due the linear dependence between the saturation current and the gate voltage (that describes an ideal triangle). For long-channel devices $\kappa = 2/3$ since the dependence between the saturation current and the gate voltage is quadratic (the area of a “triangle” composed by two parabolas is $\frac{1}{4}I_{\text{max}}(t_p - t_n)$). Therefore, the parameter $\kappa$ takes a value between 2/3 (long-channel) and 1 (deep-submicron). The parameter $\kappa$ is taken equal for both falling and rising input transitions for simplicity. For the two technologies used in this work we obtained $\kappa = 0.69$ and $\kappa = 0.75$ for the 0.35$\mu$m and the 0.18$\mu$m technologies respectively.

3.4 Overshooting effects

![Diagram](image-url)

**Fig. 2.** Geometrical derivation of $Q_{sc}'$ from $Q_{sc}^0$ and $t_{ov}$

**Effects of overshooting on the short-circuit current** Fig. 2 is a plot of the current through the short-circuiting transistor during the transition. Two cases are considered, $C_M = 0$ and $C_M \neq 0$. Simulations in Fig. 2 showed that overshooting displaces the short-circuit current to the right, maintaining the current slopes invariant. This can be described analytically using a geometrical approach as shown in Fig 2. We approximate the short-circuit current curve displacement with straight lines of equal slopes, with the displacement due to overshooting, $Q_{sc}^0$ is the area of the triangle $ABC$ (SCCT when $C_M = 0$) and is known from eq. (22), while the reduced SCCT due to overshooting ($Q_{sc}'$) is the area under the triangle $DEC$. Using this geometrical analogy $Q_{sc}'$ is derived from $Q_{sc}^0$ using $t_{ov}$ as:

$$Q_{sc}' = Q_{sc}^0 \left(1 - \frac{t_{ov} - t_n}{t_{sc}}\right)^2$$  \hspace{1cm} (23)

If $t_{ov} > t_n + t_{sc}$, eq.(23) no longer holds and $Q_{sc}' = 0$.

**Overshoot time evaluation** The solution of eq.(5) to get $t_{ov}$ is a non-linear problem with no-closed form expression. Therefore, we relate this parameter to the time at
which the overshooting current is maximum (defined as $t_{ov}^{max}$), to get an analytical expression for $t_{ov}$.

Fig. 3 shows HSPICE simulations of $t_{ov}$ and $t_{ov}^{max}$ for a 0.18$\mu m$ and a 0.35$\mu m$ process technology. A minimum sized inverter driving another minimum sized inverter are considered with different input rise times.

$$t_{ov} - t_n = \frac{3(t_{ov}^{max} - t_n)}{2}$$  \hspace{1cm} (24)

The maximum overshoot time is obtained from eq.(5) neglecting the current through the pMOS. At $t = t_{ov}^{max}$ the nMOS transistor is saturated and the output voltage is maximum ($\frac{dV_{out}}{dt} = 0$). Then eq. (5) is reduced to:

$$C_M\frac{dV_{in}}{dt} - I_{D0n} = 0$$

Eq.(25) can be expressed as:

$$C_M\frac{V_{DD}}{t_{in}} - I_{D0n} \left( \frac{V_{DD}^{max} - V_{TN}}{V_{DD} - V_{TN}} \right)^n_n = 0$$ \hspace{1cm} (26)

Solution to eq. (26) leads to:

$$t_{ov}^{max} = \frac{t_{in}}{V_{DD}} \left( V_{TN} + (V_{DD} - V_{TN}) \left( \frac{C_MV_{DD}}{I_{D0n}t_{in}} \right)^{1/n} \right)$$ \hspace{1cm} (27)

Once $t_{ov}^{max}$ is obtained, $t_{ov}$ is given by eq.(24).

Finally, the energy associated to the short-circuit current for a rising input transition is computed as:

$$E_{sc} = Q'_{sc}V_{DD}$$ \hspace{1cm} (28)
3.5 Including the resistivity of the interconnect line

We include the effect of the interconnect line by using the effective capacitance concept [7]. To obtain an analytical expression of the effective capacitance to be used in the short-circuit model we analyze the RC network of Fig. 4. The node voltage $V_1$ and $V_2$ are related by a system of differential equations.

$$\frac{dV_1}{dt} = \frac{1}{C_2} \left( \frac{V_2 - V_1}{R} - I_{D0a} \right)$$

(29)

For the case $R=0$ we have that $V_1 = V_2$.

$$V_1 = V_{DD} - t \frac{I_{D0a}}{C_L}$$

(30)

where $C_L = C_1 + C_2$. From eq.(30) we define the effective capacitance as:

$$C_{eff} \equiv \frac{I_{D0a} t}{V_{DD} - V_1(t)}$$

(31)

Using this definition we have that $C_{eff} = C_L$ when $R = 0$. For a large value of the interconnect resistance, $V_1$ takes the form:

$$V_1 = V_{DD} - t \frac{I_{D0a}}{C_1} + \frac{t^2 I_{D0a}}{2RC_1^2} + o \left( \frac{1}{R^2} \right)$$

(32)

Using this solution, the effective capacitance for large values of the interconnect resistance is:

$$C_{eff} \simeq C_1 + \frac{t}{2R}$$

(33)

We build a single expression of the effective capacitance that leads to $C_L$ for $R = 0$ and to eq.(33) for large values of the interconnect resistance as:

$$C_{eff}(t) = C_1 + \frac{(C_L - C_1) t}{2R(C_L - C_1) + t}$$

(34)

where $C_1$ represents the output capacitance of the buffer (drain capacitance of both nMOS and pMOS transistors) and $C_L$ is the total capacitance at the output. Eq.(34) is time dependent and is evaluated inside the interval we want a description of the output voltage. In our model we need an effective capacitance for the time during which the short-circuit current takes place. Therefore we take the time at the middle of the input transition so that $C_{eff} = C_{eff} \left( \frac{t}{2} \right)$. This effective capacitance is included in the short-circuit energy model (eq. (28)).
Fig. 5. Energy dissipated vs. line resistance for a 0.18\( \mu \)m technology. Different values of ratio \( W_p/W_n \) are considered.

Fig. 6. Energy dissipated per cycle period vs. input time for different \( W_p/W_n \) ratios and interconnect resistances for a 0.35\( \mu \)m technology.

4 Results

In Fig. 5 we plot HSPICE simulations of the short-circuit energy dissipated per one cycle period vs. line resistance for a 0.18\( \mu \)m technology. The interconnect line is simulated with HSPICE using the \( \pi3 \) model. Three different inverters with different values of the \( W_p/W_n \) ratio are considered. Total capacitor at the output is fixed at \( C_L = 0.2pF \). As can be appreciated the short-circuit power increases when increasing the line resistance. This is because part of the output capacitance is shielded from the buffer, thus decreasing the gate delay and increasing the short-circuit dissipation [7]. We compare HSPICE simulations (dots) with model in [5] and the model proposed. The model in [5] underestimates energy while the model proposed presents an overall good accuracy.

In Fig. 6 we show HSPICE simulations of the energy dissipated per one cycle period vs. the input time for a 0.35\( \mu \)m technology. Two different inverters are considered with different values of the \( W_p/W_n \) ratio and two values of the interconnect resistance (\( R = 0 \) and \( R = 1k\Omega \)). The total capacitor at the buffer output is fixed at \( C_L = 1pF \).
As can be appreciated, the model developed describes correctly HSPICE simulations for all the conditions considered.

Finally, in Table 1 we show the total and the mean error of the short-circuit energy dissipated for the models in [3, 5] and the model proposed respect HSPICE simulations in Figs. 5 and 6. As can be appreciated, the model proposed represents an improvement in terms of accuracy.

<table>
<thead>
<tr>
<th>Error</th>
<th>Max.(0.18μm)</th>
<th>Mean(0.18μm)</th>
<th>Max.(0.35μm)</th>
<th>Mean(0.35μm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5%</td>
<td>3%</td>
<td>30%</td>
<td>8%</td>
</tr>
<tr>
<td>[3]</td>
<td>66%</td>
<td>57%</td>
<td>90%</td>
<td>70%</td>
</tr>
<tr>
<td>[5]</td>
<td>90%</td>
<td>29%</td>
<td>84%</td>
<td>47%</td>
</tr>
</tbody>
</table>

5 Conclusions

An accurate analytical model to calculate the power consumption of CMOS buffers driving RC interconnect lines has been developed. The model is compared to HSPICE simulations (level 50) and other previously published models for a 0.18μm and a 0.35μm process technology reporting a high accuracy. It is able to describe the energy dependence of the buffer with input slew time, asymmetry of the buffer, output capacitance, input-output coupling capacitor and the resistance of the interconnect line. Non-linear problems are solved using simple formulas and avoiding time-consuming numerical procedures. The model represents an improvement respect previously published works providing a high accuracy respect HSPICE simulations.

ACKNOWLEDGMENT

This work has been supported by the Spanish government under the project CICYT-TIC98-0284.

References

Power-delay Modeling of Dynamic CMOS Gates for Circuit Optimization

J.L. Roselló and Jaume Segura
Physics Dept., Balearic Islands University, Palma de Mallorca, Spain

Abstract-We present an accurate analytical expression to compute power and delay of domino CMOS circuits from a detailed description of internal capacitor switching and discharging currents. The expression obtained accounts for the main effects in complex sub-micron gates like velocity saturation effects, body effect, device sizes and coupling capacitors. The energy-delay product is also evaluated and analyzed. Results are compared to HSPICE simulations (level 50) for a 0.18μm CMOS technology.

I. INTRODUCTION

Power dissipation has become an important concern in circuit design during the last decade due to heating problems in high-density/high-performance circuits and to the power saving required for portable applications [1]. Domino CMOS is widely used in digital VLSI circuits with the advantage of small area and delay when compared to complementary static logic [2]. The pMOS transistor network is removed and replaced by a single pMOS charging transistor to reduce layout area and interconnect capacitance, thus increasing further circuit speed and reducing power dissipation. If the timing at the inputs is such that they are not asserted until after the precharge clock has been asserted, then there is no short-circuit current during the output transition reducing power dissipation and increasing throughout. Power consumption is also decreased due to glitch-free operation. However, the large clock loads and signal transition activities due to the precharging result in an excessive power dissipation [3]. Since power consumption has become one of the biggest challenges in high-performance VLSI design [4], reducing power dissipation in domino CMOS ICs is critical for these applications. In this work we present analytical models to accurately compute power and delay of domino CMOS circuits from process parameters. These models can be used for fast timing and power estimation and circuit optimization. This paper is organized as follows. Section II presents an accurate and simple model of power dissipation in domino CMOS circuits. Section III develops the delay model, while Section IV presents a practical example of a 2NAND energy-delay optimization. The conclusions are in Section V.

II. ENERGY EVALUATION

Power consumption in domino circuits is mainly due to the charge/discharge of internal capacitors. This dynamic power is dissipated at the clock distribution network, the output buffer, the output load and the internal capacitors. The energy dissipated by domino CMOS circuits in one precharge/evaluation period can be expressed in a general form as $E = QV_{DD}$, where $Q$ is the charge provided by the supply voltage during the cycle, and $V_{DD}$ is the voltage difference through which the charge flows. The power consumption in a domino CMOS gate is related to this energy dissipated as $P = a f E$, where $f$ is the frequency of the clock, and $a$ is the switching activity factor. In this work we analyze the power dissipation of the NAND gate of Fig. 1 when the output is charged and discharged. This analysis can be extended to gates with more than two logic transistors in the discharging path. The total energy dissipated by the NAND gate is:

$$ E_{tot} = E_{clk} + E_{load} + E_{int} + E_{buffer} \quad (1) $$

where $E_{clk}$ is the energy dissipated due to the clock, $E_{load}$ is due to the capacitor at the logic output, $E_{int}$ to internal switching capacitors, and $E_{buffer}$ to the output buffer.
A. Energy dissipated by the clock

This energy component is obtained computing the charge transferred per clock cycle by the clocking buffer. When the clock node is high the charge in this node is:

\[ q_{\text{clk}}^H = (C_{\text{dk}} + C_{\text{gndT1}} + C_{\text{gndT2}} + C_{\text{gndT2}}^H) VDD \]  

where \( VDD \) is the supply voltage, \( C_{\text{dk}} \) is the output load of the clock buffer, \( C_{\text{gndT1}}^H \) and \( C_{\text{gndT2}}^H \) are the gate-to-bulk and gate-to-source capacitances of the nMOS transistor driven by the buffer (T1 in Fig.1) when the clock is high, and \( C_{\text{gndT1}} \) and \( C_{\text{gndT2}} \) are the gate-to-drain coupling capacitances of the nMOS and the pMOS transistors driven by the buffer evaluated when the clock is high. When the clock is low, the charge stored at the clock node is:

\[ q_{\text{clk}}^L = -(C_{\text{gndPd}}^L + C_{\text{gndPd}} + C_{\text{gndPd}}^H) VDD \]  

where \( C_{\text{gndPd}}^L, C_{\text{gndPd}} \) and \( C_{\text{gndPd}}^H \) are the gate-to-source, gate-to-bulk and gate-to-drain parasitic capacitance values of transistor Tpd when the clock is low. The gate capacitance of transistors T1 and Tpd is \( C_{\text{gT1}} = C_{\text{gT1}}^H + C_{\text{gT1}}^L + C_{\text{gT1}} \) and \( C_{\text{gTpd}} = C_{\text{gTpd}}^L + C_{\text{gTpd}} + C_{\text{gTpd}}^H \) respectively. These two components are dependent on transistor size as:

\[ C_{\text{gTpd}} = C_{\text{gox}} W_{\text{doff}} f_L + 2LAP_e \]  
\[ C_{\text{gT1}} = C_{\text{gox}} W_{\text{noff}} f_L + 2LAP_e \]

where \( C_{\text{gox}} \) is the gate oxide capacitance, \( W_{\text{doff}} f_L \) and \( W_{\text{noff}} f_L \) are the effective pMOS and nMOS channel widths, \( LAP_e \) are the process bias on the channel length of nMOS and pMOS, while \( LAP_e \) and \( L_{\text{noff}} \) are the effective pMOS and nMOS channel length respectively. Capacitance \( C_{\text{gTpd}}^H \) and eq.(2) is due to the side-wall capacitance between the gate and the drain of transistor Tpd, and is:

\[ C_{\text{gTpd}}^H = C_{\text{gox}} W_{\text{doff}} f_L \]

From eqs. (2) and (3) we evaluate the charge provided by the supply node \( (q_{\text{clk}}^H - q_{\text{clk}}^L) \) and then the energy dissipated per cycle:

\[ E_{\text{clk}} = (C_{\text{dk}} + C_{\text{gT1}} + C_{\text{gTpd}} + C_{\text{gTpd}}^H) VDD \]  

B. Power dissipated at the logic output

The power dissipated at the logic output is computed similarly. We assume that both gate inputs are at the high level when the clock pulse arrives (the worst-case in power dissipation). Under these conditions the charge stored at the output node of the domino gate is:

\[ q_{\text{load}}^H = -(C_{\text{gT1}}^H + C_{\text{gT2}} + C_{\text{gT2}}^H) VDD \]  
\[ q_{\text{load}}^L = (C_{\text{gpd}} + C_{\text{gT1}} + C_{\text{gT2}} + C_{\text{gT2}}^L) VDD \]  

Therefore, the energy dissipated due to the charge/discharge of the gate output is:

\[ E_{\text{load}} = (C_{\text{gpd}} + C_{\text{gT1}} + C_{\text{gT2}} + C_{\text{gT2}}^L) VDD \]

\[ + C_{\text{gT1}} + C_{\text{gT2}} + C_{\text{gT2}}^L + C_{\text{gT2}}^H + C_{\text{gT2}}^L) VDD \]

where \( C_{\text{gpd}} \) and \( C_{\text{gT2}} \) are the drain capacitances of transistors Tpd and T3, \( C_{\text{gT1}} \) and \( C_{\text{gT2}} \) are the gate capacitances of transistors T1n and T2p respectively. Finally, \( C_{\text{gT1}}^H, C_{\text{gT2}}^H, C_{\text{gT2}}^L, C_{\text{gT2}}^H, C_{\text{gT2}}^L \) and \( C_{\text{gT2}}^L \) are the gate-to-drain capacitances of transistor T3, T1, T2p and T2n respectively when their gate voltage are at the state defined by their superscripts. These capacitances have a linear dependence on the channel width of each transistor and can be easily evaluated.

C. Power dissipated by internal capacitors

If before precharge both inputs are high, then the internal nodes are grounded and the transistors of the chain are in the linear region. During precharge these nodes are charged until the upper transistor (T3 in Fig.1) is off. The final value of the voltage at internal nodes is \( V_{\text{int1}} = V_{\text{int2}} = VDD - V_{\text{th3}} \) where \( V_{\text{th3}} \) is the threshold voltage of transistor T3. As in the previous steps we evaluate the charge provided by the supply voltage to compute the energy. The charge stored at the internal capacitors when all inputs are high and the internal nodes are low is:

\[ q_{\text{int}}^L = -(C_{\text{gT2}} + C_{\text{gT2}} + C_{\text{gT2}} + C_{\text{gT2}}) VDD \]

where \( C_{\text{gT1}}^H, C_{\text{gT1}}^H, C_{\text{gT1}}^H, C_{\text{gT1}}^H, C_{\text{gT1}}^H \) and \( C_{\text{gT1}}^H \) are the gate-to-source and gate-to-drain capacitances of the i-th transistor evaluated when the input is high and their drain and source are low (that is, when they are in the linear region). The value of these capacitances can be expressed as:

\[ C_{\text{gT1}}^H = C_{\text{gox}} W_{\text{doff}} f_L + 2LAP_e \]

where \( W_{\text{doff}} f_L \) and \( LAP_e \) are the effective channel width and length of the i-th transistor.

The charge stored at the internal nodes when they are charged through the pull-up pMOS transistor is:
\[ q_{\text{int}}^H = -(C_{pT3} + C_{pT2}) V_{DD} + (C_1 + C_2 + C_{pT2} + C_{\text{int}}^H) V_{\text{int}} \]  

where \( C_1 \) is the drain capacitance of transistors \( T1 \) and \( T2 \), and \( C_2 \) is the drain capacitance of transistors \( T2 \) and \( T3 \). The side-wall capacitances of transistors \( T1 \) and \( T3 \) are neglected in eq. (11).

Based on eqs. (9) and (11), the energy associated to the charge/discharge of this internal node capacitances is:

\[ E_{\text{int}} = (C_{pT3} + C_{pT1}) V_{DD} + (C_1 + C_2 + C_{pT2} + C_{\text{int}}^H) V_{\text{int}} V_{DD} \]  

\[ (12) \]

**D. Power dissipated by the buffer output**

Neglecting short-circuit currents, the charge at the buffer output when the output is high is:

\[ q_{\text{BUF}}^H = (C_{\text{out}} + C_{\text{gph}} + C_{\text{gph}} + C_{p\text{gph}}^L + C_{p\text{gph}}^H) V_{DD} \]  

\[ (13) \]

where \( C_{\text{out}} \) is the capacitance due to the gates driven by the buffer, \( C_{\text{gph}} \) and \( C_{p\text{gph}} \) are the drain capacitances of the pMOS and the nMOS transistor of the buffer, and \( C_{p\text{gph}}^L \) and \( C_{p\text{gph}}^H \) are the gate-to-drain coupling capacitances of the pMOS and the nMOS transistors when the input voltage of the buffer is low. Similarly, the charge at the buffer output when it is low is:

\[ q_{\text{BUF}}^L = -(C_{\text{gph}}^L + C_{p\text{gph}}^H) V_{DD} \]  

\[ (14) \]

From eqs. (13) and (14) the energy dissipated by the buffer is:

\[ E_{\text{BUF}} = (C_{\text{out}} + C_{\text{gph}} + C_{p\text{gph}} + C_{p\text{gph}}^L + C_{p\text{gph}}^H) V_{DD} \]  

\[ (15) \]

**III. DELAY MODEL**

Gate level energy reduction is always considered under timing constraints. There are several ways that model delay in CMOS buffers [5, 6]. The analysis of multiple input gates is more complex due to the body effect at intermediate devices, and to charge distribution through the internal gates [5]. In this section we present a simple analytical model for delay evaluation in domino CMOS gates. We reduce the nMOS chain to an equivalent transistor. This device is then included in a simple model to compute the gate delay \( t_d \) and its output transition time \( t_f \). Finally, this simple delay model is applied to the CMOS output buffer to compute the whole delay.

\[ I_D = \begin{cases} 
0 & (V_{GS} \leq V_{TH}) \\
(2 - \frac{V_{TH}}{V_{DD}}) \frac{V_{DD}}{V_{GS}} I_D & (V_{DS} < V_{DD}) \\
I_D & (V_{DS} \geq V_{DD})
\end{cases} \]

\[ (16) \]

where \( V_{GS} \), \( V_{DD} \), \( V_{DS} \), and \( V_{TH} \) are the gate, supply, saturation and threshold voltage respectively. The threshold voltage \( V_{TH} \) is expressed as:

\[ V_{TH} = V_{TH0} + \gamma V_{ab} \]

\[ (18) \]

where \( V_{TH0} \) is the threshold voltage with no substrate or drain bias, \( \gamma \) is the linearized body effect coefficient, and \( V_{ab} \) is the source-to-bulk voltage. \( I_D \) is the drain current at \( V_{GS} = V_{DD} = V_{DD} \). The parameter \( \alpha \) is the velocity saturation index that ranges between 2 (long-channel devices) and 1 (short-channel). Parameters \( I_D \), \( V_{DD} \), \( V_{TH0} \) and \( \alpha \) are extracted from the I-V data [5].

1. **Two stacked transistors**

Fig. 2 shows two series-connected nMOS transistors. We define the parameter \( I_D^{(2)} \) as the charge that flows through the two transistors when \( V_1 = V_2 = V_{DD} \). In this state, transistor \( T2 \) is saturated while \( T1 \) is in the linear region. To compute \( I_D^{(2)} \) we assume that the internal node voltage \( V' \) is small compared to the supply voltage \( V_{DD} \) and equate the drain current of both transistors. Using \( V_2 = V_{DD} \), a first order Taylor expansion of \( T2 \) drain current is:

\[ I_{DS2} \approx I_{D02} \left( 1 - \frac{\alpha_2 V' (1 + \gamma_2)}{V_{DD} - V_{TH0}} \right) \]

\[ (19) \]

Since transistor \( T1 \) is in the linear region, its drain current when \( V_1 = V_{DD} \) is:

\[ 496 \]
\[ I_{DS_1} = I_{D0_1} \left( 2 - \frac{V'}{V_{DS_1}} \right) \frac{V'}{V_{DS_1}} \]  
(20)

Equating eqs. (20) and (19), and solving for \( V' \):

\[ V' = V_0 \left( 1 - \sqrt{1 - \frac{V_{T0_1}^2}{V_{DS_1}^2} \frac{I_{DS_1}}{I_{D0_1}} \frac{1}{V_0}} \right) \]  
(21)

where:

\[ V_0 = V_{T0_1} + \frac{\alpha_2 (1 + \gamma_2) V_{T0_1}^2}{2 (V_{DD} - V_{T0_1}) I_{D0_1}} \]  
(22)

\( I_{D0_1} \) is computed substituting eq.(21) in eq.(19).

We observed through SPICE simulations that when varying the gate voltage at transistor \( T_2 \), the threshold and saturation voltage and also the velocity saturation coefficient of the chain were equal to the values of these coefficients for transistor \( T_2 \).

Therefore, the drain saturation current equation of the two series connected transistors when varying \( V_2 \) can be expressed as:

\[ I_D = I_{D0_1} \left( \frac{V_2 - V_{T0_1}}{V_{DD} - V_{T0_1}} \right) \alpha_2 \]  
(23)

2. Three stacked \( n \)MOS transistors

In a three stacked transistors, the pair of transistors at the top of the chain (say \( T_2 \) and \( T_3 \)) are collapsed to a single transistor \( T_2' \) with parameters \( I_{D0_{T2'}} = I_{D0_2} \), \( V_{T0_{T2'}} = V_{T0_2} \), \( \alpha_{T2'} = \alpha_3 \) and \( \gamma_{T2'} = \gamma_3 \). Then, transistor \( T_2' \) and \( T_1 \) (\( T_1 \) being the transistor at the bottom of the chain) are collapsed to an equivalent transistor \( T_1' \). The drain current when \( V_1 = V_2 = V_3 = V_{DD} \) is computed as \( I_{D0_{T1'}} = I_{D0_1} I_{D0_2} \). We define this current as \( I_{D0_{T1'}} \), \( I_{D0_{T2}} \).

3. \( n \)-stacked \( n \)MOS transistor

This scheme can be applied to a series of \( n \)-stacked MOSFETs. The drain current when varying the gate voltage of the \( n \)-th transistor in the chain (the upper transistor) is:

\[ I_D = I_{D0_{Tn}} \left( \frac{V_n - V_{T0_n}}{V_{DD} - V_{T0_n}} \right) \alpha_n \]  
(24)

where \( I_{D0_{Tn}} \) is the drain current of the chain when \( V_1 = V_{DD} \) \( \forall n \in (1, 2, ..., n) \). Fig. 3 compares eq. (24) to HSPICE simulations for 4 series-connected \( n \)MOS transistors in a 0.35\( \mu \)m technology. Each input \( V_i \) is increased from 0\( V \) to \( V_{DD} \) with an input rise time of 300\( \mu \)s. The other three gates are fixed to 0\( V \). The current is fitted correctly when varying the top transistor \( T_4 \) even for this dynamic simulations. Current values when varying the other three inputs provides a more complex behavior due to coupling capacitor and saturation effects.

B. Delay model

The pull-down network is simplified to a single transistor following the process described previously. The parameter \( I_{D0_{T1}} \) represents the maximum current that can be driven by the chain of transistors. We compute the gate delay when the upper \( n \)MOS transistor switches. The input voltage is:

\[ V_n(t) = V_{T0_1} + (V_{DD} - V_{T0_1}) \frac{t - t_n}{t_{in} - t_n} \]  
(25)

where \( t_n \) is the time when the \( n \)MOS chain starts to conduct (\( t_n = V_{T0_1} I_{in}/(V_{DD}) \) and \( t_{in} \) is the input rise time. At the beginning of the transition the \( n \)MOS chain is off and \( V_{out} = V_{DD} \). At \( t = t_n \), the \( n \)MOS chain starts to conduct. An analytical expression for the output voltage is obtained solving:

\[ C_L \frac{dV_{out}}{dt} = -I_D \]  
(26)

where \( C_L \) is the total output capacitance of the gate, and \( I_D \) is the current of the chain. An analytical solution to eq. (26) from eqs. (24) and (25) is:

\[ V_{out} = V_{DD} - I_{D0_{T1}} \left( \frac{t - t_n}{C_L (t_{in} - t_n) + \alpha_{n+1} t_{in} - t_n} \right) \alpha_n \]  
(27)

Eq. (27) is used to obtain the delay from the input at 0.5\( V_{DD} \) to the output at 0.5\( V_{DD} \) (\( t_{di} \)).

\[ t_{di} = t_n + \left[ \frac{Q_f (\alpha_{n+1})}{I_{D0_{T1}}} \right] t_{in} \left( t_{in} - t_n \right) \frac{\alpha_n}{2} \]  
(28)
where \( Q_L = C_L V_{DD}/2 \) is the charge transferred by the nMOS transistors when the output reaches \( V_{DD}/2 \). Eq. (28) is valid when \( t_{d1} < t_{m}/2 \). If \( V_f \) is defined as the value of \( t_{m} \) at time \( t = t_{m} \), then eq. (28) is valid in the interval \( V_f < V_{DD}/2 \).

Otherwise, if the input node reaches the supply voltage before the output is at \( V_{DD}/2 \) then eq. (28) is not valid. The solution of eq. (28) when \( V_a = V_{DD} \) is straightforward and in this case \( t_{d1} \) is given by:

\[
t_{d1} = \frac{t_{m}}{2} + \frac{C_L (V_f - V_{DD}/2)}{I_{DD}} \tag{29}
\]

Eq. (29) is valid when \( V_f > V_{DD}/2 \).

The slope of the output voltage at \( V_{DD}/2 \) can also be obtained as:

\[
\frac{dV_{out}}{dt} |_{V_{DD}/2} =
\begin{cases} 
\frac{t_{m} - t_{f}}{t_{m}} \left( \frac{Q_f(t_{m} - t_{f})}{t_{m} - t_{f}} \right)^{\frac{1}{2}} & (V_f < V_{DD}/2) \\
\frac{t_{m} - t_{f}}{t_{m}} & (V_f > V_{DD}/2)
\end{cases}
\tag{30}
\]

In calculating the output fall time of the dynamic gate (\( t_f \)), the output waveform slope is approximated by 70% of its derivative at the half-\( V_{DD} \) point.

The delay of the output buffer in Fig. 1 is obtained similarly using the alpha-power law parameters of the pMOS transistor of the buffer obtaining:

\[
t_{d2} = t_p + \left[ \frac{Q_f}{I_{DD}} \left( \alpha + 1 \right) \right] \left( t_f - t_p \right) \frac{t_{m}}{2} \tag{31}
\]

where \( t_p = t_f |V_T|/V_{DD} \) is the time when the pMOS device starts to conduct, \( \alpha \) and \( I_{DD} \) are the velocity saturation index and the parameter \( I_{DD} \) of pMOS. Eq. (31) is valid when \( V_c > V_{DD}/2 \), where \( V_c \) is the voltage at the buffer output when \( t = t_f \).

If \( V_c < V_{DD}/2 \) then:

\[
t_{d2} = \frac{t_f}{2} + \frac{C_{out} (V_{DD} - V_c)}{I_{DD}} \tag{32}
\]

The input fall time to the CMOS buffer is given by the value of \( t_f \) provided by the nMOS chain.

The total circuit current contributions are neglected in this analysis while overshoot effects are considered. We include overshoot by computing the charge that must be transferred through the pMOS transistor to reach \( V_{DD}/2 \) at the output of the buffer. This charge is \( Q_f = \sum_{i=1}^{n} V_a (V_{DD}/2) - \sum_{i=1}^{n} V_{out}(t) \) where \( V_a \) is the output voltage of the buffer and the charge \( \sum_{i=1}^{n} V_{out}(t) \) can be approximated by \( \sum_{i=1}^{n} V_{out}(t) \) (see eqs. (13) and (14)).

Thus, the total delay is computed as the sum of the two delays \( t_d = t_{d1} + t_{d2} \).

IV. ENERGY-DELAY EVALUATION

The energy-delay relationship of each gate in an IC is controlled through design by device sizing. In the device description considered, the geometry dependence comes through parameters \( I_{DD_w} \) and \( I_{DD_b} \) (that are the parameter \( I_{DD} \) for transistors \( T_i \) and \( T_{p} \) respectively). Since \( I_{DD} \) is linearly dependent on the channel width, we use:

\[
I_{DD} = J_{DD} W + \delta I_{DD}
\tag{33}
\]

where \( J_{DD} \) and \( \delta I_{DD} \) are technology-dependent parameters. Applying eq. (33) to the energy and delay expressions derived previously we compute the energy-delay product (EDP).

The channel width of the series connected MOSFETs (we assume that all transistors in the nMOS chain are equally sized) and the nMOS transistor of the buffer \( (W_{a} \) and \( W_{b} \) can be optimized to obtain a minimum EDP. The pMOS channel width of the buffer is fixed to \( W_{p} = 2W_{a} \).

In the simulations performed we compute the energy dissipated when the output is charged in the precharge phase and discharged in the evaluation phase when the gate voltage of any of the transistors of the chain \( V_i \) changes. The delay is computed as the time from \( V_{DD}/2 \) at the rising gate voltage \( V_i \) of transistor \( T_i \) to \( V_{DD}/2 \) at the output of the CMOS buffer. Fig. 4 is a plot of EDP as a function of parameters \( W_{a} \) and \( W_{b} \) for a 0.13μm CMOS technology. A minimum is obtained for \( W_{a} = 6\mu m \) and \( W_{b} = 3.5\mu m \).

Fig. 5 compares model predictions to HSPICE simulations for EDP when varying \( W_{a} \) for a fixed value of the buffer nMOS width \( W_{b} = 3.5\mu m \). HSPICE simulations are reproduced by the model that predicts the position of the minimum.

Fig. 6 compares the model to HSPICE when varying \( W_{b} \) with \( W_{a} \) evaluated at the minimum EDP obtained previously \( (W_{a} = 6\mu m) \). A minimum at \( W_{b} = 3.5\mu m \) is obtained in accordance to HSPICE.
Finally, Fig. 7 plots EDP vs. supply voltage when $W_{na} = 3.5 \mu m$ and $W_{ni} = 6 \mu m$. As can be appreciated, a minimum EDP is obtained at 0.9V as predicted by HSPICE simulations.

V. CONCLUSIONS

We presented a simple and accurate analytical expression to compute power and delay of domino CMOS circuits from a detailed description of internal capacitor switching and discharging currents. The expression obtained account for the main effects present in complex sub-micron gates like velocity saturation effects, body effect, device sizes and coupling capacitors. The energy-delay product is also evaluated and analyzed showing a good agreement with HSPICE simulations for a 0.18um technology. Although second order effects like short-circuit contribution are not included, the results obtained suggest that their contribution is not significant when computing the minimum EDP. Additional work is required to model EDP when any of the gate inputs makes a transition.

ACKNOWLEDGMENT

This work has been supported by the Spanish Government under the project CICYT-TIC98-0294.

REFERENCES


Modeling the input-output coupling capacitor effects on the CMOS buffer power consumption

J.L. Rossello and Jaume Segura
Physics Dept., Balearic Islands University, Cra. Valsdemossa km.7.5
07071 Palma de Mallorca, Spain
email: {j.rossello,jaume.segura}@uib.es

Abstract- A detailed description of coupling capacitor effects on the short-circuit and transient power is presented. It is shown that the input-output coupling capacitor decreases the contribution of the short-circuit charge transferred (SCCT), while increases the total capacitor at the buffer output thus increasing the total power. The model is developed for sub-micron CMOS buffers and applied to a recently published short-circuit power model resulting in a considerable improvement in terms of accuracy. Results are compared to HSPICE simulations (level 50) for a 0.35µm technology.

I. INTRODUCTION

A high percentage of the power dissipated by VLSI integrated circuits is due to the clock distribution network, I/O drivers and busses which are all based on inverters. The analysis of CMOS inverters is also the most important step when a detailed description of more complex gates is required. Therefore, the analytical description of the power dissipated in a CMOS inverter is of increasing importance.

Power dissipation in CMOS circuits has a dynamic and a static component. The dynamic dissipation is due to the charge/discharge of gate output load (defined as transient energy), and to the short-circuit current due to the supply-ground conducting path created during the transition [1].

Several works have focussed on modeling the short-circuit power consumption. Veendrick [1] obtained an expression for unloaded buffers. Sakurai et al. [2] derived a model for long and short channel devices using their alpha-power law MOSFET model for unloaded buffers. Turgis et al. [3] derived an expression for the short-circuit and over-shooting dissipation and introduced the equivalent capacitance concept allowing a direct and frequency-independent comparison of the different power components. More recently, Nose et al. [4] derived a simple model for the short-circuit power which takes short-channel effects into consideration concluding that the short-circuit power to total power ratio $P_{sc}/P_{total}$ will not change with scaling if the ratio $V_{TH}/V_{DD}$ is kept constant. This model does not take into account coupling capacitor effects, of increasing importance in current submicron technologies.

As the transistor feature size is decreased, the gate oxide thickness is also decreased further increasing coupling capacitance effects. Therefore this capacitor cannot be neglected when considering sub-micron technologies. The input-output coupling capacitor reduces the short-circuit current shape and therefore the total short-circuit charge transferred. This capacitor also impacts on the transient dissipation further increasing the total power.

In this work we propose a simple model to accurately compute the coupling capacitor contribution to the power dissipated in a CMOS buffer. A closed form expression for the transient energy is developed and the impact of overshooting on the short-circuit current shape is analyzed. This analysis is applied to the short-circuit model developed by K. Nose and T. Sakurai in [4]. Model results are compared to HSPICE simulations (level 50) for a 0.35µm technology showing a significant improvement in terms of accuracy. This paper is organized as follows: in section II we obtain an expression for the transient energy. Section III analyzes the impact of the coupling capacitor on the short-circuit energy component. Section IV presents the results while Section V concludes the work.

II. TRANSIENT DISSIPATION

The transient energy ($E_{tr}$) is the energy dissipated when charging/discharging the output capacitor. In this section we analyze the power dissipated when discharging the output capacitor since the charging
case is equivalent. The output node charge is stored in both the input-output coupling capacitor \( C_M \) and the load capacitor \( C_L \) (see Fig. 1). The energy of a discharging capacitor is \( E = QV/2 \), where \( V \) is the voltage swing and \( Q \) the charge initially stored in the capacitor. The voltage swing at the output node of the buffer is equal to the supply voltage \( V_{DD} \) while the charge transferred can be derived using charge conservation by computing the charge at the beginning of the transition minus the charge at its end.

The value of the coupling capacitor \( C_M \) in the static input low state (defined as \( C_{L}^{I} \)) considering the side-wall capacitance of both transistor drains, and the gate to drain capacitance of the pMOS transistor in the linear region is given by:

\[
C_M^I = C_{ox} \left( \frac{W_{p_{eff}}}{L_{p_{eff}}} + L_{AP} W_{p_{eff}} + L_{AP} W_{n_{eff}} \right)
\]

with \( L_{AP} \) and \( L_{AP} \) being the gate drain underdiffusion for the pMOS and nMOS transistors respectively, \( W_{n_{eff}} \) and \( W_{p_{eff}} \) are the nMOS and pMOS effective channel width while \( L_{p_{eff}} \) is the effective channel length of the pMOS transistor. For a static input high the capacitance \( C_M^H \) can be obtained similarly.

In a CMOS buffer the charge at the output node is stored in both the output and the coupling capacitor. This charge (defined as \( Q_{out} \)) can be expressed as:

\[
Q_{out} = (C_M + C_L) V_{out} - C_M V_{in}
\]

where \( V_{out} \) is the output voltage and \( V_{in} \) the input voltage of the buffer.

The energy dissipated when discharging the charge stored at the buffer output can be expressed as:

\[
E_{tr} = \frac{1}{2} \left( Q_{out} (0) - Q_{out} (\infty) \right) \frac{V_{DD}^2}{2}
\]

where \( Q_{out} (0) \) is the charge at the output node at the beginning of the transition and \( Q_{out} (\infty) \) is the charge stored at the output node when the output transition is finished.

Using eq. (2) we obtain an expression for \( Q_{out} (0) \) and \( Q_{out} (\infty) \) as:

\[
Q_{out} (0) = (C_L + C_M^I) V_{DD}
Q_{out} (\infty) = -C_M^H V_{DD}
\]

From eqs. (3) and (4) the transient energy for a high to low output transition is:

\[
E_{tr} = \frac{1}{2} (C_L + C_M^H + C_M^I) V_{DD}^2
\]

For a low to high output transition the analysis is equivalent and the same expression is obtained.

### III. Coupling capacitor effects on short-circuit power

The total power dissipated in one cycle period is given by:

\[
E_{tot} = (C_L + C_M^H + C_M^I) V_{DD}^2 + (Q_{sc}^f + Q_{sc}^l) V_{DD}
\]

where \( Q_{sc}^f \) and \( Q_{sc}^l \) are the short-circuit charge transference for a rising and a falling input transition respectively. In this section we analyze the influence of coupling capacitor \( C_M \) on the short-circuit charge transference (SCCT).

The dynamic behavior of the circuit in Fig. 1 is described by:

\[
(C_L + C_M) \frac{dV_{out}}{dt} = I_p - I_n + C_M \frac{dV_{in}}{dt}
\]

where \( I_p \) and \( I_n \) are the pMOS and nMOS current respectively. Eq. (6) is non-linear and non-closed form expression can be obtained. The effect of \( C_M \) is often neglected [4] in order to obtain a simpler expression for the output voltage and the short-circuit current shape. The cases of rising and falling input transitions are equivalent and only the rising input transition is analyzed.

#### A. MOSFET model

A simple short-channel MOSFET model is developed in [5]. Neglecting channel-length modulation effects, this model is expressed as:

\[
I_{DS} = \begin{cases} 0 & (V_{GS} \leq V_{TH}) \\ \left( 2 - \frac{V_{DS}}{V_{TH}} \right) \frac{V_{TH}}{V_{DS}} I_{D0} & (V_{DS} < V_{D0}) \\ I_{D0} & (V_{DS} \geq V_{D0}) \end{cases}
\]

with

\[
I_{D0} = I_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^n
\]
The saturation voltage $V'_{D0}$ is given by:

$$V'_{D0} = V_{D0} \left( -\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^n$$

(9)

where parameter $n$ is the velocity saturation index that takes a value between 2 (long-channel devices) and 1 (short-channel) and the two parameters $I_{D0}$ and $V_{D0}$ are the drain current and saturation voltage for $V_{GS} = V_{DS} = V_{DD}$. Parameters $n$, $m$ and $V_{TH}$ are fitting parameters.

B. Overshooting effects

Fig. 2 presents HSPICE simulations of the current through the pMOS transistor during a rising input transition for various values of the coupling capacitance. This current (labeled as $I_p$ in Fig. 1) has two components clearly distinguished by the sign of the current. At the beginning of the transition, a negative pMOS current is present due to a partial discharge of the output capacitor from the output node to the supply rail. This effect appears when the input-output capacitor drives the output voltage beyond the supply value ($V_{DD}$) at the beginning of the transition and is known as overshooting. The time during which the output voltage is beyond the supply value is called the overshoot time, $t_{ov}$. When the nMOS device starts to conduct (that is, when the input voltage is equal to the nMOS threshold voltage $V_{in} = V_{TN}$), it pulls the output voltage down. Once the output voltage goes below $V_{DD}$, the pMOS current is positive corresponding to the short-circuit component due to the simultaneous conduction of both devices.

Therefore, the overshoot and short-circuit current components are related and their relative contribution is determined by the input transition time and the coupling capacitance. If the input voltage is below $V_{DD} + V_{TP}$ (the input voltage value at which the pMOS turns off; $V_{TP}$ is the pMOS threshold voltage) at $t = t_{ov}$, then there will be a short-circuit current period from this time until the pMOS is turned off. Otherwise the output voltage is still beyond $V_{DD}$ when the input reaches $V_{DD} + V_{TP}$ and the short-circuit current is not present.

Fig. 2 shows that overshooting displaces the short-circuit current to the right, maintaining the current slopes invariant. As can be appreciated, the short-circuit charge transferred (positive current) is further reduced while the overshooting charge transferred (negative current) increases. Therefore, overshoot effects cannot be neglected when computing the short-circuit component.

In this work, overshooting effects on the short-circuit current shape are included through the time at which overshooting current vanishes $t_{ov}$. An empirical relationship between the short-circuit charge transferred ($Q_{sc}^c$) and the value of this charge when the coupling capacitor is neglected ($Q_{sc}^0 = Q_{sc}(C_M = 0)$) is obtained through $t_{ov}$ as:

$$Q_{sc}^c = Q_{sc}^0 e^{-\left(2\frac{t_{in} - t_{ov}}{t_{sc}}\right)^2}$$

(10)

where $t_{in}$ is the time at which the nMOS transistor starts to conduct and $T_{sc}$ is the time during which both transistors are conducting simultaneously. If a linear transition is assumed at the input of the buffer with input rise time $t_{in} (V_{in} = V_{DD})$, we obtain $T_{sc}$ and $t_{ov}$ as:

$$t_{ov} = t_{in} \left( 1 - \frac{V_{in}}{V_{DD}} + \frac{V_{TP}}{V_{DD}} \right)$$

(11)

When no-overshooting effects are present ($t_{ov} = t_{in}$) then $Q_{sc}^0 = Q_{sc}^c$ in eq. (10). When overshooting is significant and $t_{ov} > t_{in} + T_{sc}$ the short-circuit charge vanishes $Q_{sc}^c \sim 0$. An analytical expression for $Q_{sc}^0$ can be obtained from the model in [4] where a simple closed-closed form expression for SCCT assuming $C_M = 0$ is developed.

In Fig. 3 we plot SCCT vs. coupling capacitance for various values of the output capacitor (from $C_{min}$ to 50$C_{min}$, where $C_{min}$ is the minimum capacitor allowed by the technology) for a 0.35$\mu m$ technology. As can be appreciated, the model developed is close to HSPICE simulations.

Finally, the energy associated to the SCCT in a rising input edge is computed as:

$$E_{sc}^c = Q_{sc}^c V_{DD}$$

(12)

C. Overshoot time

The solution of eq.(6) to get $t_{ov}$ is a non-linear problem and no-closed form expression can be found for this time. To avoid numerical procedures we compute the time at which the overshoot current is maximum $t_{ov}^{max}$, and relate $t_{ov}$ to this value.
Figure 3 – Short-circuit charge transference vs. coupling capacitor for various values of $C_L$ for a 0.35µm technology.

Figure 4 – HSPICE simulations of $t_{ov}$ and $t_{ov}^{max}$ for a 0.18µm and a 0.35µm process technology. A minimum sized inverter driving an other minimum sized inverter are considered with different input rise times.

Fig. 4 shows HSPICE simulations of $t_{ov}$ and $t_{ov}^{max}$ for a 0.18µm and a 0.35µm process technology obtained for a minimum sized inverter driving an other minimum sized inverter. The different simulations are obtained varying the input rise time. A linear relationship is obtained with the form:

$$t_{ov} = t_n + \frac{3 (t_{ov}^{max} - t_n)}{2}$$

The maximum overshoot time is obtained from eq.(6) neglecting the short-circuit current. At $t = t_{ov}^{max}$ the nMOS transistor is saturated and $\frac{dV_{in}}{dt} = 0$. Then eq. (6) is reduced to:

$$C_M \frac{dV_{in}}{dt} - I_{D0n} = 0$$

The input voltage can be described with a linear ramp. For a low to high input transition we have:

$$V_{in}(t) = V_{DD} \frac{t}{t_{in}}$$

Figure 5 – Energy dissipation vs. input to output time ratio for different $W_p/W_n$ ratios for a 0.35µm technology. As can be appreciated, when $t_{in}/t_{out} > 1$ the short-circuit component is dominant

where $t_{in}$ is the input rise time. Then, eq.(14) can be expressed as:

$$C_M \frac{V_{DD}}{t_{in}} - I_{D0n} \left( \frac{V_{DD}^{max} - V_{TN}}{V_{DD} - V_{TN}} \right) n_n = 0$$

where $I_{D0n}$ is the parameter $I_{D0}$ of the nMOS transistor while $n_n$ is the velocity saturation index of nMOS. Solution to eq. (16) leads to:

$$t_{ov}^{max} = t_n + \frac{t_{in}}{V_{DD}} \left( V_{DD} - V_{TN} \right) \left( \frac{C_M V_{DD}}{I_{D0n t_{in}}} \right)$$

Once $t_{ov}^{max}$ is obtained, $t_{ov}$ is given by eq.(13).

IV. Results

We plotted the model prediction vs. HSPICE level 50 (MM9) simulations for a 0.35µm technology, considering energy vs. input transition time.

In Fig. 5 we show the energy dissipated per one cycle period vs. the input to output time ratio for a 0.35µm technology. Four different inverters are considered with different values of the $W_p/W_n$ ratio and for two values of the output capacitor (each size and capacitance is indicated in the graph). The input time $t_{in}$ ranges from 20ps to 1.5ns (that is, from 0.2$t_{out}$ to 3.8$t_{out}$). We plot HSPICE simulations (symbols) and the model in [4] with the two formulations, the original model and a modified version taking into account the overshooting model developed. As can be appreciated a significant improvement can be obtained using our formulation.

V. Conclusions

A detailed description of coupling capacitor effects on the short-circuit and transient power have been
developed. The contribution of the short-circuit power dissipation for high and low speed transitions have been described with detail. Non-linear problems are solved using simple formulas and avoiding time-consuming numerical procedures. The model is applied to a recently published short-circuit power model [4] and compared to HSPICE simulations (level 50) for a 0.35μm process technology reporting a higher accuracy respect the original model.

ACKNOWLEDGMENT

This work has been supported by the Spanish “Comisión Interministerial de Ciencia y Tecnología” under the project CICYT-TIC98-0284.

REFERENCES


A Compact Charge-Based Propagation Delay Model for Submicronic CMOS Buffers

José Luis Rossello and Jaume Segura

Departament de Física, Universitat Illes Balears, 07071 Palma de Mallorca, Spain

Abstract. We provide an accurate analytical expression for the propagation delay and the output transition time of submicron CMOS buffers that takes into account the short-circuit current, the input-output coupling capacitance, and the carrier velocity saturation effects, of increasing importance in deep-submicron technologies. The model is based on the nth-power law MOSFET model and computes the propagation delay from the charge delivered to the gate. Comparison with HSPICE level 50 simulations and other previously published models for a 0.35µm and a 0.18µm process technologies show significant improvements over previously-published models.

1 Introduction

Timing analysis is one of the most critical topics in VLSI design. The nonlinear behavior of CMOS gates requires numerical procedures for accurate timing analysis at expenses of large computation times. Moreover, the impact of design parameters such as fan in, fan out or transistor sizes on the propagation delay are difficult to understand and optimize using numerical procedures.

The dynamic behavior of submicron CMOS buffers depends on several nonlinear effects like the velocity saturation of carriers due to the high electric fields in submicron technologies, the short circuit current appearing when both pMOS and nMOS transistors are conducting simultaneously [1], and the additional effect of the input-output coupling capacitance [2].

Several methods have been proposed to derive the delay of CMOS buffers [2]-[7] as a first step to describe more complex gates [8,9]. Cocchini et al. [3] obtained a piece-wise expression for the propagation delay based on the BSIM MOSFET model [10]. The model included overshooting effects (due to the input-to-output coupling capacitance) while the short-circuit current was neglected. In [2] and [4] K.O Jeppson and L. Bislounis presented a model for the output response of CMOS buffers using a quadratic current-voltage dependence for MOSFET devices, which is not longer valid for submicron technologies. Daga et. al. [5] obtained a simple empirical expression for the propagation delay taking into account both overshooting and short-circuit currents using six fitting parameters. The relative error of this model was 19% for a 0.6µm technology as reported in [5]. Hirata et al. [6] derived a delay model based on the nth-power law MOSFET model [11] considering both short-circuit and overshooting currents and
using numerical procedures. The model provides an accurate description for the propagation delay but the numerical procedures used increases the computation time considerably. Bisdounis et al. [7] developed a piece-wise solution with seven operation regions for the transient response of a CMOS inverter based on the $\alpha$-power law MOSFET model [12] including both overshooting and short-circuit currents. In [8] T. Sakurai et. al. obtained a simple expression for the propagation delay of CMOS gates based on their $n$th-power law MOSFET model neglecting both short-circuit and overshooting currents.

In this work we propose a compact analytical model to accurately compute the propagation delay and the output transition time of a CMOS buffer accounting for the main effects of submicron technologies as the input-output coupling capacitance, carriers velocity saturation effects and short-circuit currents. The model is based on an accurate physically-based $n$th-power law MOSFET model [13] and on a power dissipation model for CMOS inverters [14]. Comparisons with HSPICE level 50 simulations and previously published models for a 0.35$\mu$m and a 0.18$\mu$m process technologies are reported showing significant improvements in terms of accuracy.

![Fig. 1.](image)

**Fig. 1.** In this figure we show the CMOS current and voltage switching characteristics.

This paper is organized as follows: In Section 2 the CMOS buffer switching characteristics are analyzed with detail and the MOSFET model used is presented. The delay and the output transition time models are developed in Section 3 and compared to HSPICE simulations and other previously published models for a 0.35$\mu$m and a 0.18$\mu$m process technology in section 4. Finally in section 5 we conclude the work.
2 Analysis of the CMOS Buffer Switching Characteristics

The dynamic behavior of a CMOS buffer is described by the next equation:

\[
(C_L + C_M) \frac{dV_{out}}{dt} = I_p - I_n + C_M \frac{dV_{in}}{dt}
\]  

(1)

where \(C_L\) is the output capacitance, \(V_{out}\) and \(V_{in}\) are the output and input voltage respectively, while \(I_p\) and \(I_n\) are the current that crosses the pMOS and the nMOS transistor respectively. \(C_M\) is the input to output coupling capacitance which is voltage dependent. The static value of \(C_M\) when the input is low \((C_M^L)\) is computed considering the side-wall capacitances of both transistor drains and the gate to drain capacitance of the pMOS transistor that operates in the linear region as:

\[
C_M^L = C_{ox} \left( \frac{W_{peff} L_{peff}}{2} + L_{Dn} W_{n_{eff}} + L_{Dp} W_{p_{eff}} \right)
\]

(2)

with \(W_{p_{eff}}\) and \(W_{n_{eff}}\) being the effective channel width of pMOS and nMOS respectively, \(L_{peff}\) is the effective channel length of pMOS, while \(L_{Dn}\) and \(L_{Dp}\) are the gate-drain underdiffusion for the nMOS and pMOS transistors respectively. For a static input high the capacitance \(C_M^{H}\) is obtained similarly.

In this work a mean value for the coupling capacitance during the transition \((C_M = 0.5 (C_M^L + C_M^{H}))\) will be used.

Fig. 1 illustrates the input and output voltage evolution of the buffer along with the current through the nMOS and pMOS transistors for a low to high input transition. The current through the pMOS transistor \((I_p\) in Fig. 1) has two components clearly distinguished by the sign of the current. The negative pMOS current is due to a partial discharge of the output capacitance from the output node toward the supply rail and appears when the input-output capacitance drives the output voltage below the supply voltage \((V_{DD})\) at the beginning of the transition \(\Delta\) (this effect is known as overshooting). When the nMOS device starts to conduct, it pulls the output voltage down. Once the output voltage goes below \(V_{DD}\), the pMOS current is positive corresponding to the short-circuit component due to the simultaneous conduction of both devices.

The propagation delay (defined as \(t_{PHL}\) for a high to low output transition) is typically defined as the time interval from the 50% \(V_{DD}\) input voltage to the 50% \(V_{DD}\) output voltage. The dependence of the propagation delay with design parameters is non-linear and difficult to model given that eq. (1) can not be solved in a closed form even using the simple Shockley MOSFET model [15]. Moreover carrier saturation effects become important with technology scaling and more complex MOSFET models accounting for such effects must be considered.

The \(nth\)-power law MOSFET model [11] is a widely used short-channel drain current model, and will be used in this work to derive the propagation delay and the output transition time of CMOS inverters. The drain current is expressed as:
\[ I_D = \begin{cases} 0 & (V_{GS} \leq V_{TH}) \\ (2 - \frac{V_{DD}}{V'_{D0}}) V_{DD} I'_{D0} & (V_{DS} < V'_D) \\ (V_{DS} \geq V'_D) \end{cases} \]

with

\[ I'_{D0} = I_D \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^n \]

where \( V_{GS}, V_{DD}, \) and \( V'_{D0} \) are the gate, supply, and saturation voltage respectively and \( I'_{D0} \) is the drain current at \( V_{GS} = V_{DS} = V_{DD} \). The parameter \( n \) is the velocity saturation index that ranges between 2 (long-channel devices) and 1 (short-channel) [11]. The saturation voltage \( V'_{D0} \) is given by:

\[ V'_{D0} = V_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^m \]

The parameter \( V_{D0} \) is the saturation voltage at \( V_{GS} = V_{DD} \), while \( m \) and \( V_{TH} \) are empirical parameters [11]. These equations are mathematically simpler than physically-based MOSFET models such as BSIM3v3 or MM9 with the disadvantage that, in the original model developed by Sakurai and Newton, the relationship between the empirical and the process parameters supplied by manufacturers is not provided. Therefore the variation of \( nth \)-power law model predictions with key parameters like the supply voltage are not taken into account in the original formulation performed by Sakurai and Newton, where each parameter must be recomputed if the supply voltage or some device dimension are changed. In this work we use the physical formulation proposed in [13] and used in [14]. This physical formulation provides an analytical relationship between the \( nth \)-power law parameters and the more accurate MM9 model parameters (that take into account the parameter variations with the supply voltage, MOSFET dimensions and temperature).

### 3 Delay Model

We compute the propagation delay when the input voltage switches. The short-circuit and overshooting currents are first neglected and incorporated later. Assuming a linear variation of the input voltage with rise time \( t_{in} \), then \( V_{in}(t) \) is:

\[ V_{in}(t) = V_{TN} + (V_{DD} - V_{TN}) \frac{t - t_n}{t_{in} - t_n} \]

where \( V_{TN} \) is the nMOS threshold voltage, \( t_n \) is the time when the nMOS chain starts to conduct \( t_n = V_{TN} t_{in}/V_{DD} \) and \( t_{in} \) is the input rise time. At the beginning of the transition the nMOS is off and \( V_{out} = V_{DD} \). At \( t = t_n \), the nMOS starts to conduct and the output voltage is obtained solving:

\[ C_L \frac{dV_{out}}{dt} = -I_n \]

where \( C_L \) is the total output capacitance of the gate, and \( I_n \) is the current through the nMOS transistor. An analytical solution to [7] is possible if the
A Compact Charge-Based Propagation Delay Model

nMOS transistor is assumed to be in the saturation region (valid while $V_{out} > V'_{D0n}$):

$$V_{out} = V_{DD} - \frac{I_{D0n}}{C_L} \left( \frac{t - t_n}{t_{in} - t_n} \right)^{n+1} \frac{t_{in} - t_n}{n+1}$$  \hspace{1cm} (8)

Equation (8) is used to obtain the propagation delay from the input at 0.5$V_{DD}$ to the output at 0.5$V_{DD}$.

$$t_{pHL_1} = t_n + \left[ \frac{Q_f (n+1)}{I_{D0n}} \right]^{1/n} \left( t_{in} - t_n \right)^{\frac{n}{n+1}} - \frac{t_{in}}{2}$$  \hspace{1cm} (9)

where $Q_f = C_L V_{DD}/2$ is the charge transferred by the nMOS transistor when the output reaches $V_{DD}/2$ while parameters $I_{D0n}$ and $n$ are the maximum saturation current and the velocity saturation index of nMOS respectively. Equation (9) is valid for slow inputs (defined when $t_{pHL} \leq t_{in}/2$). If $Q_{f0}$ is defined as the total charge transferred through the nMOS transistor when $t_{pHL} = t_{in}/2$ then eq. (9) is valid in the interval $Q_f < Q_{f0}$. The parameter $Q_{f0}$ is obtained equating $t_{pHL_1} = t_{in}/2$ and solving $Q_{f0}$

$$Q_{f0} = \frac{I_{D0n}}{(n+1)} (t_{in} - t_n)$$  \hspace{1cm} (10)

If the input reaches the supply voltage before the output is at $V_{DD}/2$ then $Q_f > Q_{f0}$ and eq. (9) is no longer valid. The propagation delay when $t_{pHL} \geq t_{in}/2$ (fast input range) can be obtained by solving (7) for $I_n = I_{D0n}$ leading to:

$$t_{pHL_2} = \frac{t_{in}}{2} + \frac{Q_f - Q_{f0}}{I_{D0n}}$$  \hspace{1cm} (11)

Equation (11) is valid when $Q_f > Q_{f0}$ (fast input range). For simplicity, the proposed model for the propagation delay (eqs. (10) and (11)) does not take into account the fact that the nMOS transistor is in the linear region when $V_{out} > V'_{D0n}$. For the evaluation of the output fall time ($t_f$) we first compute the output voltage slope at $V_{DD}/2$ from (8)

$$\left. \frac{dV_{out}}{dt} \right|_{V_{DD}/2} =
\begin{cases}
\frac{-I_{D0n} V_{DD}}{2Q_f} \left( \frac{Q_f (n+1)}{(t_{in} - t_n) I_{D0n}} \right)^{\frac{n}{n+1}} & (Q_f < Q_{f0}) \\
\frac{-I_{D0n} V_{DD}}{2Q_f} & (Q_f > Q_{f0})
\end{cases}$$  \hspace{1cm} (12)

For the evaluation of the output fall time we use:

$$t_f = \left. \frac{V_{DD}}{\frac{dV_{out}}{dt}} \right|_{V_{DD}/2}$$  \hspace{1cm} (13)

The value of the output fall/rise time is important since this parameter is used as the input fall/rise time of the gates driven by the buffer.
3.1 Including Short-Circuit Currents

The model proposed cannot be used for static CMOS gates since short-circuit currents are not considered. In this work we include the short-circuit current contribution to the delay as an additional charge that must be transferred through the pull-down (pull-up) network during an output falling (rising) transition. This additional charge is computed from the short-circuit power model presented in [14].

Therefore, for an output falling transition, the charge transferred through the nMOS transistor is computed as

\[ Q_f = C_L V_{DD}/2 + q_{sc}^f, \]

where \( q_{sc}^f \) is defined as the short-circuit charge transferred during the falling output transition until \( V_{out} = V_{DD}/2 \). The analytical derivation of \( q_{sc}^f \) is complex given that this parameter depends on the relative switching speed between the input and the output. For an input transition faster than the output, \( q_{sc}^f \) can be modeled as the total short-circuit charge transferred (defined as \( Q_f^{sc} \)) given that when \( V_{out} = V_{DD}/2 \) the input is high and the short-circuit current ceased. For an input transition slower than the output then \( q_{sc}^f < Q_f^{sc} \). For both cases we assume that \( q_{sc}^f = \kappa Q_f^{sc} \), where \( \kappa \) is an empirical parameter that must be optimized for each technology. For the 0.35\( \mu \)m and the 0.18\( \mu \)m technologies considered we obtained \( \kappa = 0.45 \) and \( \kappa = 0.73 \) respectively.

3.2 Including Overshooting Effects

Similarly to the short-circuit currents case, overshooting currents effects are included into the delay model as an additional charge to be transferred through the nMOS transistor. For fast inputs, the charge injected through the coupling capacitor \( (C_M) \) when \( V_{out} = V_{DD}/2 \) is \( Q_{ov} = C_M V_{DD} \). For simplicity we assume the same value for \( Q_{ov} \) in the slow-input case. Therefore, the total charge that must be transferred through the nMOS transistor during a falling output transition is:

\[ Q_f = 0.5 [(C_L + 2C_M) V_{DD}] + \kappa Q_f^{sc} \tag{14} \]

Equation (14) must be used in eqs. (9), (11) and (13).

4 Results

We plotted model results vs. HSPICE level 50 simulations for a 0.35\( \mu \)m and a 0.18\( \mu \)m technologies. Results show the propagation delay for different values of the input transition time \( t_{in} \), the configuration ratio \( W_p/W_n \) and the supply voltage \( V_{DD} \).

In Fig. 2 we plot the propagation delay \( t_{pHL} \) vs. the input time \( t_{in} \) for different values of the \( W_n/W_p \) ratio for a 0.35\( \mu \)m technology. HSPICE simulations (dots) are compared to the model proposed and to a previous model [3]. Short-circuit currents are not taken into account in [3] leading to an underestimation of the propagation delay. The model in [3] provides a piece-wise solution of the propagation delay: depending on the input transition (fast or slow input transitions) it uses an approximated or an exact expression for the propagation delay.
The approximated propagation delay is used when the nMOS transistor changes from the saturation to the linear region and the input is rising (eq. (11) in [3]), otherwise an exact expression for the output response is used. This piece-wise solution leads to a discontinuity in the propagation delay when changing from one region to the other (see Fig. 2).

**Fig. 2.** Propagation delay vs. input rise time for different values of the configuration ratio. Short-circuit currents are not taken into account in [3].

Fig. 3 plots the propagation delay vs. the input rise time for a 0.18μm technology. When the $W_p/W_n$ ratio is small the propagation delay decreases when increasing the input rise time. The model proposed in this work (solid lines) and the previously-published in [6,7] provide a good approximation to HSPICE simulations (dots).

Fig. 4 is a plot of HSPICE simulations (dots) and model predictions of the propagation delay vs. the supply voltage. The model proposed in this work provides a better fitting than the models in [6,7]. The model in [7] uses Taylor series expansions for the output response. For large supply voltage values the input transition is slow with respect the output response and the output voltage crosses $V_{DD}/2$ when both nMOS and pMOS transistors are in saturation, (called region 3 in [7]) and the output response used to compute the propagation delay is described through a Taylor series expansion. For the voltage range $(0.6V < V_{DD} < 1V)$ the propagation delay is obtained computing the output voltage when the nMOS is saturated, the pMOS off and $V_{in} < V_{DD}$ (region 4 in [7]) and a quadratic Taylor series expansion of the output voltage around time $t_{1-p} = t_{in} \left(1 - \frac{V_{in}}{V_{DD}} - \frac{V_{tp}}{V_{DD}}\right)$ is used to compute the propagation delay. As $V_{DD}$ is further reduced, the accuracy of the approximated output voltage decreases because the time point $t_{1-p}$ used in the Taylor expansion is reduced. For low
values of the supply voltage, $V_{DD} < 0.6V$, the propagation delay is computed for the nMOS saturated and $V_{in} = V_{DD}$ (region 5A) and an exact solution for the output response is used. Therefore, there are two discontinuities: between regions 4-5A ($V_{DD} = 0.6V$) and between regions 3 and 4 ($V_{DD} = 1V$).

In Fig. 4, we plot HSPICE simulation of the output transition time $t_f$ for different values of the input rise time $t_{in}$ and the configuration ratio $W_p/W_n$. 

Fig. 3. Propagation delay vs. input rise time for different values of the configuration ratio for a 0.18μm technology.

Fig. 4. Propagation delay vs. supply voltage for a 0.18μm technology.
A maximum relative error of 15% is obtained between model predictions and HSPICE simulations. In general a good agreement is obtained with the proposed model.

![Graph showing output skew vs. input rise time for different technology sizes.]

**Fig. 5.** Propagation delay vs. supply voltage for a 0.18µm technology.

5 Conclusions

An accurate analytical expression to compute the propagation delay and the output transition time of CMOS buffers has been presented. The main effects present in current submicron CMOS technologies like the input-output coupling capacitance, carriers velocity saturation effects and short-circuit currents are taken into account in the analysis. The model is compared to HSPICE simulations (level 50) and other previously published works for a 0.18µm and a 0.35µm process technology reporting a high degree of accuracy. The model represents an improvement with respect to previously published works.

References


A Simple Analytical Description of Power Dissipation in Submicronic CMOS Buffers

Jose Luis Rossello and Jaume Segura
Physics Dept., Balearic Islands University, Cra. Valldemossa km 7.5
07010 Palma de Mallorca, Spain
email:{j.rossello.jaume.segura}@uib.es

Abstract- We present an accurate closed-form expression for the short-circuit power of sub-micron CMOS buffers based on the nth-power law MOSFET model [1]. The expression obtained accounts for the main effects in current sub-micron CMOS technologies as carrier velocity saturation effects, input-output coupling capacitor, output load, input slew time or device sizes. Results are compared to HSPICE simulations (level 50) and to other models previously published for a 0.18µm and a 0.35µm technology showing significant improvements.

I. INTRODUCTION

A growing fraction of power consumed by VLSI integrated circuits is due to the clock distribution network, I/O drivers and busses which are all based on inverters. The power description of CMOS inverters is also an important step to describe more complex gates. Hence, the analytical description of power dissipated in a CMOS inverter is of increasing importance.

Power dissipation in CMOS circuits has a dynamic and a static component. The dynamic dissipation is due to the charge/discharge of gate output load (defined as transient energy), and to the short-circuit current due to the supply-ground conducting path created during the transition [2].

Several works have focussed on modeling the short-circuit power consumption. Veedrick [2] obtained an expression for unloaded buffers. Saku- rai et al. [3] derived a model for long and short channel devices using their alpha-power law MOSFET model for unloaded buffers. Turgis et al. [4] derived an expression for the short-circuit and overshooting dissipation applicable only to deep-submicron technologies and introducing the equivalent capacitance concept allowing a direct and frequency-independent comparison of the different power components. More Recently, Nose et al. [5] derived a model for short-circuit power which takes short-channel effects into consideration concluding that the short-circuit power to total power ratio \(P_{sc}/P_{total}\) will not change with scaling if \(V_{TH}/V_{DD}\) is kept constant. In [6], Rosselló et al. develop an accurate power model for submicron CMOS buffers. Overshooting and short-circuit currents are taken into account in their analysis but the expression obtained are too complex for fast power estimation. In this work we propose a simple and accurate power model of CMOS buffers accounting for the main effects in submicronic technologies. A closed form expression for the total power is developed avoiding time-consuming numerical procedures, thus allowing the model to be applied for a fast power estimation of large circuits. The model is compared to HSPICE simulations (level 50) and to other models for a 0.18µm and a 0.35µm technology showing a significant improvement in accuracy. The paper is organized as follows: in section II we obtain an expression for the transient energy. Section III derives the short-circuit energy component. Section IV presents the results while Section V concludes the work.

II. TRANSIENT DISSIPATION

The transient energy (\(E_{tr}\)) corresponds to the energy dissipated when discharging the output capacitor. The output node charge is stored in both the input-output coupling capacitor \(C_M\) and the load capacitor \(C_L\). The energy of a discharging constant capacitor is simply \(E = QV/2\), where \(V\) is the voltage swing and \(Q\) the charge initially stored in the capacitor. The voltage swing of the output node in the buffer circuit is equal to the supply voltage \(V_{DD}\) while the charge transferred will be derived using charge conservation by computing the charge at the beginning of the transition minus the charge at its end.

The value of \(C_M\) in the static input low state (de- fined as \(C_M^{eq}\)) considering the side-wall capacitance of both transistor drains, and the gate to drain capac- itance of the pMOS transistor in the linear region
is given by:

\[ C_M^L = C_{ox} \left( \frac{W_{P,eff}L_{P,eff}}{2} + L_{A,P}W_{P,eff} + L_{A,P}W_{N,eff} \right) \]  

(1)

with \( L_{A,P} \) and \( L_{A,N} \) being the gate source/drain underdiffusion for the pMOS and nMOS transistors respectively, \( W_{N,eff} \) and \( W_{P,eff} \) are the nMOS and pMOS effective channel width and \( L_{P,eff} \) is the effective channel length of the pMOS transistor. For a static input high the capacitance \( C_M^H \) can be obtained similarly. An average value of these two expressions will be used for the coupling capacitance \( C_M = 0.5(C_M^L + C_M^H) \).

In a CMOS buffer the charge at the output node is stored in both the output and the coupling capacitor. This charge (defined as \( Q_{out} \)) can be expressed as:

\[ Q_{out} = (C_M + C_L)V_{out} - C_M V_{in} \]  

(2)

where \( V_{out} \) is the output voltage and \( V_{in} \) the input voltage.

The energy dissipated by the nMOS transistor for a high to low output transition is given by \( E_{tr} = \Delta Q_{out} V_{DD}^2/2 \) (energy dissipated for a discharging capacitor), \( \Delta Q_{out} \) is the charge transferred from the output node to ground through the nMOS transistor, and \( V_{DD} \) is the voltage swing. Thus, the transient energy is expressed as:

\[ E_{tr} = \left( Q_{out}(0) - Q_{out}(\infty) \right) \frac{V_{DD}^2}{2} \]  

(3)

where \( Q_{out}(0) \) is the charge at the output node at the beginning of the transition \( (t = 0) \) and \( Q_{out}(\infty) \) is the charge stored at the output node when the output transition is finished.

Using eq.(2) we obtain an expression for \( Q_{out}(0) \) and \( Q_{out}(\infty) \) as:

\[ Q_{out}(0) = (C_L + C_M)V_{DD} \]

\[ Q_{out}(\infty) = -C_M V_{DD} \]  

(4)

From eqs. (3) and (4) the transient energy for a high to low output transition is:

\[ E_{tr} = \frac{1}{2}(C_L + 2C_M) V_{DD}^2 \]  

(5)

III. SHORT-CIRCUIT POWER MODEL

We derive the short-circuit energy dissipated in a CMOS buffer (Fig. 1) for a low to high input transition (for a high to low input transition the model is equivalent).

The dynamic behavior of the circuit in Fig. 1 is described by:

\[ (C_L + C_M) \frac{dV_{out}}{dt} = I_p - I_n + C_M \frac{dV_{in}}{dt} \]  

(6)

where \( I_p \) and \( I_n \) are the pMOS and nMOS current respectively.

Using a simple short-channel MOSFET model from [1], and neglecting channel-length modulation effects, the drain current expression is:

\[ I_{DS} = \left\{ \begin{array}{ll} 0 & (V_{GS} \leq V_{TH}) \\ \left( 2 - \frac{V_{BS}}{V_{PD}} \right) V_{DS}^n I_{DO} & (V_{DS} < V_{DO}) \\ I_{DO} & (V_{DS} \geq V_{DO}) \end{array} \right. \]  

(7)

with:

\[ I_{DO} = I_{DO} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^n \]  

(8)

The saturation voltage \( V'_{DO} \) is given by [1]:

\[ V'_{DO} = V_{DO} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^m \]  

(9)

The parameter \( n \) is the velocity saturation index that takes a value between 2 (long-channel devices) and 1 (short-channel). Parameters \( I_{DO} \) and \( V_{DO} \) are the drain current and saturation voltage for \( V_{GS} = V_{DS} = V_{DD} \). Parameters \( n, m \) and \( V_{TH} \) are fitting parameters.

The input voltage is described with a linear ramp, for a low to high input transition it is expressed as:

\[ V_{in}(t) = V_{DD} \frac{t}{t_{in}} \]  

(10)

where \( t_{in} \) is the input rise time.

When the output capacitance is small (i.e. when the short-circuit current has a greater impact [2]) the circuit behavior is close to the inverter DC operation since \( I_p \approx I_n \). At the beginning of the transition, the pMOS transistor drives a current equal to the nMOS saturation current, while at the end of the transition the pMOS is saturated. In this particular case, the maximum current takes place when \( I'_{DO} = I'_{DO} \). Using eqs. (8) and (10) and knowing that \( V_{GS} = V_{in} \) for the nMOS and \( V_{GS} = V_{DD} - V_{in} \) for the pMOS, the time at which the short-circuit is maximum (defined as \( t_{\text{max}} \)) can be obtained solving:
\[ I_{D0\text{n}} \left( \frac{V_{DD} + I_{D0\text{n}}(t) V_{TN}}{V_{DD} - V_{TN}} \right)^{n_n} = I_{D0\text{p}} \left( \frac{V_{DD} - V_{TP} - I_{D0\text{p}}(t)}{V_{DD} - |V_{TP}|} \right)^{n_p} \]  

(11)

where \( I_{D0\text{n}} \) and \( I_{D0\text{p}} \) are the parameter \( I_{D0} \) of the nMOS and the pMOS transistor respectively, \( V_{TP} \) and \( V_{TN} \) are the pMOS and the nMOS threshold voltage while \( n_n \) and \( n_p \) are the velocity saturation index of nMOS and pMOS respectively. Equation (11) is non-linear and must be solved numerically. We obtained a good analytical approximation to the solution of eq. (11) as:

\[ t_{\text{max}}^L = t_n + t_m \left( \frac{1 - \frac{V_{TP}}{V_{DD}} - \frac{t_n}{t_m} + 1}{F_p^{t_m} + n_m} \right)^{n_p} \]  

(12)

where \( F_p \) is given by:

\[ F_p = \frac{I_{D0\text{n}}}{I_{D0\text{p}}} \left( \frac{V_{SC}}{V_{DD} - V_{TN}} \right)^{n_n} \left( \frac{V_{DD} - |V_{TP}|}{V_{SC}} \right)^{n_p} \]  

(13)

Then, the maximum short-circuit current for unloaded buffers \( I_{L0\text{max}} \) is evaluated as:

\[ I_{L0\text{max}} = I_{D0\text{p}} \left( \frac{V_{DD} - V_{DD} - I_{L0\text{max}}(t)}{V_{DD} - |V_{TP}|} \right)^{n_p} \]  

(14)

Eqs. (12) and (14) are inaccurate for large \( C_L \) values. If the output capacitor is large, then the short-circuit current is negligible. Thus, for the limit of \( C_L = \infty \) the maximum short-circuit current vanishes. To obtain an expression similar to eq. (14) for \( C_L = \infty \), the time at which the short-circuit is maximum is given by the solution of:

\[ I_{D0\text{p}} \left( \frac{V_{DD} - V_{DD} - t_m I_{L0\text{max}}(t)}{V_{DD} - |V_{TP}|} \right)^{n_p} = 0 \]  

(15)

That leads to:

\[ t_{\text{max}}^L = t_n + t_m \left( \frac{1 - \frac{V_{TP}}{V_{DD}}}{F_p^{t_m} + n_m} \right) \]  

(16)

The drain saturation current of the pMOS device evaluated at \( t_{\text{max}}^L \) leads to the expected value for the maximum short-circuit current.

Eq. (16) is in fact a limit value. In order to obtain a good description for large values of \( C_L \), we use a Taylor expansion in terms of \( 1/C_L \) of this time as:

\[ t_{\text{max}}^L = t_{\text{max}}^L - k_p I_{D0\text{n}} \frac{p_{t_m}}{C_L V_{DD}} + o \left( \frac{1}{C_L} \right) \]  

(17)

where \( k_p \) is a fitting parameter that must be extracted from SPICE simulations for each technology.

For a high-speed input transition (or a large capacitance at the output), the time at which the short-circuit current is maximum is given by eq.(17). For slow input transitions (unloaded buffers) the maximum time is given by eq. (12). We obtain a single expression that leads to eq. (12) for slow input transitions and to eq.(17) for high-speed:

\[ t_{\text{max}}^L = \frac{t_{\text{max}}^L - t_{\text{max}}^L}{t_{\text{max}}^L} \frac{I_{D0\text{n}}}{I_{D0\text{p}}} \left( \frac{V_{DD} - V_{DD} - I_{L0\text{max}}(t)}{V_{DD} - |V_{TP}|} \right)^{n_p} \]  

(18)

Using eq. (18) in the expression for the saturation current of pMOS (8), the maximum short-circuit current is obtained as:

\[ I_{\text{max}}^L = I_{D0\text{p}} \left( \frac{V_{DD} - V_{DD} - I_{L0\text{max}}(t)}{V_{DD} - |V_{TP}|} \right)^{n_p} \]  

(19)

Finally, the energy associated to the short-circuit current for a rising input transition is computed as:

\[ E_{\text{sc}}^L = \frac{1}{2} E_{\text{max}}^L I_{L0\text{max}} \]  

(20)

IV. RESULTS

We plotted the model prediction vs. HSPICE level 50 (MM9) simulations for a 0.18\( \mu \)m and a 0.35\( \mu \)m technology considering energy vs. input transition time. In each graph we also include the models from [4] and [5] which presents the better agreement with HSPICE. For the technology used, the empirical parameters \( k_p \) and \( k_n \) are taken as \( k_p = k_n = 0.03 \) and \( k_p = k_n = 0.1 \) for the 0.18\( \mu \)m and the 0.35\( \mu \)m technology respectively.

Fig.2 shows the energy dissipated per one cycle period vs. the input to output time ratio for a 0.18\( \mu \)m technology. The output time is taken from HSPICE simulations as an average value of the rise and fall times at the buffer output \( (t_f + t_r)/2 \), being \( t_f \) and \( t_r \) proportional to the time from 0.9\( V_{DD} \) to the output \( (t_90) \) to 0.1\( V_{DD} \) at the output \( (t_10) \) (i.e. \( t_f = 0.8(t_90 - t_10) \) ). Different \( W_p/W_n \) ratios are considered \( (W_p/W_n = 3\mu m/1.5\mu m, W_p/W_n = 2.25\mu m/2.25\mu m \) and \( W_p/W_n = 1.5\mu m/3\mu m) \) with \( C_L = 5C_{\text{min}} \). The output capacitance is the output capacitance of a minimum sized inverter driving another minimum sized inverter. The area of the three inverters is kept constant \( (W_p + W_n = 4.5\mu m) \).
range from 20ps to 1ns (from 0.5t_{out} to 4t_{out}). The model presented reports an improvement with respect to previous works (we include plots from the model in [4] since it showed the better fitting among the works considered) maintaining an accurate description of power for different values of W_p/W_n.

In Fig. 3 we show the energy dissipated per one cycle period vs. the input to output time ratio for a 0.35µm technology. Four different inverters are considered with different values of the W_p/W_n ratio and two output capacitance values (each size and capacitance is indicated in the graph). The input time t_{in} ranges from 20ps to 1.5ns. As can be appreciated, the model developed describes correctly HSPICE simulations for all the conditions considered.

We calculated the mean deviation of the Energy vs. input time from HSPICE simulations for the models in the literature [4, 5] and the model proposed. The percentage deviations are presented in Table I for the two technologies used showing the improvement achieved by the model.

V. CONCLUSIONS

An accurate analytical expression to calculate the power consumption of CMOS buffers has been developed. The model is compared to HSPICE simulations (level 50) and other previously published models for a 0.18µm and a 0.35µm process technology reporting a high accuracy. It is able to describe the energy dependence of the buffer with input slew time, asymmetry of the buffer and output capacitance. Power dissipation for high and low speed transitions have been described with detail. Nonlinear problems are solved using simple formulas and avoiding time-consuming numerical procedures. This model presents an average error of 6% respect HSPICE simulations for a wide range of parameter variation and represents an improvement over previously published models.

![Energy vs. input time](image1.png)

Figure 2 – Power dissipation vs. input time to output time ratio for a 0.18µm technology. Different W_p/W_n ratios are considered.

![Energy vs. input time](image2.png)

Figure 3 – Energy dissipation vs. input to output time ratio for different W_p/W_n ratios for a 0.35µm technology.

<table>
<thead>
<tr>
<th>Model</th>
<th>0.18µm</th>
<th>0.35µm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>0.18µm</td>
<td>0.35µm</td>
</tr>
<tr>
<td>Model (Max.)</td>
<td>14%</td>
<td>8%</td>
</tr>
<tr>
<td>Model (Mean)</td>
<td>6%</td>
<td>2%</td>
</tr>
<tr>
<td>4 (Max.)</td>
<td>33%</td>
<td>65%</td>
</tr>
<tr>
<td>4 (Mean)</td>
<td>12%</td>
<td>16%</td>
</tr>
<tr>
<td>5 (Max.)</td>
<td>62%</td>
<td>23%</td>
</tr>
<tr>
<td>5 (Mean)</td>
<td>19%</td>
<td>12%</td>
</tr>
</tbody>
</table>

Table I

**Mean and maximum error of energy estimation vs. HSPICE for a 0.18µm and a 0.35µm technology**

**REFERENCES**


Ph.D. Thesis entitled “Power and Timing Modeling of Sub-micron CMOS Gates”
developed by Jose L. Rosselló, j.rossello@uib.es. Directed by: Jaume Segura.

Power dissipation emerged as a major concern in digital IC design during the last decade becoming a design parameter as important as performance and area. Many low power design techniques at different levels (technological, design and architectural) were developed, all of them requiring accurate power estimation tools to achieve high-performance low-power digital systems.

Electrical level simulation methods are impractical to evaluate full chip performance for huge designs due to the enormous computation time required. The complex non-linear dependencies of logic gates delay with process and design parameters in traditional technologies are aggravated in deep-submicron ICs due to new physical effects that come into the picture. The traditional description of logic gates delay based on simplified linear models used by logic simulators (that are much more faster than electrical simulators) are not accurate enough to describe the behavior of present technology ICs.

In the Ph.D. thesis we develop new power and timing analytical models for sub-micron and deep sub-micron CMOS gates. These analytical models are of high interest for low-power digital design since they provide an explicit relationship between power/delay and process/design parameters (identifying the parameters with greatest influence on the circuit performance and power dissipation). These analytical models can be also incorporated in simulation tools to obtain accurate (very close to SPICE simulations) power/delay estimations for huge circuits.

The Thesis is organized in three parts devoted to: MOSFET devices, CMOS buffers and complex CMOS gates respectively. In the first part, a simple physically based MOSFET model [1] is developed. This model is a variation of the widely-used nth-power law MOSFET model developed by Sakurai and Newton that was oriented to circuit analysis. The proposed formulation allows a direct and simple relationship between physical and nth-power law parameters.

In the second part of the thesis, power and delay models [2,3] for sub-micron CMOS buffers are developed based on the physical nth-power law MOSFET model and on the charge transfer mechanisms involved in gate transitions. These models take into account short-channel effects of MOSFETs and can incorporate the effect of the interconnect line resistance [4]. Finally, in the third part we develop an nth-power law model for series-connected MOSFETs. This model is used to obtain a delay model of complex CMOS gates [5] and for power-delay optimization of dynamic gates [6].

The accuracy of the power/delay models developed in this thesis are within 5% of SPICE simulations for a wide range of circuit configurations and switching conditions.

Development of power and timing models for sub-micron CMOS gates

- Physical α-power law model for series-connected transistors
  - The α-power law model is suitable for circuit analysis
  - The empirical nature of parameters involved implies a lack in physical meaning
  - A relationship between physical and α-power law parameters is provided

Charge-based analytical models for the propagation delay and the output transition time of CMOS gates taking into account:
- Short-channel effects
- Short-circuit power
- Overshooting effects
- Temperature of operation

Current and future work

- Development of a compact IC electro-thermal simulation (Power, delay, crosstalk and temperature estimation inside the IC)
- Development of optimization techniques (Power, delay, crosstalk and temperature optimization inside the IC)

Basic bibliography

POWER


DELAY


APPLICATIONS


Development of power and timing models for sub-micron CMOS gates

- Development of CAD tools using all the analytical models developed.
- Technology transfer to the IC industry

APPLICATONs

- Development of optimization techniques (Power, delay, crosstalk and temperature optimization inside the IC)
Abstract. In this work we propose a compact analytical model to compute the crosstalk induced delay from a charge-based propagation delay model for submicronic CMOS gates. Crosstalk delay is described as an additional charge to be transferred through the pMOS (nMOS) network of the gate driving the victim node during its rising (falling) output transition. The model accounts for time skew between the victim and aggressor input transitions and includes submicronic effects. It provides an intuitive description of crosstalk delay showing very good agreement with HSPICE simulations for a 0.18μm technology.

1 Introduction

The constant scaling of feature sizes and supply voltage with the increase in both operating frequency and signal rise/fall times has made digital designs more vulnerable to noise. In modern ICs, interconnect coupling noise (crosstalk) becomes a performance limiting factor that must be analyzed carefully during the design process. If not considered, crosstalk can cause extra delay, logic hazards and logic malfunction.

Fig.1 illustrates the well known crosstalk induced delay effect that appears when two lines (the aggressor and the victim) switch simultaneously. For a victim node falling transition, the gate delay (defined as $t_{p\text{H} \rightarrow \text{L}}$) can be reduced ($t_{p\text{H} \rightarrow \text{L},\text{su}}$) or increased ($t_{p\text{H} \rightarrow \text{L},\text{sd}}$) depending on if the aggressor makes a falling or a rising transition respectively. Since large circuits can handle tens of millions of interconnect lines on a single chip, simple and accurate analytical descriptions for crosstalk delay are of high importance for a fast timing verification of the whole chip.

Crosstalk delay has been analyzed in [1] using a waveform iteration strategy. Unfortunately, the time skew between transitions are not considered and no-closed form expression for the worst-case victim delay is provided. More recently crosstalk delay has been modeled analytically in [2]. Analytical expressions are derived that quantify the severity of crosstalk delay and describe qualitatively the dependence on circuit parameters, the rise/fall times of transitions and the skew between transitions. The nMOS (pMOS) network is substituted by a pull-down (pull-up) resistance and short-circuit currents are neglected. The main limitation is that the complex equations obtained for the victim voltage variation cannot be used to obtain a closed-form expression of the propagation delay. Additionally, other effects as short-circuit currents, with a great impact on propagation delay [3] are not considered in their analysis.
Fig. 1. An additional delay in the victim is induced by the switching transition of \( V_{\text{out},v} \) (crosstalk induced delay).

In this work we propose a simple and accurate analytical model to compute the crosstalk delay \( (t_{\text{phL,v}} \) and \( t_{\text{phL,v}}) \) in submicron CMOS gates. A simple propagation delay model for CMOS inverters [3] is modified to include more complex gates than inverters using the collapsing technique developed in [4]. The model is compared to HSPICE simulations for a 0.18\( \mu \text{m} \) technology showing an excellent agreement.

2 Crosstalk delay model

Consider the circuit in Fig.1 where the aggressor and the victim gates drive an output capacitance \( C_{L,a} \) and \( C_{L,n} \) respectively, and there is a coupling capacitance \( C_c \) between \( V_{\text{out},a} \) and \( V_{\text{out},v} \) (that causes the crosstalk between the two lines). The time skew \( t_s \) is defined as the time interval from \( V_{\text{DD}}/2 \) at the victim input to \( V_{\text{DD}}/2 \) at the aggressor input. The output interconnect lines are modeled as lumped capacitances. This is a valid approach for medium size interconnects as it has been reported that the relative error of this capacitive model is below 10% for a 2\( \mu \text{m} \) wire in a 0.18\( \mu \text{m} \) technology [5].

The structure of this paper is as follows, first we present a compact charge-based propagation delay model of CMOS gates that is based on the propagation delay model developed in [3]. Finally, crosstalk effects are included in the model as additional charges to be transferred through the corresponding nMOS/pMOS block.
2.1 Propagation delay model for complex gates

The propagation delay is usually computed as the time interval between the input and the output crossing $V_{DD}/2$. The crosstalk-free propagation delay ($t_{pHLO}$) can be described as a function of the charge transferred through the gate using the model in [3] that is based on the nth-power law MOSFET model [6], and on an efficient and accurate transistor collapsing technique for complex gates developed in [4].

For a high to low output transition (for a low to high transition the analysis is equivalent) the propagation delay is given by:

$$t_{pHLO} = \begin{cases} t_n + \left[ \frac{Q_f (1 + n)}{I_{D0n}^2} (t_m - t_n)^n \right]^{\frac{1}{n}} - \frac{t_p}{2} & Q_f < Q_{f0} \\ t_n + \frac{Q_f - Q_{f0}}{I_{D0n}} & Q_f \geq Q_{f0} \end{cases}$$

(1)

where $t_n$ is the input transition time (i.e. the time during which the input is changing), parameter $n$ is the velocity saturation index of the nMOS transistors, $Q_f = C_L V_{DD}/2$ is the charge transferred through the nMOS block until $V_{out} = V_{DD}/2$ (where $V_{DD}$ is the supply voltage), $C_L$ is the output load of the gate, $t_n = (V_{TS}/V_{DD}) t_m$ is the time when the nMOS block starts to conduct ($V_{TS}$ is the threshold voltage of nMOS), and $Q_{f0}$ is the charge transferred through the nMOS block when the input transition is finished, given by:

$$Q_{f0} = \frac{I_{D0n} (1 + N_n) (t_m - t_n)}{1 + n}$$

(2)

Parameter $I_{D0n}^{(1,N_n)}$ is defined as the maximum current that the nMOS block can deliver (modeled as a chain of $N_n$ series-connected transistors) and is obtained from the transistor collapsing technique developed in [4]. For the simple case of a chain with $N_n$ identical transistors, the collapsing technique provides a simple expression for $I_{D0n}^{(1,N_n)}$:

$$I_{D0n}^{(1,N_n)} = \frac{I_{D0n} [1 + \frac{n}{2} (V_{DS} - V_{DD})]}{1 + (N_n - 1) K_n}$$

where $V_{DS}$ is the saturation voltage of nMOS, $I_{D0n}$ is the maximum saturation current of each nMOS (drain current when $V_{GS} = V_{DS} = V_{DD}$) and parameter $\lambda$ accounts for channel length modulation. The parameter $K_n$ is a technology-dependent parameter given by:

$$K_n = \frac{3 V_{DS} n (1 + \gamma_n)}{5 (V_{DD} - V_{TS})}$$

(3)

where $\gamma_n$ is the body effect parameter of nMOS [6].

The transition time at the output $t_{out0}$, required to evaluate crosstalk delay, can also be expressed as a function of $Q_f$ as [3]:

$$t_{out0} = \begin{cases} \frac{2 Q_f}{I_{D0n}^{(1,N_n)}} \left[ \frac{I_{D0n}^{(1,N_n)}}{Q_f (1 + n)} (t_m - t_n) \right]^{\frac{1}{n}} & Q_f < Q_{f0} \\ \frac{2 Q_f}{I_{D0n}^{(1,N_n)}} & Q_f \geq Q_{f0} \end{cases}$$

(4)
Fig. 2. Crosstalk delay is dependent on the time between transitions $V_{o_{in,a}}$ and $V_{o_{in,v}}$.

Eqs. (1) and (4) are valid for a high to low output transition. Equivalent expressions can be obtained for a low to high output transition.

2.2 Including short-circuit currents

The delay model expressed in eq. (1) is valid for CMOS gates when short-circuit currents are neglected. These currents are included in [3] as an additional charge to be transferred through the pull-down (pull-up) network for an output falling (rising) transition. For a high to low output transition, the charge transferred through the nMOS block is computed as $Q_f = C_L V_{DD}/2 + q_{dc}^f$, where $q_{dc}^f$ is defined as the short-circuit charge transferred during the falling output transition until $V_{out} = V_{DD}/2$. For simplicity we compute this charge as $q_{dc}^f = Q_{dc}^f/2$, where $Q_{dc}^f$ is the total short-circuit charge transferred. For the evaluation of $Q_{dc}^f$ we use a previously developed model described in [7].

2.3 Crosstalk delay

The impact of crosstalk on the propagation delay is computed accounting for the additional charge injected by the aggressor driver through the coupling capacitance $C_c$ that must be discharged by the victim driver. We consider a falling transition at $V_{out,v}$ slower than a rising transition at $V_{out,a}$ (for other cases the analysis is similar).

Fig.2 shows a high to low transition of the victim node $V_{out,v}$ when the aggressor output switches from low to high for six different skew times represented as $a, b, $, $f$. For case ‘a’, the aggressor transition ($A_{g_a}$) does not impact the victim propagation delay ($t_{p_{HLO}}$) since the absolute value of the time skew between
transitions is too large. The victim output $V_{\text{out},v}$ (transition $V_{\text{out},v(\text{on})}$) rises until $V_{\text{out},a} = V_{DD}$ and then falls back to $V_{DD}$ before the rising transition at the victim input is initiated. In this case the crosstalk coupling effect is a glitch that lasts for a time that is defined as $t_D$ (see Fig.3). This characteristic time ($t_D$) is computed assuming that the ON transistors of the victim gate pMOS block (that discharge the output) are in the linear region and that their drain-source voltage is small. Under these conditions, the current through the pMOS block is:

$$I_p = I_{DO_p} \frac{V_{DD} - V_{\text{out}}}{V_{DO_p}}$$  \hspace{1cm} (5)$$

where $V_{DO_p}$ is the saturation voltage of pMOS and $I_{DO_p}$ is computed from all the ON transistor chains that connect the supply and the output node. Each of these pMOS chains is collapsed to a single equivalent transistor with maximum saturation current $I_{DO_p}^{(1,N_i)}$ (where $N_i$ is the number of transistors in each chain). Finally $I_{DO_p}$ is expressed as the sum of these contributions as:

$$I_{DO_p} = \sum_{i=1}^{m} I_{DO_p}^{(1,N_i)}$$  \hspace{1cm} (6)$$

where $m$ is the number of active chains connecting $V_{DD}$ and $V_{\text{out},v}$. The time point at which $V_{\text{out},a} = V_{DD}$, the victim line output voltage evolution ($V_{\text{out},v}$) is described as:

$$V_{\text{out},v} = V_{DD} + (V_{\text{max}} - V_{DD}) e^{-\frac{t - t_0}{\tau_{\text{out}}}}$$  \hspace{1cm} (7)$$

where $C_L = C_{L,v} + C_c$, $t_0$ is the time at which $V_{\text{out},a} = V_{DD}$ and $V_{\text{max}}$ is the maximum voltage at $V_{\text{out},v}$. The characteristic time $t_D$ (relaxation of $V_{\text{out}}$ back to $V_{DD}$) is obtained from (7) leading to:

$$t_D = 2 \ln \left(\frac{V_{DO_p} C_L}{I_{DO_p}}\right)$$  \hspace{1cm} (8)$$

It is well known that crosstalk may have an impact on delay only when the aggressor and victim transitions occur within a time window. Otherwise the effect of crosstalk is simply a glitch at the victim line. The time interval during which the coincidence of input transitions may give a crosstalk delay can be defined in terms of a limit time skew value ($t_s$). This limit value will depend on the propagation delay of the aggressor driver, the characteristic time $t_D$, and the aggressor output transition time. Fig.3 illustrates this time relationships graphically. The limit time $t_s$ is obtained equating the time point at which the glitch at $V_{\text{out},v}$ is finished to the beginning of the high to low transition at $V_{\text{out},v}$, i.e.:

$$t_s + t_{pLH0,a} + \frac{t_{\text{out}0,a}}{2} + t_D = t_{pHH0,v} - \frac{t_{\text{out}0,v}}{2}$$  \hspace{1cm} (9)$$

where $t_{pLH0,a}$ and $t_{\text{out}0,v}$ are the propagation delay and the output transition time of the victim gate respectively that are obtained from (1) and (4) when crosstalk is neglected ($Q_f = C_L V_{DD}/2 + Q_{sc}/2$). If $t_s < t_s$, the effect of crosstalk
Fig. 3. The aggressor starts to affect the victim voltage variation when \( t_s = t_{s1} \).

is a glitch at the victim node and the voltage value is restored by the pMOS devices. When \( t_s > t_{s1} \) the crosstalk will impact delay as an additional charge (defined as \( Q_c \)) that must be drained by the nMOS devices during the transition.

When \( V_{\text{out,}\alpha} \) and \( V_{\text{out},v} \) start switching at the same time, the value of \( Q_c \) is maximum. Defining \( t_{s2} \) as the time skew at this time point, then we have:

\[
t_{s2} + t_{\text{pLH}0,\alpha} - \frac{t_{\text{out}0,\alpha}}{2} = t_{\text{pLH}0,v} - \frac{t_{\text{out}0,v}}{2}
\]

(10)

For \( t_s = t_{s2} \), the coupling charge can be taken as \( Q_c = C_rV_{DD} \) since the voltage variation at \( V_{\text{out,}\alpha} \) occurs during the transition at \( V_{\text{out},v} \). For the case at which \( t_{s1} < t_s < t_{s2} \) we use a linear variation of \( Q_c \) with \( t_s \).

\[
Q_c = \frac{t_s - t_{s1}}{t_{s2} - t_{s1}} C_r V_{DD}
\]

(11)

The coupling charge \( Q_c \) is equal to \( C_rV_{DD} \) for a time skew greater than \( t_{s2} \). This condition holds if \( V_{\text{out,}\alpha} \) transition finishes before \( V_{\text{out},v} \) reaches \( V_{DD}/2 \). This limit value for time skew is defined as \( t_{s3} \) and obtained solving:

\[
t_{s3} + t_{\text{pLH}0,\alpha} + \frac{t_{\text{out}0,\alpha}}{2} = t_{\text{pLH}0,v}
\]

(12)

If \( t_{s1} < t_s < t_{s3} \) (case d in Fig. 2) then \( Q_c \) is maximum (\( Q_c = C_rV_{DD} \)). When \( t_s > t_{s3} \) \( Q_c \) starts to decrease until \( Q_c = 0 \) at \( t_s = t_{s4} \), where \( t_{s4} \) is defined as the time at which the beginning of the transition at \( V_{\text{out,}\alpha} \) and the time at which \( V_{\text{out},v} = V_{DD}/2 \) are equal.

\[
t_{s4} + t_{\text{pLH}0,\alpha} + \frac{t_{\text{out}0,\alpha}}{2} = t_{\text{pLH}0,v}
\]

(13)
In the interval $t_{s_3} < t_s < t_{s_4}$ we use a linear variation of $Q_c$ with $t_s$ as:

$$Q_c = C_c V_{DD} \frac{t_s - t_{s_4}}{t_{s_3} - t_{s_4}}$$  \hspace{1cm} (14)

Finally, for $t_s > t_{s_4}$ (transitions e and f in Fig.2) we use $Q_c = 0$. For the case at which $V_{out,a}$ makes a high to low transition, speeding up the propagation delay, a similar model can be obtained changing the sign of $Q_c$. Once $Q_c$ is obtained, the propagation delay ($t_{pHL_{su}}$ and $t_{pHL_{ld}}$) at $V_{out,v}$ is obtained using $Q_f = C_L V_{DD}/2 + Q_{Ic}/2 + Q_c$ in (1).

3 Results

We compare the model results to HSPICE simulations for a 0.18μm technology. In Fig.4 we plot the propagation delay limit bounds for a 3-NAND gate for different values of the coupling to load capacitance ratio $C_c/C_L$. For these cases $t_{s_2} < t_s < t_{s_3}$ and therefore $Q_c = C_c V_{DD}$. The output load value ($C_L$) is obtained by adding the fan-out of a 3-NAND gate to the coupling capacitance ($C_L = C_{L,o} + C_c$). Therefore, as $C_c$ increases, the delay $t_{pHL_{LO}}$ is larger. Fig.4 shows an excellent agreement between HSPICE simulations (using the BSIM3v3 MOSFET model) and the proposed crosstalk delay model.

In Fig.5 we plot the propagation delay of the 3-NAND gate when varying the time skew between $V_{in,v}$ and $V_{in,a}$. Different channel width values of the nMOS transistors of the 3-NAND gate are selected (the pMOS sized appropriately) showing that the proposed model can describe a wide range of design choices.
Fig. 5. In this figure we show the influence of the time skew between transition $t_a$ and the propagation delay $t_{pHL}$. Model predictions are HSPICE simulations while solid line is the proposed model.

4 Conclusions and Future work

A simple description for the evaluation of crosstalk delay has been presented. The model is useful for a fast and accurate signal integrity simulation of large ICs. A very good agreement is achieved between model predictions and HSPICE simulations for a 0.18μm technology. The description accounts for short-circuit currents and short-channel effects and can be applied to multiple input gates. Experimental measurements for a more faithful validation of the proposed model are under development.

ACKNOWLEDGMENT

This work has been supported by the Spanish Government under the project CICYT-TIC02-01238.

References


A physically-based nth power law MOSFET model for efficient CAD implementation

Jose L Rosselló and Jaume Segura
Physics Dept. Universitat Illles Balears, 07122 Palma de Mallorca, Spain

Abstract

In this paper we develop a simple description for the nth-power law MOSFET model to incorporate this analytical description within CAD tools in an efficient way. This fully analytical physically-based description allows fast and accurate computation of circuit level delay and power without the need to extract the empirical nth-power law parameters at different device sizes, supply voltages or temperatures. The model is compared with respect to HSPICE simulations (level 50) for a 0.18μm technology showing excellent accuracy.

Keywords: Delay models, MOSFET models, Circuit modeling

I. INTRODUCTION

The nth-power law MOSFET model [1] is a short-channel drain current model widely used to derive circuit delay or power dissipation [1]-[5] due to its simplicity. The model takes into account the velocity saturation effect, which becomes dominant in short-channel devices, giving an excellent intuitive understanding of the relationship between the drain saturation current and the gate voltage. However, the empirical nature of the parameters included in the model (that must be extracted from measurements or SPICE simulations) makes this model unfeasible for high-level circuit simulations since an additional computational effort for parameter extraction is required when key variables like device sizes, supply voltage or temperature change. A physical alpha-power law MOSFET model was developed in [6] providing a more complicated relationship between the drain saturation current and the gate voltage. The complexity of the expressions obtained in [6], although accurate, cannot be used to compute delay or power using the models based on the traditional power law model [1]-[5].

In this work we present a physical description of the nth-power law MOSFET model that provides a physical meaning to many gate delay and power models. The main objective of this work is not to develop a new MOSFET model but to provide a basic formulation that would allow direct computation of delay and power at the circuit level without the need to recompute the empirical parameters when key variables like device sizes, the supply voltage or temperature change.

II. THE N-TH POWER LAW MOSFET MODEL

Current MOSFET models are too complex to derive analytical expressions for the propagation delay or the power dissipation of CMOS gates. For this reason, T. Sakurai and R. Newton derived a single short-channel MOSFET model suitable for circuit analysis (the nth power law MOSFET model). This model [1] is expressed as:

\[
I_D = \begin{cases} 
0 & (V_{GS} \leq V_{TH}) \\
(2 - \frac{V_{DS}}{V_{DD}}) \frac{V_{DS}}{V_{DD}} I_D' & (V_{DS} < V_{DD}) \\
I_D' & (V_{DS} \geq V_{DD})
\end{cases}
\]  

where \( I_D' = I_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^m (1 + \lambda (V_{DS} - V_{DD})) \)

and \( V_{GS}, V_{DD} \) and \( V_{D0}' \) are the gate, supply and saturation voltages respectively. \( I_{D0} \) is the drain current at \( V_{GS} = V_{DS} = V_{DD} \). The parameter \( n \) (which is empirical) is the velocity saturation index that ranges between 2 (long-channel devices) and 1 (short-channel) [1], and \( \lambda \) describes the channel length modulation. The saturation voltage \( V_{D0}' \) is expressed as:

\[
V_{D0}' = V_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^m
\]

The parameter \( V_{D0} \) is the drain saturation voltage at \( V_{GS} = V_{DD} \), while \( m \) and \( V_{TH} \) are empirical parameters [1]. This simple MOSFET model is the basis of many delay and power models of CMOS gates.
Figure 1 – Drain current vs. drain voltage for a short channel nMOS transistor (symbols are SPICE simulations and solid lines are model predictions).

\[
W_{eff} = W + W_{VAR} - 2W_{OT} \\
L_{eff} = L + L_{VAR} - 2L_{AP}
\]  

where \( W_{VAR} \) and \( L_{VAR} \) are the process bias on the channel width and length respectively while \( W_{OT} \) is the isolation reduction of channel width.

\[ E_{sat} \text{ corresponds to the critical electric field at which the carrier velocity becomes saturated and } \theta \text{ is the gate field mobility reduction coefficient. A physical description for } V'_{DO} \text{ is given by [7]:} \]

\[
V'_{DO} = \frac{2}{1+\delta} \left( 1 + \left( \frac{2V_{DD}}{V_{VTR}^2 + 1} \right) \right) \]

where \( \delta \) accounts for body bias effects.

Our goal is to relate the nth power law parameters \( (I_{DO}, V_{DO}, n, V_{TH}, m) \) to the physical parameters in (3) and (5). Parameters \( I_{DO} \) and \( V_{DO} \) can be obtained directly from (3) and (5) respectively when \( V_{GS} = V_{DD} \), while for the other parameters the next procedure is used.

In a first approach we assume that \( V'_{DD} \) has a linear dependence with \( V_{GS} \), i.e. \( m=1 \) in (2), and substitute (2) in (3). Neglecting channel length modulation and with some algebra we get:

\[
I_{DS} = I_{DO} \left( \frac{V_{GS}^2}{V_{DD}^2} \right) \left( 1 + \theta V_{DR} \right) \frac{1 + \frac{1}{\theta} \frac{V_{DD}}{V_{VTR}^2}}{1 + \frac{1}{\theta} \frac{V_{DD}}{V_{VTR}^2}} \frac{V_{DD}}{V_{DS}}
\]

where \( V_{DR} = V_{DD} - V_{th} \).

Equating (6) to \( I_{DO}^n \) in (1) (with \( \lambda = 0 \)) we obtain an expression for \( n \):

\[
n = 2 \left( \ln \left( \frac{(1+\theta V_{DR})\frac{V_{DD}}{V_{VTR}^2}}{(1+\theta V_{GR})\frac{V_{DD}}{V_{VTR}^2}} \right) \right)
\]

Eq. (7) depends on the gate voltage complicating the current expression. When \( V_{GR} = V_{VR} \) we have:

\[
n (V_{GS} = V_{DD}) = 1 + \frac{1}{1+\theta V_{DR}} + \frac{1}{1+\frac{1}{\theta} \frac{V_{DD}}{V_{VTR}^2}} V_{DD}
\]

this is the general expression that we use for \( n \).

Now we choose a \( V_{TH} \) value that provides a better fitting of the \( I_{DS} - V_{GS} \), as:
Figure 2 – Delay time of a CMOS buffer vs. input rise time for different aspect ratios (symbols are SPICE simulations and solid lines are model predictions).

\[ V_{TH} = \left( \frac{i_{DQ3}}{I_{DQ3}} \right)^{\frac{1}{m}} \frac{V_{DD}}{V_{GS2}} - V_{DD} \]  

(9)

where \( I_{DQ3} \) is the drain saturation current for \( V_{GS} = V_{DD} / 3 \) and \( V_{DS} = V_{DD} \) that can be obtained from (3). Eq. (9) can be obtained by equating \( I_{DQ3} \) to \( I_{DQ0} \) in (1).

Equations (8) and (9) provide a direct analytical relationship between \( n \) and \( V_{TH} \) with physical parameters.

Parameter \( m \) is obtained following the same procedure to get the velocity saturation index \( n \). Equating (2) and (5) and computing the limit for \( V_{GT} = V_{DT} \) we obtain:

\[ m = \frac{x}{2(1+x-\sqrt{1+x})} \]  

(10)

where \( x \) takes the form:

\[ x = 2 \frac{V_{DT}}{(1+\delta) E_{sat} L_{eff}} \]  

(11)

One of the main advantages of this new formulation of the nth-power law MOSFET model is the incorporation of additional effects to the model that were not included explicitly in the original formulation of Sakurai [1]. This is the case of the temperature, which is not included explicitly in (1) but can be taken into account implicitly from the temperature-dependence of the physical parameters used in this new formulation, as shown in the next section.

Figure 3 – Temperature dependence of propagation delay of a CMOS buffer.

IV. RESULTS

In this section we compare the physically based nth-power law MOSFET model to accurate HSPICE level 50 (MM9 model [7]) simulations for a 0.18 \( \mu \)m CMOS technology.

Fig.1 shows the \( I_D - V_{DS} \) curves of a nMOS transistor with dimensions \( W/L = 6 \mu \text{m}/0.18 \mu \text{m} \). It is observed that the model proposed provides an accurate description of the device characteristics for the whole gate and drain-source voltage ranges. Similar results were obtained for pMOS devices. Table 1 shows the model parameters for the nMOS transistor shown in Fig.1, the nth-power law parameters are directly extracted from technological parameters. In Fig.2 we compare HSPICE simulations of the propagation delay for CMOS inverters with different aspect ratios to the delay model developed in [2] (based on the nth-power law MOSFET model) using the physical formulation proposed in this work. We also apply the physical model of the nth power law to other different delay model for CMOS inverters [5] in Fig.3. Different temperatures are used in the simulations, showing a very good agreement with respect to HSPICE. This figure show the advantage of the physical formulation of the nth-power law model since the delay dependence with a target parameter as the temperature is implicitly included in the physical formulation (and not included in the original nth-power law model). The computation of the gate delay at different supply voltages and temperatures using the original device model would require a previous computation of the original parameters for each \( V_{PD} \) and temperature. With the proposed physical description, these
parameters are included into the model, thus allowing its implementation in CAD tools.

V. Conclusions

A new formulation of the nth-power law MOSFET model has been presented. The model takes into account physical parameters to describe carrier saturation effects in a simple way showing good accuracy for a deep submicron technology (0.18μm). The model can be used in many handy formulae developed from the nth-power law MOSFET model [1]-[5].

VI. Acknowledgment

This work has been supported by the Comisión Interministerial de Ciencia y Tecnología under project CICYT-TIC02-01238 and from CRL, Microprocessor Research Laboratories from Intel Corporation.

REFERENCES


A Compact Propagation Delay Model for Deep-Submicron CMOS Gates including Crosstalk
J. L. Rosselló and J. Segura
Electronics Technology Group, Balearic Islands University, Spain.
emails: {j.rossello,jaume.segura}@uib.es

Abstract—We present a compact, fully physical, analytical model for the propagation delay and the output transition time of deep-submicron CMOS gates. The model accounts for crosstalk effects, short-circuit currents, the input-output coupling capacitance and carrier velocity saturation effects. It is based on the nth-power law MOSFET model and computes the propagation delay from the charge delivered to the gate. Comparison with HSPICE simulations and other previously published models for different submicron technologies show significant improvements in terms of accuracy.

I. INTRODUCTION
Timing analysis is one of the most important topics in VLSI design. The nonlinear behavior of CMOS gates requires numerical procedures for accurate timing analysis that together with today’s circuit complexity results in large computation times. Moreover, the constant scaling of feature sizes and supply voltages with the increase in both operating frequency and signal rise/fall times has made digital designs more vulnerable to noise. In modern ICs, interconnect coupling noise (crosstalk) becomes a performance limiting factor that must be analyzed carefully during the design process. If not considered, crosstalk can cause extra delay, logic hazards and logic malfunction.

Traditionally gate delay and crosstalk have been treated as independent problems and modeled separately. The impact of crosstalk on gate delay for scaled technologies requires a compact gate delay model that incorporates this effect for an optimized analysis.

The propagation delay of CMOS inverters has been extensively studied in the past [1]-[7] as a first step to describe more complex gates [8, 9]. Cocchini et al. [3] obtained a piece-wise expression for the propagation delay based on the BSIM MOSFET model [10]. The model included overshooting effects (due to the input-to-output coupling capacitance) while short-circuit currents were neglected. In [2] and [4] K.O Jeppson and L. Bisdominis presented a model for the output response of CMOS inverters using a quadratic current-voltage dependence for MOSFET devices, which is not longer valid for submicron and deep-submicron devices. Daga et. al. [5] obtained a simple empirical expression for the propagation delay taking into account both overshooting and short-circuit currents. Although the simplicity of their model, CMOS inverter switching characteristics are anad...
and the MOSFET model used is presented. The delay and the output transition time models are developed in Section 3 and compared to HSPICE simulations and other previously published models for different technologies in section 4. Finally in section 5 we present the conclusions.

II. ANALYSIS OF THE CMOS INVERTER SWITCHING CHARACTERISTICS

The dynamic behavior of a CMOS inverter (see Fig.1) is described by:

\[
(C_L + C_M) \frac{dV_{out}}{dt} = I_p - I_n + C_M \frac{dV_{in}}{dt}
\]

(1)

where \(C_L\) is the output capacitance, \(V_{out}\) and \(V_{in}\) are the output and input voltage respectively, while \(I_p\) and \(I_n\) is the current that crosses the pMOS and the nMOS transistor respectively. \(C_M\) is the input to output coupling capacitance which is voltage dependent. The static value of \(C_M\) when the input is low (\(C_{M,\text{off}}\)) is computed considering the side-wall capacitances of both transistor drains and the gate to drain capacitance of the pMOS transistor that operates in the linear region as:

\[
C_{M,\text{off}} = C_{M,\text{off}} \left( \frac{W_{p,\text{eff}}}{2} + L_{D,\text{p}} W_{p,\text{eff}} + L_{D,\text{n}} W_{n,\text{eff}} \right)
\]

(2)

with \(W_{p,\text{eff}}\) and \(W_{n,\text{eff}}\) being the effective channel width of pMOS and nMOS respectively, \(I_{p,\text{eff}}\) is the effective channel length of pMOS, while \(L_{D,\text{n}}\) and \(L_{D,\text{p}}\) are the gate-drain underdiffusion for the nMOS and pMOS transistors respectively. For a static input high the capacitance \(C_{M,\text{on}}\) is obtained similarly. In this work a mean value for the coupling capacitance during the transition (\(C_M = 0.5 (C_{M,\text{off}} + C_{M,\text{on}})\)) is used.

The propagation delay (defined as \(t_{p,\text{off}}\) for a high to low output transition) is typically defined as the time interval from the 50% \(V_{DD}\) input voltage to the 50% \(V_{DD}\) output voltage. The dependence of the propagation delay with design parameters is non-linear and difficult to model given that (1) can not be solved in a closed form even using the simple Shockley MOSFET model. Moreover carrier saturation effects become important with technology scaling and more complex MOSFET models accounting for such effects must be considered. The influence of a low-pass coupled line (crosstalk), would require an additional differential equation coupled to (1), thus complicating the analysis more severely.

The \(n\)th-power law MOSFET model [11] is a widely used short-channel drain current model, and will be used in this work to derive the propagation delay and the output transition time of CMOS inverters. The drain current is expressed as:

\[
I_D = \begin{cases} 
0 & (V_GS \leq V_{TH}) \\
(2 - \frac{V_{GS}}{V_{TH}}) V_{DS}^n L_{p,\text{off}} & (V_{DS} < V_{D0}) \\
V_{DS}^n & (V_{DS} \geq V_{D0})
\end{cases}
\]

with

\[
I_{D0} = I_D \left( \frac{I_{G} - V_{TH}}{V_{DD} - V_{TH}} \right)^n
\]

(4)

where \(V_{GS}, V_{DD}\), and \(V_{D0}\) are the gate, supply, and saturation voltage respectively and \(I_{D0}\) is the drain current at \(V_{GS} = V_{DS} = V_{DD}\). The parameter \(n\) is the velocity saturation index that ranges between 2 (long-channel devices) and 1 (short-channel) [11]. The saturation voltage \(V'_{D0}\) is given by:

\[
V'_{D0} = V_{D0} \left( \frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}} \right)^m
\]

(5)

The parameter \(V_{D0}\) is the saturation voltage at \(V_{GS} = V_{DD}\), while \(m\) and \(V_{TH}\) are empirical parameters in [11]. These equations are mathematically simpler than physically-based MOSFET models such as BSIM3v3 or MM9 with the disadvantage that, in the original model developed by Sakurai and Newton, the relationship between the empirical and the process parameters supplied by manufacturers is not provided. The variation of the \(n\)th-power law model predictions with key parameters like the supply voltage or temperature is not taken into account in the original formulation performed by Sakurai and Newton, where each parameter must be recomputed if the supply voltage, the temperature of operation, or some device dimension are changed. In this work we use the physical formulation for the \(n\)th-power law developed previously [15]. Using this formulation, the empirical parameters used by Sakurai et al can be related to physical parameters like the supply voltage or temperature.

III. PROPAGATION DELAY MODEL FOR CMOS INVERTERS

We compute the propagation delay when the input voltage rises (for a falling transition the analysis is equivalent). The short-circuit, overshoot, and crosstalk effects are first neglected and incorporated later. Assuming a linear variation of the input voltage with rise time \(t_n\) and using the \(n\)th-power law model the analytical solution for the output voltage valid while the nMOS is saturated (\(V_{out} > V'_{D0}\)) is:

\[
V_{out} = V_{DD} - \frac{I_{D0}}{C_L} \left( \frac{t - t_n}{t_{in} - t_n} \right)^{n+1} \frac{t_{in} - t_n}{n+1}
\]

(6)

where parameters \(I_{D0}\) and \(t_{in}\) are the maximum saturation current and the velocity saturation index of \(n\) respectively. Parameter \(t_n\) is the time at which...
nMOS starts to conduct. From (6) we obtain the time at which the output voltage is \( V_{DD}/2 \). Then the propagation delay will be this time minus the time at which the input is also at \( V_{DD}/2 \) (i.e. \( t_{in}/2 \)):

\[
    t_{pHL_1} = t_{in} + \left[ \frac{Q_f}{I_{DO_o}} \left( n + 1 \right) \right]^{\frac{n}{n+1}} \frac{t_{in}}{n} - \frac{t_{in}}{2}
\]

(7)

where \( Q_f = C_t V_{DD}/2 \) is the charge transferred by the nMOS transistor when the output reaches \( V_{DD}/2 \). Equation (7) is valid for slow inputs (defined when \( t_{pHL} \leq t_{in}/2 \)) since for fast inputs the nMOS transistor is not saturated when the output reaches \( V_{DD}/2 \). If \( Q_{Io} \) is defined as the total charge transferred through the nMOS transistor when \( t_{pHL} = t_{in}/2 \) (the limit between fast and slow input transitions), then (7) is valid when \( Q_f \leq Q_{Io} \). The value of \( Q_{Io} \) is obtained equating \( t_{pHL_1} = t_{in}/2 \) and solving for \( Q_{Io} \):

\[
    Q_{Io} = \frac{I_{DO_o}}{(n+1)} t_{in} \left( 1 - \frac{V_{TN}}{V_{DD}} \right)
\]

(8)

where \( V_{TN} \) is the threshold voltage of nMOS. For fast input transitions, the input reaches the supply voltage before the output is at \( V_{DD}/2 \) and then \( Q_f > Q_{Io} \). For this case (7) is not valid and the propagation delay is obtained by solving (1) for \( I_o = I_{DO_o} \) (also neglecting the short-circuit and the overshooting currents as a first analysis) leading to:

\[
    t_{pHL_1} = \frac{t_{in}}{2} + \frac{Q_f - Q_{Io}}{I_{DO_o}}
\]

(9)

Equation (9) is valid for the interval \( Q_f > Q_{Io} \) (fast input range). For simplicity, the proposed model for the propagation delay (eqs. (7) and (9)) does not take into account the fact that the nMOS transistor enters in its linear region in the interval \( V_{DD}/2 < V_{out} < V_{DD} \) for those cases in which \( V_{DD} > V_{DD}/2 \).

The output fall time (\( t_f \)), defined as the 70% of the output voltage slope at \( V_{DD}/2 \) [8], is obtained in a similar way as (7) and (9). The output voltage slope takes the form:

\[
    \frac{dV_{out}}{dt} = \left\{ \begin{array}{ll}
        \frac{Q_f}{I_{DO_o} V_{DD}} \left( n + 1 \right) \left( n_{m+1} \right)^{\frac{n}{n+1}} & \text{if } Q_f < Q_{Io} \\
        \frac{Q_f}{I_{DO_o} V_{DD}} \left( n_{m+1} \right) \left( n_{m+1} \right)^{\frac{n}{n+1}} & \text{if } Q_f > Q_{Io} 
        \end{array} \right.
\]

(10)

For the evaluation of the output fall time we use:

\[
    t_f = \frac{V_{DD}}{dV_{out}/dt} |_{V_{DD}/2}
\]

(11)

The value of the output fall/rise time is needed for the evaluation of the input fall/rise time of the gates driven by the inverter.

A. Including short-circuit currents

The model developed cannot be used for static CMOS gates, since the short-circuit current was not considered \( I_p \) was neglected when solving (1)). In general, the impact of the short-circuit current on the gate delay can be considerable, specially for large values of the input time. In Fig.2 we plot the propagation delay \( t_{pHL} \) vs. the input time \( t_{in} \) for different values of the \( W_o/W_p \) ratio for a 0.35um technology. HSPICE simulations (dots) are compared to a previous model [3]. Short-circuit currents were not taken into account in [3], leading to severe underestimations of the propagation delay for large values of \( t_{in} \). The difference between simulations and the model developed in [3] can be associated to the delay increment due to short-circuit currents.

In this work, the short-circuit current contribution to the gate delay is modeled as an additional charge transferred through the pull-down (pull-up) transistor during an output falling (rising) transition. We develop an analytical expression for the short-circuit charge transferred when the output capacitance is negligible (the case where the impact of short-circuit current on the propagation delay is higher). In this case, the circuit behavior is close to the inverter DC operation since \( I_p \approx I_n \). At the beginning of the transition, the pMOS transistor drives a current equal to the nMOS saturation current, while at the end of the transition the pMOS is saturated. The time point at which the maximum current takes place is such that \( I_{DO_o} = I_{DO_n} \). We compute the short-circuit charge transferred until this switching point. Given that \( V_{GS} = V_{in} \) for the nMOS and \( V_{GS} = V_{DD} - V_{in} \) for the pMOS, the time at which the short-circuit is maximum (defined as \( t_{SC_{max}} \)) is obtained solving:

\[
    I_{DO_o} \left( \frac{V_{DD} - V_{in} - V_{TR_o}}{V_{DD} - V_{TR_o}} \right)^{n_o} = I_{DO_n} \left( \frac{V_{DD} - V_{in} - V_{TR_n}}{V_{DD} - V_{TR_n}} \right)^{n_p}
\]

(12)

where \( I_{DO_o} \) and \( I_{DO_n} \) are the parameters \( I_{DO} \) of the nMOS and the pMOS transistor respectively, and \( V_{TR_o} \) is the threshold voltage of the pMOS. We obtained a analytical approximation to the solution of (12) as...
\[ t_{\text{max}}^{CL=0} = t_n + t_{\text{in}} \left( 1 - \frac{V_{T_N}}{V_{DD}} - \frac{|V_{TP}|}{V_{SC}} \right) \frac{n_p}{1 + F_p^{n_p+n_n}} \]  

(13)

where \( F_p \) is given by:

\[ F_p = \frac{I_{D\text{on}}}{I_{D\text{off}}} \left( \frac{V_{SC}}{V_{DD} - V_{TN}} \right)^{n_p} \left( \frac{V_{DD} - |V_{TP}|}{V_{SC}} \right)^{n_n} \]  

(14)

where \( V_{SC} = V_{DD} - V_{TN} - |V_{TP}| \). The short-circuit charge transferred until \( V_{\text{out}} = V_{DD}/2 \) is computed integrating the nMOS current from the time point at which the nMOS starts to conduct \( (t_n) \) until the time point at which the short-circuit is maximum \( (t_{\text{max}}^{CL=0}) \). From this integration we have:

\[ q_t^{\text{LC}} = \frac{I_{D\text{on}} t_n}{1 + n_n} \left( 1 - \frac{V_{T_N}}{V_{DD} - V_{TN}} \frac{|V_{TP}|}{V_{SC}} \right)^{n_n+1} \left( \frac{V_{DD} - |V_{TP}|}{V_{DD} - V_{TN}} \right)^{n_n} \]  

(15)

Finally, the value of the charge transferred through the nMOS during the falling output transition used in the delay model is \( I_f = C_L V_{DD}/2 + q_t^{\text{LC}} \).

Note that \( q_t^{\text{LC}} < Q_{\text{f}} \). For large values of the output capacitance \( (C_L) \) such that the transitions is in the fast input range \( (Q_I > Q_{\text{f}}) \), eq. (15) is not valid. The additional charge in the fast input range due to (15) in \( Q_I \) is negligible with respect \( C_L V_{DD}/2 \). For this reason we decide to use (15) for all cases since when (15) is not valid the impact on the delay due to the inclusion of \( q_t^{\text{LC}} \) in \( Q_I \) is negligible and the total delay is practically not affected.

B. Modeling CMOS Gates

Complex CMOS gates are modeled through a gate collapsing technique. Each stack of transistors is collapsed into a single equivalent transistor. The maximum current of this single equivalent transistor would be the maximum current of the stack. For the case of a stack of nMOS transistors we define \( J_{D\text{on}}^{(1,N_n)} \) as the maximum current that can be driven by the stack (modeled as a chain of \( N_n \) series-connected transistors) that is obtained from the transistor collapsing technique developed in [9]. For the simple case of a chain with \( N_n \) identical transistors, the collapsing technique provides a simple expression for \( J_{D\text{on}}^{(1,N_n)} \):

\[ J_{D\text{on}}^{(1,N_n)} = \frac{I_{D\text{on}}}{1 + (N_n - 1) K_n} \]  

(16)

where \( I_{D\text{on}} \) is the maximum saturation current of each nMOS (drain current when \( V_{GS} = V_{DS} = V_{DD} \)). The parameter \( K_n \) is a technology-dependent parameter given by:

\[ K_n = \frac{3V_{DD} n_n (1 + \gamma_n)}{5(V_{DD} - V_{TN})} \]  

(17)

C. Including overshooting and crosstalk

Similarly to short-circuit current, the impact of the overshooting current on the delay is included as an additional charge to be transferred through the nMOS transistor. For fast inputs, the charge injected through the coupling capacitance \( (C_M) \) when \( V_{\text{out}} = V_{DD}/2 \) is \( Q_o = C_M V_{DD} \). For simplicity we assume the same value for \( Q_o \) in the slow-input case. Therefore, the total charge to be transferred through the nMOS transistor during a falling output transition is computed as:

\[ Q_I = 0.5[(C_L + 2C_M) V_{DD}^2] + q_t^{\text{LC}} \]  

(18)

The crosstalk induced delay is also considered as an additional charge to be transferred through the charging (discharging) transistor. Crosstalk affects delay when two lines (the aggressor and the victim) switch simultaneously. For a victim node falling transition, the gate delay can be reduced (this lower bound is defined as \( t_{\text{pH,L,v}} \) or increased (defined as \( t_{\text{pH,L,v}} \)) depending on if the aggressor makes a falling or a rising transition respectively. Since large circuits can handle hundreds of millions of interconnect lines on a single chip, simple and accurate analytical descriptions for crosstalk delay are of high importance for a fast timing verification of the whole chip. Consider the circuit in Fig.3 where the aggressor and the victim gates drive an output capacitance \( C_{L,a} \) and \( C_{L,v} \) respectively, and there is a coupling capacitance \( C_c \) between \( V_{out,a} \) and \( V_{out,v} \) (that causes the crosstalk between the two lines). The total capacitance at the output of the victim gate that must be charged and discharged is the addition of \( C_{L,v} \) and \( C_c \). We estimate the crosstalk delay bounds considering that the whole aggressor switching happens within the switch interval of the victim line. For the case of opposite (identical) transitions in both nets, a charge \( C_c V_{DD} (-C_c V_{DD}) \) is injected through the coupling capacitance from the aggressor to the victim line. We add these additional charges to the charge-based delay model developed to obtain the crosstalk delay bounds.

\[ Q_I = 0.5[(C_L + C_c + 2C_M) V_{DD}] + q_t^{\text{LC}} \]  

(18)
Equation (19) must be used in eqs. (7), (9) and (11) to estimate the propagation delay and the transition time of the gate.

IV. RESULTS

We plotted model results vs. HSPICE simulations for different technologies, and provide the propagation delay for different values of the input transition time $t_{in}$, the configuration ratio $W_p/W_n$, and the supply voltage $V_{DD}$.

In Fig.2 we plot the crosstalk-free propagation delay $t_{p,n}$ vs. the input time $t_{in}$, and for different values of the $W_n/W_p$ ratio for a 0.35um technology. HSPICE simulations (dots) are compared to the model proposed obtaining a good accuracy.

In Fig.5 we show the propagation delay variation with temperature for three different supply voltage values. Different trends are observed for each supply voltage value when increasing temperature. The propagation delay increases for $V_{DD} = 1.8V$ while for $V_{DD} = 0.9V$ decreases. This effect is described properly by the proposed model since we are using the physically-based nth-power law MOSFET model developed in [15] that relates the nth-power law parameters to physical parameters (that are temperature-dependent) as the carrier mobility. For a more detailed description of this model the reader is referred to [15].

Fig.6 is a plot of HSPICE simulations (dots) and model predictions of the propagation delay vs. the supply voltage (crosstalk is not considered). The model proposed in this work provides a better fitting than the models in [6, 7]. The model in [7] present discontinuities due to the use of different expressions for the propagation delay.

In Fig.7 we plot HSPICE simulation of the output transition time $t_f$ for different values of the input rise time $t_{in}$ and the configuration ratio $W_p/W_n$. A maximum relative error of 15% is obtained between model predictions and HSPICE simulations. In general a good agreement is obtained with the proposed model.

In Fig.8 we plot the propagation delay limit bounds for a 3-NAND gate for different values of the coupling to load capacitance ratio $C_L/C_{G,n}$. Fig.8 shows an excellent agreement between HSPICE simulations (dots) and the proposed crosstalk delay model (lines).

V. CONCLUSIONS

We have developed an accurate analytical expression to compute the propagation delay and the output delay.
sition time of CMOS gates including crosstalk effects. The main effects present in current submicron CMOS technologies like the input-output coupling capacitance, carriers velocity saturation effects and short-circuit currents are also taken into account in the analysis. The model is compared to HSPICE simulations and other previously published works for different submicron technologies reporting very high accuracy. The model presents a simple way to compute the crosstalk induced delay bounds showing an excellent accuracy with respect HSPICE simulations. The fully analytical description of the delay allows a quick estimation of this magnitude since no numerical analysis is required. This provides a model that can be incorporated in CAD tools and provide a quick estimation of the delay. The model is also useful to investigate the impact of design parameters on the delay.

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish Ministry of Science and Technology, the Regional European Development Funds (FEDER) from the EU project TIC2002-01238, and an Intel Laboratories-CRL research grant.

REFERENCES


