AN ANALYTICAL CURRENT, DELAY, AND POWER MODEL FOR THE SUBMICRON CMOS INVERTER

Anas A. Hamoui and Nicholas C. Rumin

Department of Electrical and Computer Engineering, McGill University
3480 University Street, Montreal, Quebec, Canada H3A 2A7
E-mail: hamoui@macs.ece.mcgill.ca

ABSTRACT

We present an analytical model for computing the supply current, delay, and power in a submicron CMOS inverter using a modified version of the 'n-th power law' MOSFET model [2]. By first computing definable reference points on the output voltage waveform and then using linear approximations through these points to find the actual points of interest, the desired speed and accuracy of the inverter model are achieved. The most important part of the analysis is a three-step process for computing the time and output voltage when the 'short-circuit' transistor changes its mode of operation. The model has been validated using an accurate, physically-based, submicron MOSFET model for a wide range of inverter sizes, input transition times, and capacitive loads: it can predict the delay, peak supply current, and power dissipation to within a few percent of simulation results, while offering about two orders of magnitude gain in CPU time.

1. INTRODUCTION

A number of methods for computing the delay and/or power dissipation in CMOS inverters has been recently presented [5]-[10]. The emphasis on modeling the inverter stems from the fact that a growing fraction of the power consumed by VLSI integrated circuits is due to the clock distribution network, I/O drivers, and busses, which are all based on inverters. Furthermore, a number of efficient transistor-level techniques for reducing CMOS logic gates to equivalent inverters are available [1]. For reliability design, the peak supply current values are needed to properly size the power and signal lines in order to avoid electromigration failures and voltage drop problems.

In this paper, an analytical model for computing the supply current, delay, and power of a submicron CMOS inverter is presented. The effect of the Miller capacitance is included. A modified version of the n-th power law MOSFET model [2] is proposed and used to relate the terminal voltages to the drain current in submicron transistors.

The outstanding feature of the inverter model proposed in this paper is its comprehensiveness: it computes the maximum currents in addition to both the delay and power, and the same model is used regardless of whether the input-voltage switching transition is fast or slow.

2. SUBMICRON MOSFET MODEL

The n-th power law MOSFET model, proposed by Sakurai and Newton [2], offers a simple, yet reasonably accurate, empirical model for the MOSFET drain current. However, it neglects the threshold-voltage variations due to the short-length, narrow-width, and drain-induced barrier lowering (DIBL) effects, which are significant in submicron MOS technologies. To model these variations, the threshold voltage at zero body-bias can be expressed as a linear function of the effective channel length-to-width ratio [3]. Thus, we use a modified n-th power law MOSFET model based on the model in [2] and augmented [3] with the following equation:

$$V_t = V_{tw}\left(1 + f\,\frac{L_e}{W_e}\right) + \gamma\, V_{SB} \qquad (1)$$

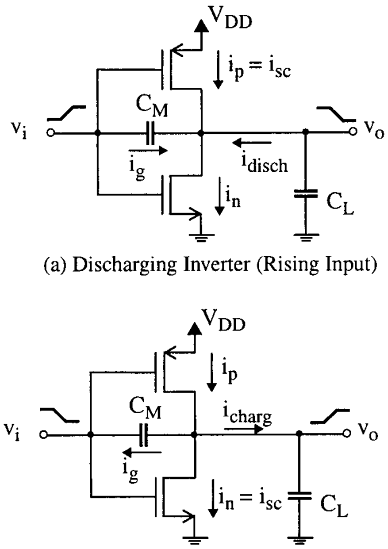
V_t denotes a threshold voltage and V_{tw} is the corresponding zero body-bias threshold voltage for wide-channel transistors. The empirical factor f describes the dependence of V_t on the effective channel length and width (L_e and W_e), while γ models the body effect (V_{SB} is the source-bulk voltage).

For a given feature size, a simple one-time procedure is used to optimize the set of seven parameters in the modified n-th power law model to yield a best fit to the measured I_D-V_{DS} characteristics over the range of channel widths used in circuit design [3]. Note that, although V_t was assumed independent of V_{DS} in (1), the DIBL effect is still implicitly accounted for to some extent by extracting a V_{tw} value optimized over the full range of operating V_{DS} voltages.

3. THE CMOS INVERTER

Consider the CMOS inverter circuit in Fig. 1. The effective load, C_L, includes the NMOS and PMOS drain-bulk junction capacitances, the gate-to-source capacitances of the fanout gates, and the interconnect capacitances. The Miller capacitance, C_M, consists of the NMOS and PMOS gate-to-drain capacitances. The nonlinear voltage-dependent MOSFET parasitic

0-7803-5682-9/99/$10.00©1999 IEEE.



(a) Discharging Inverter (Rising Input)

(b) Charging Inverter (Falling Input)

Fig. 1: CMOS inverter circuit.

capacitances are replaced by equivalent constant capacitances. Over each MOSFET mode of operation, the intrinsic gate capacitance is assumed to be a constant fraction of the effective gate-oxide capacitance.

For the discharging inverter, the input voltage waveform is assumed to be a rising ramp with transition time T_r:

$$v_i(t) = \begin{cases} s_r t & 0 \le t \le T_r \\ V_{DD} & t > T_r \end{cases} \qquad (2)$$

where $s_r \equiv V_{DD}/T_r$ is the slope of the rising input voltage ramp. The differential equation describing the discharging of the CMOS inverter ($0 \le t \le T_r$) is then

$$\frac{dv_o}{dt} = -\frac{i_n - i_p}{C_L + C_M} + \frac{C_M}{C_L + C_M} s_r. \qquad (3)$$

In the following analysis, the current, delay, and power are derived for the discharging inverter case. The analysis for the charging inverter case is symmetrical.

4. POWER DISSIPATION

For the CMOS inverter circuit in Fig. 1, the dynamic energy dissipation per switching event is

$$E = C_L V_{DD}^2 + 2\, C_M V_{DD}^2 + E_{sc}. \qquad (4)$$

The main challenge in computing the power dissipation is the determination of the short-circuit energy dissipation, E_{sc}, due to the direct-path current from supply to ground when both the NMOS and PMOS devices are on. As shown in Table 1, E_{sc} can account for more than 35% of E. Indeed, this percentage is increasing as the power supply voltage and the minimum IC feature size continue to be scaled down.

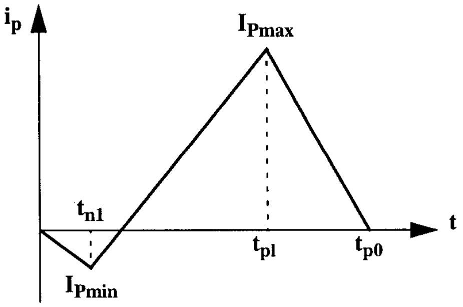


Fig. 2: Piecewise linear approximation of the short-circuit current, i_p, in the discharging inverter.

For the purpose of evaluating E_{sc}, the short-circuit current (i_n in the charging inverter and i_p in the discharging inverter) can be approximated by a piecewise linear function of time (Fig. 2). Thus, the component of E_{sc} due to the discharging of the inverter can be expressed as

$$E_{sc} = \frac{V_{DD}}{2} [\, I_{Pmax} (t_{p0} - t_{n1}) + I_{Pmin}\, t_{pl} \,] \quad (5)$$

where $t_{n1} \equiv V_{ten}/s_r$ and $t_{p0} \equiv (V_{DD} - |V_{tep}|)/s_r$ are the times when the NMOS and PMOS devices turn, respectively, on and off. Here, V_{ten} and $|V_{tep}|$ are the effective NMOS and PMOS threshold voltages, respectively, and are extracted from the I_D-$|V_{GS}|$ characteristics at $|V_{DS}| = 0.01 V_{DD}$ [3].

Note that, as discussed in [3], the empirical parameter V_t in (1) is significantly larger than the threshold voltage as it is normally defined (i.e. the $|V_{GS}|$ needed to induce a strongly inverted channel under the gate).

Hence, the objective is to determine I_{Pmin} and I_{Pmax} as well as their times of occurrence. The determination of the former is straightforward and is discussed in [3].

Maximum Short-Circuit Current

Let t_{pl} be the time when the PMOS transistor leaves the linear region and enters saturation. Simulation results have shown that, for the purpose of computing E_{sc}, it is valid to assume that the short-circuit current reaches it maximum value of I_{Pmax} at $t = t_{pl}$. The special case of $t_{pl} = t_{p0}$ corresponds to very fast input ramps where the PMOS device turns off before entering saturation. This occurs if v_i reaches $V_{DD} - |V_{tp}|$ (switching the PMOS transistor off) before the output voltage waveform has completed its overshoot and v_o has dropped below $V_{DD} - |V_{tp}|$.

Let t_{ns} be the time when the NMOS device leaves saturation and enters the linear region. Since at $v_i = v_o$

both the NMOS and PMOS transistors must be in saturation, the PMOS device must enter saturation before the NMOS device leaves it. Therefore, we have $t_{pl} \leq t_{nm1}$, where $t_{nm1} \equiv min(t_{ns}, t_{p0})$.

During the time interval $t_{n1} \leq t \leq t_{nm1}$, the PMOS transistor operates in its linear region until time t_{pl}, when it saturates. The NMOS device, on the other hand, remains saturated over the entire time interval. A three-step approach is used to evaluate t_{pl}: First, the short-circuit current i_p is neglected and an approximation to t_{pl} is computed. Second, this approximate time is corrected for the short-circuit current (neglected in the first step), yielding a point on the inverter's switching trajectory close to t_{pl}. Finally, the tangent to the output voltage waveform at this point is used to compute t_{pl} and $v_o(t_{pl})$.

Since Step1 is relatively straightforward, it is omitted here in the interest of brevity.

. Step 2:

Since the short-circuit current i_p was neglected in Step 1, the computed values of t_{pl} and $v_o(t_{pl})$, denoted $\tilde{t}_q$ and V_{Oq}, respectively, are only approximations to the true values. The effective current available to discharge the load is actually only $i_n - i_p$ because the PMOS transistor is on during the time interval $t_{n1} \leq t \leq t_{p0}$. Hence, for the output voltage to actually drop to V_{Oq}, the output node must be discharged by

$$Q_q = \int_{t_{n1}}^{\tilde{t}_q} i_n \, dt = \int_{t_{n1}}^{t_q} (i_n - i_p) \, dt \qquad (6)$$

where t_q is the actual time required for the output voltage to drop to V_{Oq}. Hence, defining

$$Q_{sc} \equiv \int_{t_{n1}}^{\tilde{t}_q} i_p \, dt \quad \text{and} \quad Q_{add} \equiv \int_{\tilde{t}_q}^{t_q} (i_n - i_p) \, dt, \qquad (7)$$

it follows from (6) that $Q_{add} = Q_{sc}$.

Here, Q_{sc} represents the amount of charge which leaked from the power supply through the short-circuit PMOS transistor during the time interval $t_{n1} \leq t \leq \tilde{t}_q$. To compensate for Q_{sc}, the output node must be discharged, during the time interval $\tilde{t}_q \leq t \leq t_q$, by a net additional charge Q_{add} to allow the output voltage to actually drop to V_{Oq}.

To compute Q_{sc} and Q_{add}, the drain currents $i_n(t)$ and $i_p(t)$ are represented by piecewise linear functions of time, as shown in Fig. 3. The current values at $t = \tilde{t}_q$ are calculated from the NMOS and PMOS drain current equations based on their respective terminal voltages, with the approximation $v_o(\tilde{t}_q) \approx V_{Oq}$. For $t \geq \tilde{t}_q$, the drain currents are described by linear functions of time with rates of change equal to those at $t = \tilde{t}_q$. Thus, equating Q_{sc} and Q_{add} yields:

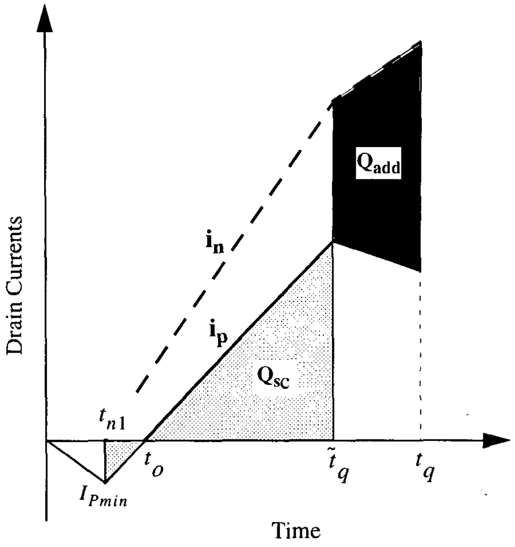


Fig. 3: Piecewise linear approximations of the discharging current i_n (---) and the short-circuit current i_p (—), used in Step 2 of the derivation of the maximum short-circuit current, I_{Pmax}, for the discharging inverter.

$$t_q = \tilde{t}_q + \frac{-\Delta I + \sqrt{\Delta I^2 + 2\, Q_{sc}\, \Delta di}}{\Delta di} \qquad (8)$$

with $\Delta I \equiv i_n(\tilde{t}_q) - i_p(\tilde{t}_q)$ and $\Delta di \equiv \frac{d}{dt}(i_n - i_p)\Big|_{t = \tilde{t}_q}$.

Note that, from Fig. 3:

$$Q_{sc} = \frac{1}{2} I_{Pmin} (t_o - t_{n1}) + \frac{1}{2} i_p(\tilde{t}_q) (\tilde{t}_q - t_o) \,. \quad (9)$$

where $t_o \equiv \dfrac{i_p(\tilde{t}_q)\, t_{n1} - I_{Pmin}\, \tilde{t}_q}{i_p(\tilde{t}_q) - I_{Pmin}}$.

. Step 3:

Now, $[t_q, V_{Oq}]$ represents an actual point on the output voltage waveform very close to the desired point $[t_{pl}, v_o(t_{pl})]$. Therefore, one can approximate the output voltage waveform near t_{pl} by a linear function of time through $[t_q, V_{Oq}]$. This linear approximation yields an improved value of t_{pl}, which takes into account the short-circuit current. The corresponding output voltage, $v_o(t_{pl})$, can now be determined, and the maximum value, I_{Pmax}, of the short-circuit current is computed with $V_{SGp} = V_{DD} - s_r t_{pl}$ and $V_{SDp} = V_{DD} - v_o(t_{pl})$.

5. PROPAGATION DELAY AND MAXIMUM DISCHARGING CURRENT

For delay time calculations, the output voltage transition can be properly characterized by the tangent

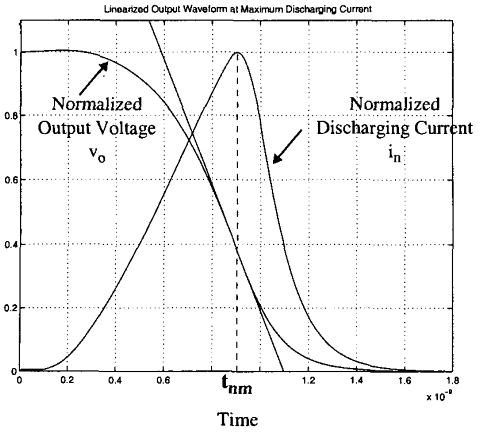


Fig. 4: Linear approximation of the output voltage waveform for a discharging inverter. The output voltage transition can be fully characterized by the tangent line to the output voltage waveform at time t_{nm} when the discharging current i_n reaches its maximum.

line to the output voltage waveform at the time when the charging/discharging current reaches it maximum (Fig. 4). The derivation of the delay time using this approach is straightforward and is discussed in [3].

However, to evaluate the delay, the time t_{nm} and output voltage $v_o(t_{nm})$ when the discharging current i_n reaches its maximum value of I_{Nmax} must be first computed, as described below.

Maximum Discharging Current

The discharging current i_n reaches its maximum when the NMOS transistor leaves saturation and enters the linear region (at $t = t_{ns}$), but not later than the time when V_{GSn} attains its maximum value of V_{DD} (at $t = T_r$). Defining $t_{nm1} \equiv min(t_{ns}, t_{p0})$ and $t_{nm2} \equiv min(t_{ns}, T_r)$, it follows that time t_{nm} must occur within one of the following two intervals:

- **Time Interval 1:** $t_{pl} \le t \le t_{nm1}$

Both the PMOS and NMOS devices are saturated. For $t \le t_{nm}$, $V_{DSn} = v_o(t)$ is larger than $V_{SDp} = V_{DD} - V_{DSn}$, because $v_o(t)$ is a falling signal and $v_o(t_{nm})$ is close to $V_{DD}/2$. Thus, a possible simplifying assumption, to be used in the drain current equations for the NMOS and PMOS devices, is: $1 + \lambda_n V_{DSn} \approx 1 + \lambda_n V_{DD}$ and $1 + \lambda_p V_{SDp} \approx 1$.

An expression for the output voltage waveform during the time interval $t_{pl} \le t \le t_{nm1}$ can be obtained by solving equation (3), with i_n and i_p expressed in terms of their respective terminal voltages using the modified n-th power law equations and with initial condition $v_o(t_{pl})$ (computed in Section 4). This, combined with the value of the output voltage when the NMOS device changes its mode of operation

$$v_o(t_{ns}) = V_{DSn_{sat}} = K_n(s_r t_{ns} - V_{tn})^{m_n}, \quad (10)$$

yields t_{ns}. If $t_{ns} \le t_{p0}$, then $t_{nm} = t_{ns}$. Otherwise, time interval 2 must be used to compute t_{ns}.

- **Time interval 2:** $t_{p0} \le t \le t_{nm2}$

The PMOS device is off ($i_p = 0$), while the NMOS device is saturated. Steps similar to those above yield the output voltage waveform during the time interval $t_{p0} \le t \le t_{nm2}$ and the time t_{ns}. If $t_{ns} \le T_r$, then $t_{nm} = t_{ns}$. Otherwise, $t_{nm} = T_r$.

The maximum value, I_{Nmax}, of the discharging current is finally computed with $V_{GSn} = s_r t_{nm}$ and $V_{DSn} = v_o(t_{nm})$.

6. RESULTS AND CONCLUSION

The proposed model, implemented in MATLAB, has been tested with a wide range of inverters designed in a 5V 0.8µm BiCMOS technology. Various switching conditions of input transition time and capacitive load were considered. To validate the model, the delay, peak supply current, and power dissipation are compared with the 'exact' values obtained by simulating the analyzed circuits in the ELDO simulator using Nortel's MISNAN MOSFET model [11]. Some typical results are presented in Fig.6, Fig.7, and Table 1. Note that in all cases the error is less than 8%.

CPU Time

To determine a reasonable estimate of the speedup achieved by the proposed inverter model over ELDO simulation, the time step in ELDO was set to the minimum needed to capture the delay and the time of the peak supply current to the nearest 0.01 nsec, and the stop time was selected to correspond to the shortest duration of the transient analysis which yields the supply energy dissipation to within 5%. Results show that the inverter model, run in MATLAB, offers about two orders of magnitude improvement in CPU time over ELDO.

7. REFERENCES

[1] A. Nabavi-Lishi and N. C. Rumin, "Inverter models of CMOS gates for supply current and delay evaluation," *IEEE Trans. Computer-Aided Design*, vol.13, pp. 1271-1279, Oct. 1994.

[2] T. Sakurai and A. R. Newton, "A simple MOSFET model for circuit analysis," *IEEE Trans. Electron Devices*, vol. 38, pp. 887-894, Apr. 1991.

[3] A. Hamoui, "Current, delay, and power analysis of submicron CMOS circuits," M.Eng. Thesis, McGill University, Montreal, 1998.

[4] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, pp. 584-594, Apr. 1990.

[5] S. H. K. Embabi and R. Damodaran, "Delay models for CMOS, BiCMOS and BiNMOS circuits and their applications for timing simulations," *IEEE T. Computer-Aided Design*, vol. 13, pp. 1132-1142, Sep. 1994.

[6] H-J Park and M. Soma, "Analytical model for switching transitions of submicron CMOS logics," *IEEE J. Solid-State Circuits*, vol. 32, pp. 880-889, Jun. 1997.

[7] L. Bisdounis, S. Nikolaidis, and O. Koufopavlou, "Analytical transient response and propagation delay evaluation of the CMOS inverter for short-channel devices," *IEEE J. Solid-State Circuits*, vol. 33, pp. 302-306, Feb. 1998.

[8] S. R. Vemuru and N. Scheinberg, "Short-circuit power dissipation estimation for CMOS logic gates," *IEEE Trans. Circuits Syst. I*, vol. 41, pp. 762-765, Nov. 1994.

[9] A. Hirata, H. Onodera, and K. Tamaru, "Estimation of short-circuit power dissipation for static CMOS gates," *IEICE Trans. Fund.*, vol. E79-A, pp. 304-311, Mar. 1996.

[10]S. Turgis and D. Auvergne, "A novel macromodel for power estimation in CMOS structures," *IEEE Trans. Comp.-Aided Design*, vol. 17, pp. 1090-1098, Nov. 1998.

[11]A. R. Boothroyd, S. W. Tarasewicz, and C. Slaby, "MISNAN-a physically-based continuous MOSFET model for CAD applications," *IEEE Trans. Comp.-Aided Design*, vol.10, pp. 1512-1529, Dec. 1991.

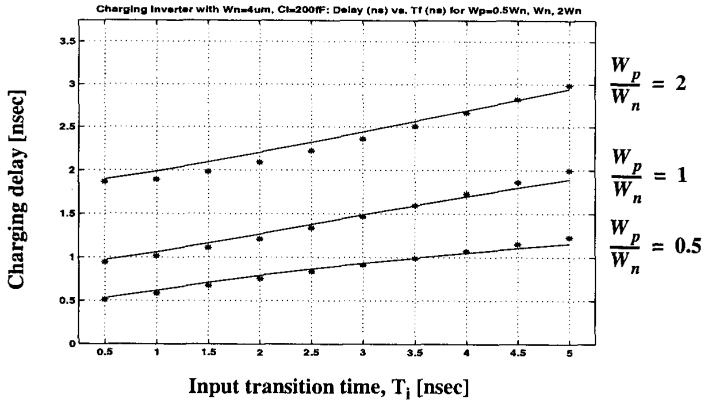


Fig. 6: Delay plot versus input transition time for several ratios of transistor sizes, W_p/W_n. ($W_n = 4\mu m$, $C_L = 0.2pF$).

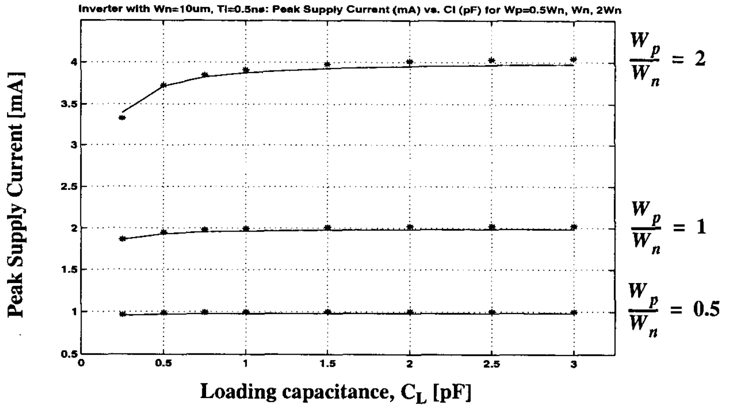


Fig. 7: Plot of the peak power-supply current versus loading capacitance CL for several ratios of transistor sizes, W_p/W_n. ($W_n = 10\mu m$, $T_i = 0.5nsec$).

T_i	W_n	W_p	C_L	**Short-Circuit Energy E_{sc} (pJ)**		**Total Dynamic Energy E (pJ)**		$\% \frac{E_{sc}}{E}$	**% Error in E**
(ns)	(μm)	(μm)	(fF)	Analytical Model	ELDO Simulator	Analytical Model	ELDO Simulator		
2	6	6	250	0.837	0.900	8.07	7.88	11.4	2.4
		12		1.50	1.55	9.22	8.87	17.4	3.9
	12	12	200	1.94	2.65	8.878	9.079	29.1	2.2
		36	350	3.26	3.63	14.92	14.49	25.0	2.9
4	6	12	250	3.42	4.52	11.2	11.80	38.3	5.0