Deep Submicron Switching Current Modeling for CMOS Logic Output Transition Time Determination

Philippe Maurine, Nadine Azemard, Daniel Auvergne

DEEP SUBMICRON SWITCHING CURRENT MODELING FOR CMOS LOGIC OUTPUT TRANSITION TIME DETERMINATION

P. Maurine, N. Azémard, D. Auvergne

LIRMM, UMR CNRS/Université de Montpellier II, (C5506),
161 rue Ada, 34392 Montpellier, France
pmaurine, azemard, auvergne@lirmm.fr

Abstract. Non zero signal rise and fall times contribute significantly to CMOS gate performances such as propagation delay or short circuit power dissipation. We present a closed form expression to model output rise and fall times in deep submicron CMOS structures. The model is first developed for inverters considering fast and slow input ramp conditions. It is then extended to gates through a reduction procedure considering the maximum current available in the serial transistor array. Validation of this modeling is obtained by comparing calculated gate output transition time to simulated ones (HSPICE level and foundry card model on 0.18µm process).

1. INTRODUCTION

The use of safe gate level characterization of performances over the full design space is the only way to maintain timing relationships between functional blocks when designs approach complexity of millions of transistors. To control or drive design alternatives, technology migration, as well as process variation it appears necessary to get available design oriented models to evaluate the performances of specific structures. The traditional representation of delay associates a constant "inertial" delay characteristic of the cell to an output load dependent delay characterizing the cell size and structure.

However input-to-output coupling effects associated to speed saturation of the carriers induce non linearity for the propagation delays which are important enough to be considered for accurate cell delay-performance characterization. Great sensitivity of the delay to the edge of the input controlling signal has been observed in submicron processes. These edges are generally defined as the controlling gate output-voltage transition time measured between appropriate voltage levels. These signal rise and fall times contribute significantly to the delay and are responsible of the nonlinear variation of real delay values. As a result, gate delay characterization implies consideration of propagation and output transition times.

The modeling of the gate output transition time has been the object of numerous works. Due to the difficulty in solving the complete differential equation representing the discharge (charge) of the gate output node, various attempts have been done to characterize this output transition time, including step [1], ramp [2] and exponential models [3]. In [4], a submicron delay and output slope modeling is given, still limited to fast input transitions. Recently, as an extension of the work proposed in [4], S. Dutta [5], considered very slow input ramp effects. Both the delay and the output ramp duration are obtained by curve fitting between two extreme points corresponding to infinitely fast and infinitely slow inputs. As an improvement of his initial

work Sakurai [6] considered extremely fast and slow ramp conditions and solved intermediate cases from smooth interpolation between the two extremes. In [7] Bisdounis proposed a fast and slow input slope definition from the operating mode of the switching transistor however, no clear design oriented definition of both fast and slow input transition range, based on the size and the load of the switching and controlling devices appears available. Hirata in [8] proposed a piece wise linear representation of the current available in the switching structure. This approach necessitates a great number of calibrations with Spice simulations of the different technological parameters used in the representation.
In fact the output ramp duration of a CMOS structure depends on its current possibility (I_{MAX}) and of the amount of charge to be transferred ($C.V_{DD}$). As proposed in [9] it can be obtained from:

$$t_{OUT} = \frac{C_L.V_{DD}}{I_{MAX}} \quad (1)$$

where V_{DD} represents the node voltage variation and C_L it's output loading capacitance.
As shown the key parameter in modeling the output transition time is the current available in the switching structure of which determination depends on the structure, its size and the duration time of the input controlling edge. In order to complete an analytical model of delays developed for submicron CMOS structures [9], we present in this paper a design oriented macro modeling of the CMOS structure output transition time. In section 2 we present the method we used to obtain the value of the maximum current available in CMOS inverter and gates considering both fast and slow input ramp conditions. The modeling and the validation of output transition time is given in section 3. Section 4 draws a conclusion on this model.

2. INVERTER MAXIMUM CURRENT

Depending on the strength of the controlling structure two design conditions have to be considered, fast and slow input ramp conditions. Let us consider an inverter with a load C_L controlled by a rising linear input ramp of duration τ_{IN}. As shown in Fig.1, the current sunk from the load by the N transistor depends on the value of τ_{IN}:

- in region 1 the set up of the current of the N transistor follows the input ramp variation and exhibits a constant maximum value during all the discharge process, this defines the fast input range,
- in region 2 the maximum current is obtained before the input ramp reaches its maximum value, resulting in a smaller value of the charge evacuated by unit time. This defines the slow input range where the maximum value of the discharging current decreases when the input transition time increases.

2.1 Maximum current value for fast input range

During all the input ramping process the N transistor is saturated, its current maximum value is defined for $V_{IN} = V_{DD}$, resulting in:

$$I_{MAX}^{fast} = K_N.W_N.(V_{DD} - V_{TN}) \quad (2)$$

where K_N is the transistor conduction factor defined in [4] for α=1, V_{TN} and W_N the N transistor threshold voltage and width respectively.

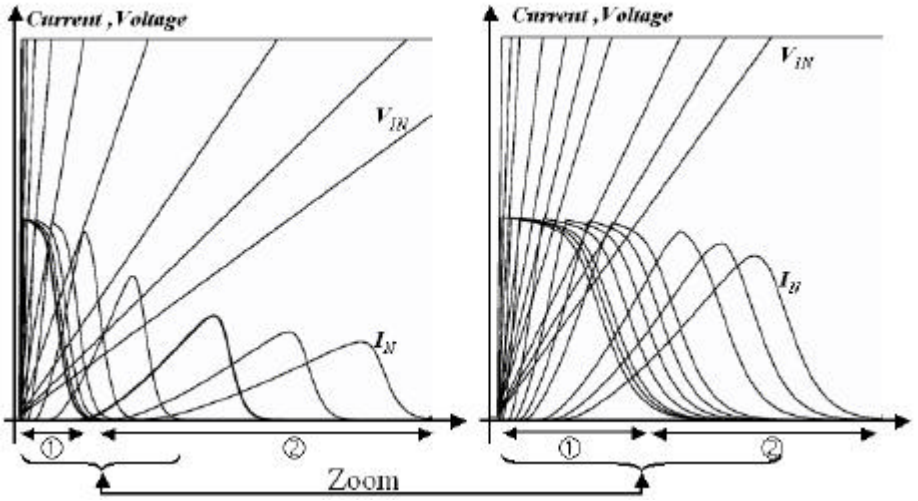

Fig. 1. Illustration of the fast • and slow • input controlling ranges of an inverter.

2.2 Maximum current value for slow input range

As in the preceding case the transistor is still in saturation when its current reaches the maximum value but its gate driving voltage is smaller and its value must be defined. For that we consider that in the time interval t_{VTN} - t_{MAX}, (Fig.2), the current exhibits a linear variation. This gives:

$$I_N(t) = K_N.W_N.(\frac{V_{DD}.t}{t_{IN}} - V_{TN}) \quad (3)$$

and:

$$\frac{\Delta I}{\Delta t} = \frac{I_{MAX}}{\Delta t} = \frac{K_N.W_N.V_{DD}}{t_{IN}} \quad (4)$$

where the input ramp duration time τ_{IN} is the output ramp duration of the controlling structure, as defined in eq.1.
Under the approximation that the current variation is symmetric with respect to its maximum value we can evaluate the total charge removed at the output node as:

$$\frac{C.V_{DD}}{2} = \frac{I_{MAX}.\Delta t}{2} \quad (5)$$

Combining eq. 4 and 5 we obtain the value of the maximum current resulting from a slow rising input controlling edge as:

$$I_{MAX}^{slow} = \sqrt{\frac{K_N.W_N.V_{DD}^2.C}{t_{IN}}} \quad (6)$$

where $C.V_{DD}$ represents the total charge to be removed from the output node, where: $C=C_L+C_{SC}+C_{PAR}$ in which C_L and C_{PAR} represent the inverter active load (output loading gates) and the output parasitic capacitance respectively, C_{SC} is the short circuit equivalent capacitance which represents the charge by volt unit between the supply rail during the discharge process as defined in [10,11].

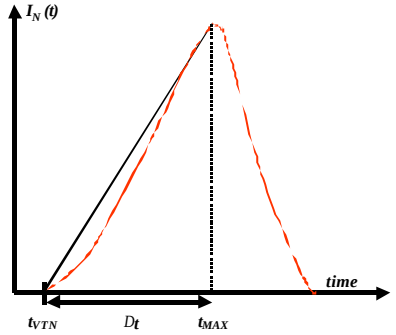

Fig. 2. Discharging current evolution.

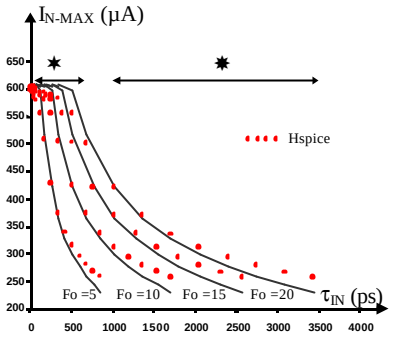

Fig. 3. Comparison between calculated and simulated maximum discharging current value; • and • label the fast and slow input ranges.

We compare in Fig.3 the maximum current values deduced from eq.2 and 6 to the values simulated with Hspice for an inverter defined by W_N=1µm, W_P=2.2µm, L=0.18µm for different loading conditions (5,10,15 and 20 times its input capacitance C_{IN}=4.5fF). As shown we obtain a very good agreement between simulated and calculated values (less than 10% discrepancy) over the considered full design range.

2-3 Maximum current value for a simple gate

To evaluate the maximum current available in a gate it is necessary to consider the current limitation effect produced by the serial array of transistors, together with the multiplication effect produced by the dual parallel array. The current possibility of this parallel array is input vector dependent, but bounds can be easily defined considering, for an n input gate, one or n times the maximum current of an inverter with identically sized transistors. The reduction of the serial array to an equivalent transistor has been the object of numerous works [8,12-14]. To reduce a gate to an equivalent inverter we present here a new reduction method by considering the serial array of n transistors as an input voltage controlled current generator, as illustrated in Fig.4.

If we consider a control on the top input (Bot and Mid inputs connected to V_{DD}) of the Nand3, we can see easily that the voltage dropt through the Mid and Bot transistors, working in linear mode, reduces the voltage swing of the controlling gate. This results in a transistor size dependent reduction of the available current in the network with can be modeled as a reduction factor equal to the ratio of the currents available in the array and in the inverter with identically sized transistors. This gives:

$$\mathrm{Re}\,d_{fast} = 1 + K_N . W_N . R_N \qquad (7)$$

where R_N represents the sum of the resistance of the bottom transistors.

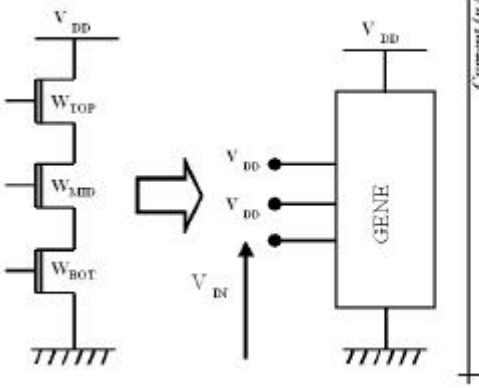

Fig. 4. Reduction of the serial array of transistors to a multiple input voltage controlled current generator.

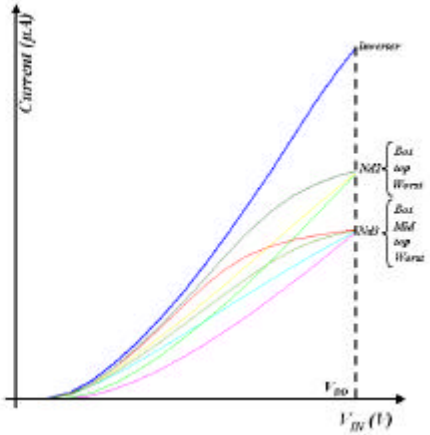

Fig. 5. Static I/V characteristics of the current generator with respect to the controlling input (Top, Mid, Bot).

For a control on the bottom input, for fast input ramps, the intermediate nodes are discharged faster than the output one. In this case the current is still limited by the top transistor and the reduction factor is given by eq.7.

For slow input ramp condition the bottom and top transistors operate in saturated mode and the current is limited by the bottom transistor working with a reduced drive and drain source voltage. In this condition it appears necessary to calibrate, from simulations on the process, the conduction factor of the bottom transistor in the serial array [15]. For the process under study (0.18µm) values of Red_{SLOW}= 1.2, 1.48 and 1.78 have been obtained for NAND 2, 3 and 4 respectively, which are quite different from the values obtained for fast edge conditions (1.55, 2.1 and 2.6 for NAND 2,3,4 respectively) or from a direct reduction based on the number of serial transistors [4].

Controlling the middle input we obtain a superposition of the contributions of the preceding effects. The reduction factor can easily be deduced from the preceding cases considering the middle transistor in top or bottom position for the bottom or top transistor of the array, respectively, resulting in a reduction factor:

$$\mathrm{Re}\,d = \mathrm{Re}\,d_{FAST} . \mathrm{Re}\,d_{SLOW} \qquad (8)$$

3. OUTPUT TRANSITION TIME

The output transition time can be obtained easily from eq.1 by replacing I_{MAX} by the expressions previously developed.

3.1 Inverters

Considering fast and slow input ramp conditions results in:

$$t_{OUT} = MAX\left\{ t_{OUT}^{fast};\ \sqrt{\frac{V_{DD}-V_{TN}}{V_{DD}}}.\sqrt{t_{IN}\, t_{OUT}^{fast}} \right\} \quad (9)$$

with:

$$t_{OUT}^{fast} = t_{ST}.\frac{C_L}{C_N} = 2.T_{HLS} \quad (10)$$

$$t_{ST} = \frac{V_{DD}.L_{GEO}.C_{OX}}{(V_{DD}-V_{TN}).K_N} \quad (11)$$

In these equations T_{HLS} represents the step response of the inverter, and τ_{ST} the shorter switching time of the process, as defined in [10]. Validation of these expressions has been realized on different configurations of inverters in various loading and controlling conditions by comparing simulated (Hspice BSIM3 level 49) and calculated (eq.9) output duration time values. The results obtained are illustrated in Fig. 6-7. The output transition time evolution is given versus the ratio τ_{IN}/T_{HLS} used as a metric for input transition times. The expression for an output rising edge can be obtained by exchanging N and P suffixes.

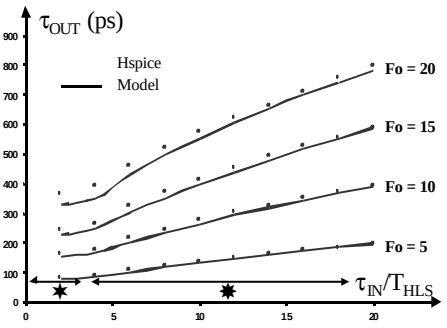

Fig. 6. Inverter output transition time (W_N=1µm, k=2, L=0.18µm) loaded by 5 to 20 C_{IN}.

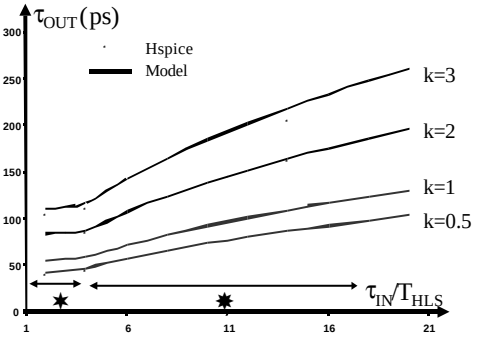

Fig. 7. Inverter output transition time (W_N=1µm, W_P=2µm, L=0.18µm) loaded by 5 to 20 C_{IN}.

As shown we obtain a very good agreement between simulated and calculated values (less than 10% discrepancy) over the considered full design range.

3.2 Gates

Considering the current reduction factors defined in eq.7-8, the generalization to gates is straightforward, we obtain:

- **for a Top input control:**

$$t_{OUT} = MAX\left\{ \mathrm{Re}\, d_{FAST}.t_{OUT}^{fast}; \sqrt{\frac{\mathrm{Re}\, d_{FAST}.(V_{DD}-V_{TN})}{V_{DD}}}.\sqrt{t_{IN}\, t_{OUT}^{fast}} \right\} \quad (12)$$

- **for a Bot input control:**

$$t_{OUT} = MAX\left\{ \mathrm{Re}\, d_{FAST}.t_{OUT}^{fast}; \sqrt{\frac{\mathrm{Re}\, d_{SLOW}.(V_{DD}-V_{TN})}{V_{DD}}}.\sqrt{t_{IN}.t_{OUT}^{fast}} \right\} \quad (13)$$

- **for a Mid input control:**

$$t_{OUT} = MAX\left\{ \mathrm{Re}\, d_{FAST}.t_{OUT}^{fast}; \sqrt{\frac{\mathrm{Re}\, d.(V_{DD}-V_{TN})}{V_{DD}}}.\sqrt{t_{IN}.t_{OUT}^{fast}} \right\} \quad (14)$$

Validation has been done following the same procedure than for inverters. Table 1 and 3 are relative to Top and Bottom controlled Nand2,3 (W_N=W_P=1µm) loaded by 10.C_{IN} and implemented in a 0.18µm process. As shown we obtain a very good agreement between simulated and calculated values of the output transition time.

τ_{IN}/T_{HLS}	Nand2_Top			Nand3_Top			Nand2_Bot			Nand3_Bot		
	SIM	**CAL**	D**%**	**SIM**	**CAL**	D**%**	**SIM**	**CAL**	D**%**	**SIM**	**CAL**	D**%**
2	144	154	7%	240	223	5%	134	121	10%	197	204	6%
6	178	172	3%	256	252	8%	166	155	7%	209	204	0%
10	227	222	2%	315	325	3%	201	199	1%	233	217	1%
16	291	281	3%	392	385	2%	275	268	3%	296	291	5%
20	329	315	4%	439	436	2%	292	282	4%	312	306	4%

Table 1. Comparison between simulated and calculated values of output transition time for NAND2, 3 with top and bot input control.

3.3 Discussion on slow and fast ranges of input transition times

Valuable criteria in evaluating the quality of designs or in defining metric for design performance optimization is to clearly identify the limit condition between fast and slow transition times. As previously illustrated this limit (fig.1) corresponds to the threshold between the availability of constant discharging (charging) current and varying one. It depends of the relative values of the input and output transition times.

For example let us identify this limit for inverters in equalizing the two terms of eq.9, (eq.12-14 for gates). This gives the limit at which input ramps must be considered as slow as:

$$t_{IN} \geq \left(\frac{V_{DD}}{V_{DD} - V_{TN}} \right) t_{OUT}^{FAST} \qquad (15)$$

Remembering that on an array τ_{IN} represents the input transition time of the controlling inverter (i-1) and τ_{OUT} the output transition time of the switching device (i) we obtain from eq.15:

$$Fo(i-1) \geq \frac{V_{DD} - V_{TN}}{V_{DD}} \left(\frac{k}{R_m}\right).Fo(i) \qquad (16)$$

$$Fo(i-1) \geq \frac{V_{DD} - |V_{TP}|}{V_{DD}} \left(\frac{R_m}{k}\right).Fo(i) \qquad (17)$$

for output falling and rising edges, respectively, where Rμ is the ratio of the conduction factors of N and P transistors, k and Fo have been previously defined. Extension to gates can be easily obtained from eq.12-14, including the reduction factors.

In table 2, we compare the limit value of the τ_{IN} separating fast and slow input range as defined in eq.15, to the values deduced from the simulation on an inverter for different configuration ratio values and loading conditions. As shown the limit previously defined is in very good agreement with the values obtained from the simulations.

t_{IN} limit (ps)	Fo=5		Fo=20	
	Sim	Cal	Sim	Cal
k=1	63.1	60	243	240
k=2	94.6	90	365	360
k=3	126	120	487	480

Table 2. Comparison between simulated and calculated (eq.15) values of the limit value between fast and slow range defined for τ_{IN}.

In the figure 8 we illustrate the relative character of the definition of the limit between fast and slow input ranges. The curves represent the output voltage and discharging

current controlled by an input ramp of 50 and 500ps of duration with different output loads. As shown, (fig.8.a) while the input duration ramp is quite short (50ps), due to the weakness of the load (Fo=0.5) the input control must be considered as slow. On the other hand (fig8.b) a heavy loaded inverter (Fo=80) controlled by a quite long duration ramp (τ_{IN}=500ps) is controlled under fast input ramp conditions. This results is very important, this justifies why, as well for defining design validation range than look up tables, it is necessary to define the input control range relatively to the output transition time of the considered cell.

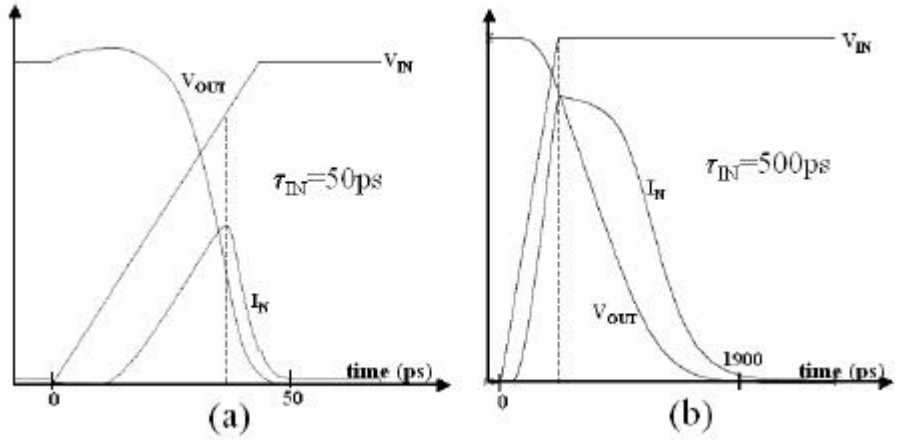

Fig. 8. Illustration of the relative definition of slow ((a) τ_{IN}=50ps) and fast ((b) τ_{IN}=500ps) input range, obtained with short and long duration ramps respectively.

4. CONCLUSION

We derived design oriented simple and closed form formula for the output transition time of CMOS gates. We showed that the proposed expression reproduces the sensitivity to the design and process parameters. Based on a metric defined on inverter for fast input ramp conditions the formula includes deep submicron effects by considering the variation of the maximum current available with the input edge. Extension has been done to gates by reduction to an equivalent inverter, considering the different input control conditions. Clear evidence of different reduction factor values for fast and slow input edges has been demonstrated. Validations through Hspice simulations for a 0.18µm process confirmed the validity of the proposed expressions which can easily be used to replace look up tables in timing estimator.

Clear definition of the slow and fast input control range is clearly defined and demonstrated. Application to edge control for low power buffer design is under development.

5. REFERENCES

[1] J.R. Burns, "Switching response of complementary symmetry MOS transistor logic circuits" RCA Review, vol. 25, pp627-661, 1964.

[2] N. Hedenstierna and K.O. Jepson, " CMOS circuit speed and buffer optimization", IEEE Trans. Computer-Aided Design, vol. 6, pp. 270-281, March 1987.

[3]I. Kayssi Ayman, A. Sakallah Karem, M. Burks Timothy, " Analytical transient response of CMOS inverters" IEEE Trans. on circuits and Syst. Vol. 39, pp. 42-45, 1992

[4]T. Sakurai and A.R. Newton, "Alpha-power model, and its application to CMOS inverter delay and other formulas", J. of Solid State Circuits vol. 25, pp. 584-594, April 1990.

[5]Santanu Dutta, Shivaling S. Mahant Shetti, and Stephen L. Lusky, "A Comprehensive Delay Model for CMOS Inverters" J. of Solid State Circuits, vol. 30, no. 8, pp. 864-871, 1995.

[6] T. Sakurai, A. R. Newton “A simple MOSFET model for circuit analysis” IEEE Trans. On electron devices, vol.38, n°4, pp. 887-894, April 1991.

[7] L. Bisdounis, S. Nikolaidis, O. Koufopavlou "Analytical transient response of propagation delay evaluation of the CMOS inverter for short channel devices” J. of Solid State Circuits vol. 33, n°2, pp. 302-306, Feb.1998.

[8]A. Hirata, H. Onodera, K. Tamaru "Proposal of a Timing Model for CMOS Logic Gates Driving a CRC π load" in proc. of the Int. Conf. On CAD 1998 (San Jose), pp 537-544.

[9] D. Auvergne, J. M. Daga, M. Rezzoug, “Signal transition time effect on CMOS delay evaluation ”IEEE Trans. on Circuit and Systems-1, vol.47, n°9, pp.1362-1369, sept.2000

[10] J. M. Daga, D. Auvergne "A comprehensive delay macromodeling for submicron CMOS logics" IEEE J. of Solid State Circuits Vol.34, n°1, pp.42-55, 1999.

[11] S. Turgis, D. Auvergne "A novel macromodel for power estimation for CMOS structures" IEEE Trans. Computer-Aided-Design vol.17, n°11, pp1090-1098, nov.98.

[12] A. Chatzigeorgiou , S. Nikolaidis "Collapsing the Transistor Chain to an Effective Single Transistor" Date 1998.

[13] A. Nabavi-Lishi "Inverter Models of CMOS Gates for Supply Current and Delay Evaluation" IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, Vol. 13, N° 10, October 1994.

[14] Q. Wang and S. B. K. Vrudhula "A New Short Circuit Power Model for Complex CMOS Gates", in Proc. IEEE Alessandro Volta Memorial Workshop on Low Power Design (Volta99), pp. 98-106, Como Italy, Mar. 4-5, 1999.

[15] K. Nose, T. Sakurai “Analysis and future trend of short circuit power” IEEE Trans. Trans. Computer- Aided-Design vol. 19, n°9, pp.1023-1030, Sept. 2000.