Physical Extension of the Logical Effort Model

B. Lasbouygues[1], R. Wilson[1], P. Maurine[2], N. Azémard[2], and D. Auvergne[2]

[1] STMicroelectronics Design Department 850 rue J. Monnet, 38926, Crolles, France
{benoit.lasbouygues,robin.wilson}@st.com
2 LIRMM, UMR CNRS/Université de Montpellier II, (C5506),
161 rue Ada, 34392 Montpellier, France
{pmaurine,azemard,auvergne}@lirmm.fr

Abstract. The logical effort method has appeared very convenient for fast estimation and optimization of single paths. However it necessitates a calibration of all the gates of the library and appears to be sub-optimal for a complex implementation. This is due to the inability of this model in capturing I/O coupling and input ramp effects. In this paper, we introduce a physically based extension of the logical effort model, considering I/O coupling capacitance and input ramp effects. This extension of the logical effort model is deduced from an analysis of the supply gate switching process. Validation of this model is performed on 0.18µm and 0.13µm STM technologies. Application is given to the definition of a compact representation of CMOS library timing performance.

1 Introduction

In their seminal book [1], I. E. Sutherland and al introduced a simple and practical delay model. They developed a logical effort method allowing designers to get a good insight of the optimization mechanisms, using easy hand calculations. However this model suffers of a limited accuracy when strong timing constraints have to be imposed on combinatorial paths. An empirical extension of the logical effort model has been proposed in [2] for considering the input ramp effect. However the I/O coupling capacitance effect, first introduced by Jeppson [3], remains neglected in this method.

Moreover, with the dramatic increase of the leakage current, designers may be willing to design with multiple threshold voltage CMOS (MTCMOS) and several supply voltage values. As a consequence, it is necessary to develop a delay model allowing the designers to deal with multiple V_{TH} and V_{DD} values, with reduced calibration time penalty.

In this paper we introduce a physical extension of the logical effort. Both the I/O coupling and input ramp effects are considered together with the timing performance sensitivities to the V_{TH} and V_{DD} values.

The rest of the paper is organized as followed. In section 2, starting from the Sakurai's alpha power law [4], we first physically justify and then extend the logical effort model. Section 3 is devoted to the validation of the extended model and to its application to the definition of a new timing performance representation. Section 4 discusses briefly how to increase the accuracy in choosing the sampling points to be reported in traditional look up tables before to conclude in section 5.

E. Macii et al. (Eds.): PATMOS 2004, LNCS 3254, pp. 838–848, 2004.

2 Timing Performance Model

In the method of logical effort [1] the delay of a gate is defined by

$$d = \tau \cdot (p + gh) \tag{1}$$

τ is a constant that characterizes the process to be used, p is the gate parasitic delay. It mainly depends on the source/drain diffusion capacitance of the transistors. g, the logical effort of the gate, depends on the topology of the gate. h, the electrical effort or gain, corresponds to the ratio of the gate output loading capacitance to its input capacitance value. It characterizes the driving possibility of the gate.

Using the inverter as a reference these different parameters can be determined from electrical simulation of each library cell. However, if this simple expression can be of great use for optimization purpose no consideration is given to the input slew effect on the delay and to the input-to-output coupling [3]. As a result, the different parameter values cannot be assumed constant over the design space.

As it can be shown from [5], eq.1 characterizes the gate output transition time. However, timing performance must be specified in terms of cell input (output) transition time and input-to-output propagation delay. A realistic delay model must be input slope dependent and must distinguish between falling and rising signals. In order to get a more complete expression for the gate timing performance, let us develop eq.1, starting from a physical analysis of the switching process of an inverter. We first consider a model for the transition time.

2.1 Inverter Transition Time Modeling

Following [5], the elementary switching process of a CMOS structure, and thus of an inverter, can be considered as an exchange of charge between the structure and its output loading capacitance. The output transition time (defining the input transition time of the following cell) can then be directly obtained from the modeling of the charging (discharging) current that flows during the switching process of the structure and from the amount of charge ($C_{L\text{-}TOT} \cdot V_{DD}$) to be exchanged with the output node as

$$\tau_{outHL} = \frac{C_{L-TOT} \cdot V_{DD}}{I_{NMax}} \quad \tau_{outLH} = \frac{C_{L-TOT} \cdot V_{DD}}{I_{PMax}} \tag{2}$$

V_{DD} is the supply voltage value and $C_{L\text{-}TOT}$ the total output load which is the sum of two contributions: $C_{L\text{-}TOT} = C_{INT} + C_{EXT}$. C_{INT} is the internal load proportional to the gate input capacitance, constituted of the coupling capacitance, C_M, between the input and output nodes [3] and of the transistor diffusion parasitic capacitance, C_{DIFF}. Note that C_M can be evaluated as one half the input capacitance of the P(N) transistor for input rising (falling) edge [6], or directly calibrated on Hspice simulation. C_{EXT} is the load of the output node, including the interconnect capacitive component.

In eq.2 the output voltage variation has been supposed linear and the driving element considered as a constant current generator delivering the maximum current available in the structure. Thus the key point here is to determine a realistic value of this maximum current. Two controlling conditions must be considered.

Fast input control conditions

In the Fast input range, the input signal reaches its maximum value before the output begins to vary, in this case the switching current exhibits a constant and maximum value:

$$I_{MAX}^{Fast} = K_{N,P} \cdot W_{N,P} \cdot \left(V_{DD} - V_{TN,P}\right)^{\alpha_{N,P}} \tag{3}$$

This is directly deduced from the Sakurai's representation [4], $\alpha_{N,P}$ being the velocity saturation indexes of N, and P transistors, $K_{N,P}$ is an equivalent conduction coefficient to be calibrated on the process.

From (2, 3) we obtain the expression of the transition time for a Fast input control condition as

$$\begin{aligned} \tau_{outHL}^{Fast} &= \tau \cdot (1+k) \cdot \frac{C_{L-TOT}}{C_{IN}} \equiv \tau \cdot \left(p_{HL} + g_{HL} \cdot h\right) \\ \tau_{outLH}^{Fast} &= \tau \cdot \frac{(1+k)}{k} \cdot R \cdot \frac{C_{L-TOT}}{C_{IN}} \equiv \tau \cdot \left(p_{LH} + g_{LH} \cdot h\right) \end{aligned} . \tag{4}$$

τ and R, defined below, are respectively a unit delay characterizing the process, and the dissymmetry between N and P transistors while k is the transistor P/N width ratio. C_{IN} is the gate input capacitance.

$$\tau = \frac{C_{OX} \cdot L_{GEO} \cdot V_{DD}}{K_N \cdot \left(V_{DD} - V_{TN}\right)^{\alpha_N}} \qquad R = \frac{K_N \cdot \left(V_{DD} - V_{TN}\right)^{\alpha_N}}{K_P \cdot \left(V_{DD} - V_{TP}\right)^{\alpha_P}} \tag{5}$$

Finally $p_{HL,LH}$ and $g_{HL,LH}$ are the parameters characterizing the topology of the inverter under consideration. They are defined by:

$$\begin{aligned} p_{HL} &= \frac{(1+k) \cdot \left(C_{MHL} + C_{DIFF}\right)}{C_{IN}} & g_{HL} &= (1+k) \\ p_{LH} &= R \cdot \frac{(1+k) \cdot \left(C_{MLH} + C_{DIFF}\right)}{k \cdot C_{IN}} & g_{LH} &= R \cdot \frac{(1+k)}{k} \end{aligned} \tag{6}$$

As clearly shown the eq.4-6 constitute an explicit representation of the logical effort model [1] for non-symmetrical inverters.

Slow input control conditions

In the slow input range, the transistor is still in saturation when its current reaches the maximum value but its gate source voltage is smaller. This results in a smaller value of the maximum switching current. Let us evaluate this value. From the alpha power law model we can write

$$I_N(t) = K_N \cdot W_N \cdot \left(\frac{V_{DD} \cdot t}{\tau_{IN}} - V_{TN}\right)^{\alpha_N} \tag{7}$$

where τIN is the transition time of the signal applied to the gate input. This leads to

$$\frac{dI}{dt} = \frac{\alpha_N \cdot K_N \cdot W_N \cdot V_{DD}}{\tau_{IN}} \cdot \left(\frac{V_{DD} \cdot t}{\tau_{IN}} - V_{TN}\right)^{\alpha_N - 1} \tag{8}$$

Defining Δt as the time spent by the transistor to deliver its maximum current, (8) becomes:

$$\frac{\Delta I}{\Delta t}=\frac{\alpha_N\cdot K_N\cdot W_N\cdot V_{DD}}{\tau_{IN}}\cdot\left(\frac{V_{DD}\cdot\Delta t}{\tau_{IN}}\right)^{\alpha_N-1}\equiv\frac{I_{NMAX}}{\Delta t} \tag{9}$$

Then 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:

$$\frac{C_{L-TOT}\cdot V_{DD}}{2}=\frac{I_{MAX}\cdot\Delta t}{2} \tag{10}$$

Combining eq.9 and 10, we obtain the value of the maximum current for a slow input ramp applied to the input as

$$I_{N-MAX}^{Slow}=\left\{\frac{\alpha^{\frac{1}{\alpha_N}}\cdot K_N^{\frac{1}{\alpha_N}}\cdot W_N^{\frac{1}{\alpha_N}}\cdot C_{L-TOT}\cdot V_{DD}^2}{\tau_{IN}}\right\}^{\frac{\alpha_N}{1+\alpha_N}} \tag{11}$$

Finally, by combining eq.2 and 11, the expression of the transition time for slow input control condition is obtained as

$$\tau_{outHL,LH}^{Slow}=\left\{\left(\frac{V_{DD}-V_{TN,P}}{\alpha_{N,P}^{\frac{1}{\alpha_{N,P}}}\cdot V_{DD}}\right)^{\alpha_{N,P}}\cdot\tau_{outHL,LH}^{Fast}\cdot\tau_{IN}^{\alpha_{N,P}}\right\}^{\frac{1}{1+\alpha_{N,P}}} \tag{12}$$

Defining the supply voltage effort as

$$U_{N,P}=\frac{V_{DD}-V_{TN,P}}{\alpha^{\frac{1}{\alpha_{N,P}}}\cdot V_{DD}} \tag{13}$$

results in a logical effort like expression of the output transition time for a slow input ramp condition

$$\tau_{outHL,LH}^{Slow}=\left\{U_N^{\alpha_{N,P}}\cdot\tau_{IN}^{\alpha_{N,P}}\cdot\tau\cdot\left(p_{HL,LH}+g_{HL,LH}\cdot h\right)\right\}^{\frac{1}{1+\alpha_{N,P}}} \tag{14}$$

Note here that for advanced processes, in which the carrier speed saturation dominates ($\alpha_{N,P}$=1), this expression can be simplified

$$\tau_{outHL,LH}^{Slow}=\sqrt{U_{N,P}\cdot\tau_{IN}\cdot\tau\cdot\left(p_{HL,LH}+g_{HL,LH}\cdot h\right)} \tag{15}$$

with

$$U_{N,P}=\frac{V_{DD}-V_{TN,P}}{V_{DD}} \tag{16}$$

Unifying Fast and Slow domain representation

Let us now consider the case of a rising edge applied to a gate input. As it can be deduced from (4) and (14), the logical effort delay model

$$\tau_{outHL,LH}^{Fast} = \tau \cdot \left(p_{HL,LH} + g_{HL,LH} \cdot h\right)$$

can be used as a metric for both Fast and Slow input ramp domains. Indeed, considering the sensitivity of the different expressions to the input slope, we can express the normalized inverter output transition time as

$$\frac{\tau_{outHL,LH}}{\tau_{outHL,LH}^{Fast}} = Max\begin{bmatrix} 1 \\ \left\{U_{N,P} \cdot \sigma_{HL,LH}\right\}^{\frac{\alpha_{N,P}}{1+\alpha_{N,P}}} \end{bmatrix} \tag{17}$$

where $\sigma_{HL,LH}$, is equivalent to an input slew effort

$$\sigma_{HL,LH} = \frac{\tau_{IN}}{\tau_{outHL,LH}^{Fast}} \tag{18}$$

Eq.17 is of great interest, since it clearly shows that the output transition time of any inverter of a given library can be represented by a single expression. As shown, the right part of eq.17 is design parameter independent. As a result this gives the opportunity, in a library, to characterize the complete set of drive of an inverter by one look up table of one line. This is obtained by using a representation relative to the slew effort parameter σ. As a validation of these results we represent in Fig.1, the output transition time variation (with respect to σ_{HL}) of 7 inverters of a 0.13µm process. As expected all the curves pile up on the same one, representing the output transition time sensitivity to the slew effort σ_{HL}.

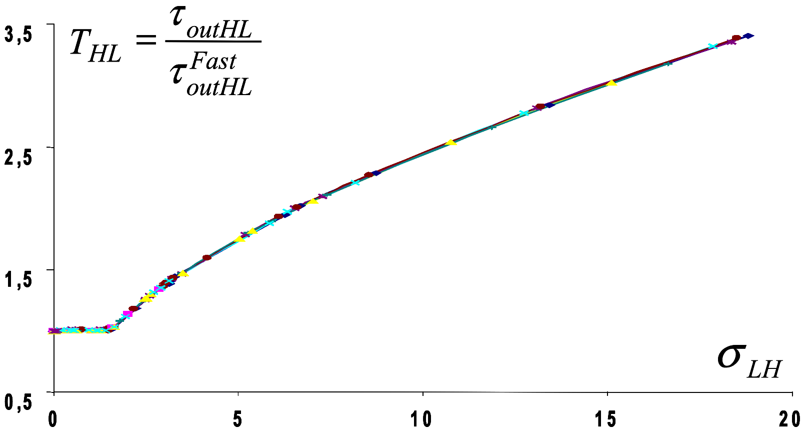

Fig. 1. Output transition time of the 7 inverters of a 0.13µm process

2.2 Extension to Gates

Following [7], the extension to gates is obtained by reducing each gate to an equivalent inverter. For that we consider the worst-case situation. The current capability of the N (P) parallel array of *m* transistors is evaluated as the maximum current of an inverter with identically sized transistors. The array of N (P) series-connected transistors is modeled as a voltage controlled current generator with a current capability reduced by a factor $\mathrm{DW}_{\mathrm{HL,LH}}$. This reduction factor (DW) is defined as the ratio of the current available in an inverter to that of a series-connected array of transistors of identical size.

$$DW_{HL,LH} = \frac{I_{N,P}(Inv)}{I_{N,P}(Gate)} \cdot \frac{W_{N,P}(Inv)}{W_{N,P}(Gate)} \tag{19}$$

DW corresponds to the explicit form of the logical effort [3]. Let us evaluate the expression of $\mathrm{DW}_{\mathrm{HL'LH}}$. In the Fast input range, the maximum current that can provide an array of *n* serially connected transistors is defined by:

$$\frac{\mathrm{I_R}}{\mathrm{K}_{\mathrm{N},P} \cdot \mathrm{W}_{\mathrm{N},P}} = \left\{\mathrm{V_{DD}} - \mathrm{V}_{\mathrm{TN},P} - (n-1)\mathrm{R}_{\mathrm{N},P} \cdot \mathrm{I_{ARRAY}}\right\}^{\alpha_{N,P}} \tag{20}$$

where $\mathrm{R_{N,P}}$ is the resistance of the bottom transistors working in linear mode. Using a first order binomial decomposition of (20), the reduction factor for Fast input controlling conditions is obtained as

$$DW_{HL,LH} = 1 + \alpha_{N,P} \cdot K_{N,P} \cdot W_{N,P} \cdot (n-1) \cdot R_{N,P} \cdot \left(V_{DD} - V_{TN,P}\right)^{\alpha_{N,P}-1} \tag{21}$$

which is a generalization of the result introduced in [7] for a DSM process, with full carrier speed saturation (α=1), where the value of the reduction factor reduces to

$$DW_{HL,LH} = 1 + K_{N,P} \cdot W_{N,P} \cdot (n-1) \cdot R_{N,P} \tag{22}$$

Using this reduction factor of the maximum current (21) to evaluate the output transition time (2) results in the normalized gate output transition time expression

$$\frac{\tau_{outHL,LH}}{\tau_{outHL,LH}^{Fast}} = Max\left[\frac{1}{\left\{U_{N,P} \cdot \sigma_{HL,LH}\right\}^{\frac{\alpha_{N,P}}{1+\alpha_{N,P}}}}\right] \tag{23}$$

where

$$\tau_{outHL,LH}^{Fast} = \tau \cdot \left(p_{HL,LH} + g_{HL,LH} \cdot h\right) \tag{24a}$$

and

$$\begin{aligned} p_{HL} &= DW_{HL} \cdot \frac{(1+k) \cdot (C_{MHL} + C_{DIFF})}{C_{IN}} & \qquad g_{HL} &= DW_{HL} \cdot (1+k) \\ p_{LH} &= R \cdot DW_{LH} \cdot \frac{(1+k) \cdot (C_{MLH} + C_{DIFF})}{k \cdot C_{IN}} & \qquad g_{HL} &= R \cdot DW_{LH} \cdot \frac{(1+k)}{k} \end{aligned} \tag{24b}$$

As shown p and g, the parasitic contribution to the gate delay and the logical effort, strongly depend on the topology of the considered cell through the parameters k and DW. Expression (23) constitutes an extension of the logical effort model considering the input ramp effect.

2.3 Gate Propagation Delay Model

A realistic delay model must be input slope dependent and must distinguish between falling and rising signals. As developed by Jeppson in [3], considering the input-to-output coupling effect, the input slope effect can be introduced in the propagation delay $\theta_{HL,LH}$ as

$$\begin{aligned} t_{HL} &= \frac{v_{TN}}{2}\tau_{IN} + (1 + \frac{2C_M}{C_M + C_L + C_{DIFF}})\frac{\tau_{outHL}}{2} \\ t_{LH} &= \frac{v_{TP}}{2}\tau_{IN} + (1 + \frac{2C_M}{C_M + C_L + C_{DIFF}})\frac{\tau_{outLH}}{2} \end{aligned} \qquad (25)$$

τ_{IN} ($\tau_{out,HL,LH}$) is the transition time of the input (output) signal, generated by the controlling gate. As shown, for each edge, the delay expression is a linear combination of the output transition time of the controlling and the switching gate.

Normalizing (25) with respect to the metric τ_{out}^{Fast}, gives a generic gate propagation delay expression for all the drives of a cell

$$\Theta_{HL,LH} = \frac{t_{HL,LH}}{\tau_{outHL,LH}^{Fast}} = \frac{v_{TN,P}}{2}\sigma_{HL,LH} + \frac{2 \cdot \alpha_{HL,LH}}{(\alpha_{HL,LH} + A) + h} + \frac{\tau_{outHL,LH}}{2 \cdot \tau_{outHL,LH}^{Fast}} \qquad (26)$$

In this equation $\alpha_{HL,LH}$ are the Meyer coefficients [6] for falling and rising edges that can be calibrated on the process the average value is $\alpha = 0.5$). The parameter A is related to the drain diffusion capacitance (C_{DIFF}=A.C_{IN}).

For a typical cell, the second term of eq.26 could be neglected for value of the electrical effort h greater than 3 or equivalently for important value of the input slew effort $\sigma_{HL,LH}$.

Although this term becomes quickly negligible, we have to note here that it can have a significant effect when imposing a small value of the electrical effort. It directly shows the minimum limit of the electrical effort value to be reasonably imposed on a path. Indeed, for small values of h (<2), the first and third terms of (25) become smaller than the second one. Any further increase of the transistor width along the path is inefficient in improving the corresponding gate speed that is limited by the I/O coupling effect and the parasitic content of the gate.

Combining finally (25) and (23) gives a general expression of the normalized propagation delay of any combinational gate as:

$$\Theta_{HL,LH} = \left[\begin{array}{c} \dfrac{1 + v_{TN,P} \cdot \sigma_{HL,LH}}{2} \\ \\ \dfrac{v_{TN,P} \cdot \sigma_{HL,LH} + \left\{ U_{N,P} \cdot \sigma_{HL,LH} \right\}^{\frac{\alpha_{N,P}}{1+\alpha_{N,P}}}}{2} \end{array} \right] + \frac{2 \cdot \alpha_{HL,LH}}{(\alpha_{HL,LH} + A) + h} \qquad (27)$$

This expression is of great interest. A part the last term, that is negligible in a typical design range (h between 3 and 6), eq.27 clearly indicates that the propagation delay of any gate of a library can be represented with only one equation.

3 Validation

In order to validate this model we first compare the estimation of performance of an inverter, designed in a 0.18µm process, calculated from eq.3, 11 and 17, to values obtained from Hspice simulations. Different supply voltage conditions have been considered. The value of the index saturation index has been obtained from direct calibration on the transistor current simulation. Fig.2-3 give some example of the calculated and simulated evolutions of both the inverter maximum switching current and output transition time with respect to σ_{HL}.

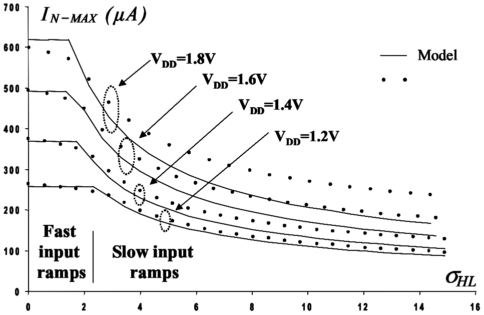

Fig. 2. Comparison between simulated and calculated values of the maximum switching current provided by an inverter (W_N=1µ k=2, l=0.18µm, $F_O=C_L/C_{IN}$=5) for different supply voltage values.

As shown the accuracy of the model is satisfactory. Note that the underestimation of the switching current for high supply voltage is mainly due to the short circuit current. This has a minor impact on the evaluation of the output transition time that is determined from the 40% and 60% points of the signal voltage swing.

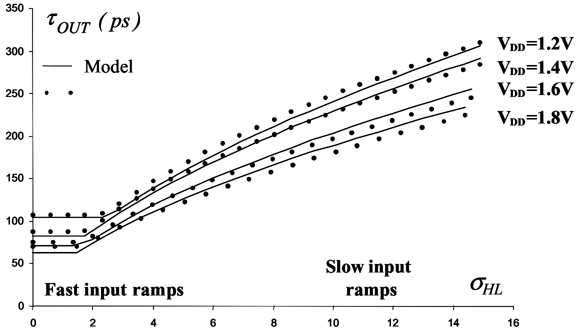

Fig. 3. Comparison between simulated and calculated values of the output transition time of an inverter (W_N=1µ k=2, l=0.18µm, $F_O=C_L/C_{IN}$=5) for different supply voltage values.

A more global validation has also been performed. Following eq.23 and 26, and using the appropriate representation, it is easy to deduce from the model that the transition time and the propagation delay variations of the different drive possibilities of a gate can be represented by a set of four laws, one by edge of each performance parameter. We validate this representation by representing in Fig.1, 4-6 the simulated input transition time sensitivities of the output transition time and delay of inverter, Nand and Nor gates, from a 0.13μm library (at least five drive each). All the values are normalized with respect to τ_{out}^{Fast}. As shown, all the simulated sampling points extracted from the Timing Library Format, corresponding to a particular gate, pile up on one curve. This clearly demonstrates that, using τ_{out}^{Fast} as a metric of performance, it is possible not only to simplify the library timing performance representation [8], but also to minimize the number of data to be simulated for the complete library characterization.

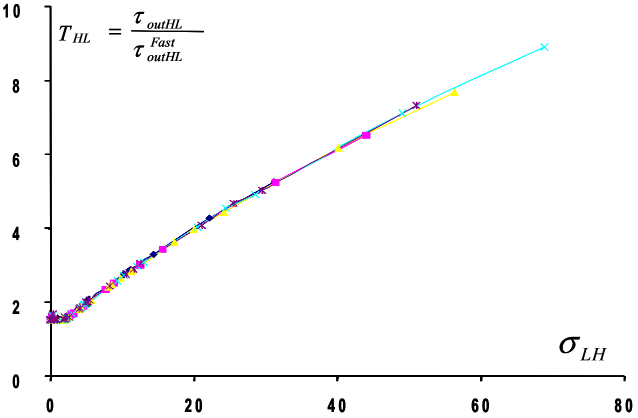

Fig. 4. Output transition time representation of the 5 Nand2 of a 0.13μm process

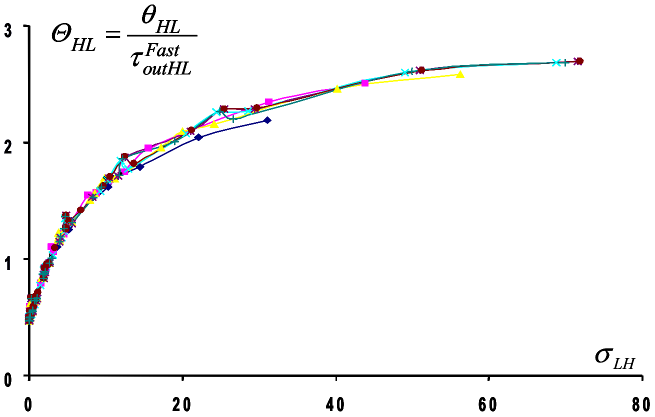

Fig. 5. Propagation delay representation of the 7 inverters of a 0.13μm process.

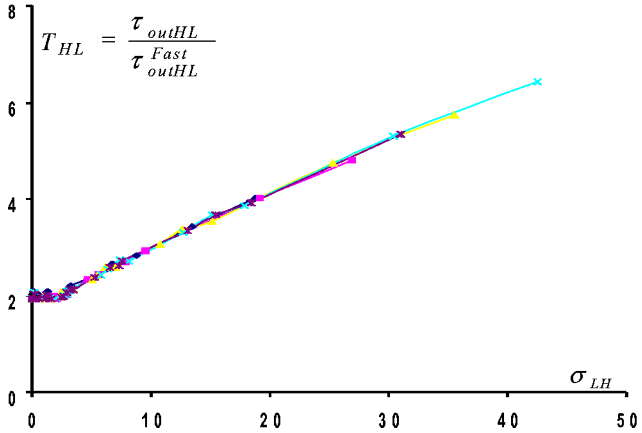

Fig. 6. Output transition time representation of the 5 Nor2 of a 0.13µm process

4 Discussion

In submicron process the transition time and the propagation delay exhibit a non-linear variation with respect to the controlling and loading conditions (Fig. 1,4-6). This non-linear range must clearly be determined with closest simulation steps because interpolating in this range may induce significant errors. It is obvious that the relative accuracy, obtained with a tabular method, is strongly dependent on the granularity of the table that is not necessarily constant but must cover a significant part of the design range. For that, indications for defining the granularity and the coverage of the design space must be available. As illustrated in Fig. 1,4-6, the non-linearity corresponds to the Fast /Slow input ramp domain boundary. Thus the evaluation of this boundary is of great interest to determine the data points to be sampled and reported in the look up tables. The proposed model offers an easy way to determine this boundary. It is just necessary to find the controlling and loading conditions for which eq.1 and 17 have the same value. This gives

$$\tau_{IN}^{lim} = \frac{\tau}{U_{N,P}^{\frac{1}{\alpha_{N,P}}}} \left\{ p_{HL,LH} + g_{HL,LH} \cdot \frac{C_L}{C_{IN}} \right\} \qquad (28)$$

Then to increase the accuracy of the tabular approach it is necessary to fix the granularity of the table in such a way that the data points belonging to the diagonal of the table respect the condition defined by eq.28.

5 Conclusion

We have introduced an explicit extension of the logical effort model in order to consider both the I/0 coupling and the input ramp effects and defined a simple but accurate representation of the timing performance of the simple CMOS structures. Validation of the model has been done on 0.18µm and 0.13µm processes. Application has been given to the definition of a compact representation of CMOS library timing performance. As discussed this representation must be of great interest in defining the granularity and the evolution of the look up tables used to represent the timing performance of CMOS library.

References

[1] I. Sutherland, B. Sproull, D. Harris, "Logical Effort: Designing Fast CMOS Circuits", Morgan Kaufmann Publi., California, 1999.

[2] X. Yu, V. G. Oklobdzija, W/ Walker, "Application of the logical effort on design arithmetic blocks" 35th Asilomar Conf. On Signals, Systems and Computers, California, Nov. 2001,pp 872-874

[3] K. O. Jeppson, "Modeling the influence of the transistor gain ratio and the input-to-output coupling capacitance on the CMOS inverter delay", IEEE J. SSC, vol.29, pp.646-654, 1994.

[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] C. Mead and L. Conway, “Introduction to VLSI systems”, Reading MA: Addison Wesley 1980.

[6] J. Meyer "Semiconductor Device Modeling for CAD" Ch. 5, Herskowitz and Schilling ed. Mc Graw Hill, 1972.

[7] "Not given for anonymous review" IEEE trans. on Computer Aided Design, 2002.

[8] Cadence open book "Timing Library Format References" User guide, v. IC445, 2000.