Improving Analytical Delay Modeling for CMOS Inverters

Felipe S. Marranghello, André I. Reis, and Renato P. Ribas

PGMicro, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
e-mail: fsmarranghello@inf.ufrgs.br

ABSTRACT

Analytical methods for gate delay estimation are very useful to speedup timing analysis of digital integrated circuits. This work presents a novel approach to analytically estimate the CMOS inverter delay. The proposed method considers the influence of input slope, output load and I/O coupling capacitance, as well as relevant effects such as channel length modulation and drain induced barrier lowering. Experimental results are on good agreement with HSPICE simulations, showing significant accuracy improvement compared to published related work. The delay model error has an average value of 3%, and the worst case error is smaller than 10%.

Index Terms: Analytical method, CMOS inverter, delay modeling, digital integrated circuits, timing analysis.

I. INTRODUCTION

The CMOS inverter is an essential element in digital VLSI integrated circuit (IC) design. Among other applications, inverter chains (buffers) are used in the distribution of clock signals and to drive large loads as those occurring in I/O pads. Because the analysis and optimization of buffers relying on electrical simulations tends to be a very time consuming task, many works have proposed analytical delay models targeting the CMOS inverter [1]-[25], and explored such models in buffer design optimization [34]-[36].

Even though the inverter is the simplest CMOS gate, defining an efficient and precise delay prediction is quite difficult due to the non-linear behavior of the circuit. The influence of load capacitance, input transition time, I/O coupling capacitance, short circuit current (SCC), velocity saturation, channel length modulation and drain-induced barrier lowering (DIBL) represent relevant challenges to obtain expressions for the inverter delay. However, most differential equations describing the inverter transient behavior do not have an analytical solution, requiring simplifications to be solved. However, such simplifications can impact model accuracy and fitting parameters may be needed to compensate errors.

Currently, no CMOS inverter delay estimation method considers effectively all effects mentioned above. In this work, a novel approach for estimating the propagation delay of a CMOS inverter is presented to cover such a lack. The proposed method provides accurate results for both fast and slow input transitions, requiring only transistor parameters. Simulation results have shown promising improvements in accuracy in comparison to previous works.

The rest of the paper is organized as follows. Section 2 discusses existing inverter delay models. Section 3 reviews some technical background useful for a better understanding of this work. The proposed CMOS inverter delay modeling is presented in Section 4. Section 5 provides simulation results and compares the proposed method to previously published works. Finally, Section 6 outlines the conclusions.

II. RELATED WORK

CMOS inverter delay models available in the literature can be roughly divided into three categories: (a) differential equation solving approaches, (b) charge based approaches, and (c) RC network approaches. Nevertheless, some methods can present characteristics from different categories.

2.1 Differential Equation Solving Modeling

This kind of inverter delay modeling relies on solving differential equations in order to obtain a precise description of CMOS inverter transient behavior. The resulting formulation is usually complex, although it is able to reproduce the entire output voltage waveform. Still, in most cases, the differential equations do not have an analytical solution and simplifications to the inverter behavior are needed.

Burns presents an expression for the inverter delay under a step input signal [1]. Hedenstierna and Jeppson consider a ramp with finite slew as input, which is assumed to be fast [2]. Jeppson improves the model from [2] by adding the influence of I/O coupling capacitance [3]. Bisdounis et al., provide explicit delay expressions for any input transition time considering both SCC and I/O coupling capacitance [3]. These models are based on long channel MOS transistors, being unsuitable for DSM devices [1]-[4].

Sakurai and Newton, in [5], propose the α-power transistor model for short channel devices. This transistor model is the standard choice for analytical inverter delay estimation. They also derive a timing model for the CMOS inverter, although neglecting SCC and I/O coupling capacitance. This model is extended to consider slow input transitions in [6] where an empirical estimation of SCC is performed by assuming that the output voltage is constant until the input signal reaches the inverter threshold voltage. In [7], Chandra et al. use the same formulation applied in [5] but considering an improved α-power law transistor model. In [12], another improved α-power model is presented by Consoli et al., and the differential equations for CMOS inverter transient behavior are solved using Taylor series. However, a closed-form formulation for the inverter delay is not provided and DIBL effect is modeled similarly to channel length modulation.

Bisdounis et al. consider the influence of both I/O coupling capacitance and SCC [9]. However, to determine if an input is fast, the output voltage must be evaluated as a function of time. In [10], Rossello and Segura solve the differential equations only considering the NMOS device, and correct the results by adding the influence of SCC using the model described in [30], which requires fitting parameters. In [11], a similar approach is applied by Chatzigeorgiou and Nikolaidis although only presenting expressions for the output voltage as time function. Wang and Zwolinski, in [12], consider channel length modulation but neglect SCC, and a closed formulation for inverter delay estimation is only presented when the input transition is considered fast. In [26], Alam et. al., consider channel length modulation although neglecting both SCC and I/O coupling capacitance. In [8], Cocchini et al. apply the BSIM3 transistor model but neglect the SCC impact.

From the works on this group, only the approaches presented in [6], [8], [12], [13] and [26] consider channel length modulation and only the ones proposed in [8] and [13] take into account DIBL effect.

2.2 Charge Based Inverter Delay Modeling

The main goal of models in this this category is to predict only the inverter delay rather than the whole output voltage waveform. Therefore, the resulting formulation tends to be simpler compared to differential equation solving approaches. The most adopted strategy is to determine the delay by estimating the total charge to be added (removed) from the output node considering an average (dis)charging current.

In [14], Deschacht et al. assume mean charge conservation to derive the delay expression for long channel devices, obtaining an expression similar to [2]. In [15], the model described in [14] is extended to include SCC through the use of fitting parameters. In [16], Daga and Auvergne adapt the modeling presented in [15] to short channel devices. Embabbi and Damodaran propose the utilization of an iterative approach to improve modeling of SCC, but still neglecting I/O coupling capacitance [17]. Such an iterative model is improved by Hamoui and Rumin, in [18], with better SCC estimation and considering I/O coupling capacitance. A different approach is exploited by Dutta et al., in [19]. They use the DC transfer curve to estimate the inverter transient response for very slow inputs. Nevertheless, fitting parameters are required and I/O coupling capacitance is neglected. In [20], Kabbani et al. add the influence of I/O coupling capacitance to the model described in [6]. However, this capacitance is only considered until the output voltage reaches the highest value.

In [21], Wang and Markovic adopt a slope correction term to include the impact of finite input slew. This correction term may be extracted through transient electrical simulations and depends on the transistor dimensions. Furthermore, such approach represents the gate delay behavior as a linear function of input transition time, so being only accurate for fast input transitions. Finally, in [22], Huang et al. divide the inverter response into overshoot period and discharging time, but SCC impact is neglected. The delay estimation for slow inputs considers that the discharging time rises linearly with the input transition time. In this category, none of the models consider channel length modulation and DIBL effects.

2.3 RC Based Modeling

In this category, the CMOS inverter is modeled as an RC network. The advantage of this strategy is that related equations are straightforward and easily

extendable to complex gates. However, RC based gate delay approaches fail to reproduce the non-linearity of inverter transient behavior. The well-known Elmore delay is widely adopted due to its simplicity [23]. In [24], Uebel and Bampi propose an exploit fitting parameters to improve accuracy of RC modeling. In [25], Mehri et al. analytically obtain average values for the transistor resistance considering the influence of input transition time. However, assuming similar input and output transitions and neglecting SCC.

2.4 General Considerations

In general, according to the discussion above, existing CMOS inverter delay methods exhibit at least one of the following drawbacks for application in modern (short channel) MOS technologies, as summarized in Table I:

a) Use long channel transistor models – The gate delay modeling is tied to the accuracy of the transistor model applied. For this reason, the utilization of long channel transistor models is not recommended because they cannot accurately predict the impact of short channel effects.

b) Neglect important parasitic effects – As MOS technology dimension shrinks, the influence of second order effects becomes even more important. Channel length modulation is one effect that deserves special attention for nanometer technologies [38]. If a particular effect presents significant influence on transistor behavior for a certain technology, gate delay modeling that neglect such effect tend to present loss in accuracy.

c) Use fitting parameters – Adding fitting parameters to the method is a way to account for an effect without needing to derive an expression to evaluate it. However, the extraction procedure of these parameters may require extensive electrical simulations. Moreover, there are several ways to include fitting parameters in the delay modeling, becoming difficult to determine the portability of the method to different technology nodes.

d) Do not provide a closed form expression for the delay – Several works must evaluate the output voltage as function of time in order to estimate the delay. Similarly, some approaches present equations for which there are no analytical solutions.

As shown in Table 1, the inverter delay estimation methods proposed in [1]-[4] are not valid for submicrometer technologies because they use long channel transistor model. Several works neglect SCC [5][8][12][21][22][25][26], I/O coupling capacitance [6][7][17][19], or channel length modulation and DIBL [10]-[12][14]-[16][18][20][22][25]. In order to correct inadequate modeling, fitting parameters are often employed [12][15]-[17][21][24]. Furthermore, some approaches do not provide explicit delay formulation for the entire range of input transition time [8][12][13][20][25]. The inverter delay model proposed in this work aims to overcome these drawbacks.

Table 1. Overview of CMOS inverter delay models characteristics.

Work	Slow input transitions	Short channel transistor model	Short circuit current	I/O coupling capacitance	Channel length modulation and DIBL	Only transistor parameters	Closed form delay expression for all cases
[1,2]	No	No	No	No	No	Yes	No
[3]	No	No	Yes	Yes	No	No	Yes
[4]	Yes	No	Yes	Yes	No	Yes	Yes
[5]	No	Yes	No	No	No	Yes	No
[6]	Yes	Yes	Yes	No	Yes	Yes	Yes
[7]	No	Yes	No	No	No	Yes	Yes
[8,12]	Yes	Yes	No	Yes	Yes	Yes	No
[9]	Yes	Yes	Yes	Yes	No	Yes	Yes
[10,19]	Yes	Yes	Yes	Yes	No	No	Yes
[11]	Yes	Yes	Yes	Yes	No	Yes	No
[13]	Yes	Yes	Yes	Yes	Yes	Yes	No
[14]	No	Yes	No	Yes	No	No	Yes
[15,16]	Yes	Yes	No	Yes	No	No	Yes
[17,20]	Yes	Yes	Yes	Yes	No	Yes	Yes
[18]	Yes	Yes	Yes	Yes	No	Yes	Yes
[21]	No	Yes	No	Yes	No	No	No
[22]	Yes	Yes	No	Yes	No	Yes	Yes
[23]	No	No	No	No	No	Yes	No
[24]	Yes	No	Yes	Yes	No	No	Yes
[25]	No	Yes	No	Yes	No	Yes	No
[26]	Yes	Yes	No	No	Yes	Yes	Yes
This work	Yes	Yes	Yes	Yes	Yes	Yes	Yes

III. PRELIMINARIES

The CMOS inverter schematic is shown in Fig. 1, where *Vin* and *Vout* are the input and output voltages, respectively. *Vdd* is the supply voltage. *Cl* represents the sum of the output load and diffusion capacitances of NMOS and PMOS transistors, and *Cm* represents the I/O coupling capacitance.

The coupling capacitance can be divided into two components: the bias independent component (*Cov*) that can be directly obtained from the fabrication process parameters, and the bias dependent component which depends on the transistor operating condition. In CMOS inverter delay analysis, only the gate-to-drain capacitance (*Cgd*) is a concern. Typically, *Cgd* is considered to be half the gate capacitance for a transistor operating in linear region and zero for a device with the channel pinch off or in off state [26]. Hence, the coupling capacitance for a static low input (Cm_{low}) can be written as follows:

$$Cm_{low} = \frac{Cg_p.W_p}{2} + Cov_{const} \quad (1)$$

being W_p the PMOS transistor effective channel width and Cg_p the gate capacitance per meter for a fixed transistor channel length, which can be obtained both from fabrication process parameters and from electrical simulations [27]. Cov_{const} is the sum of the bias independent coupling capacitance of both transistors:

$$Cov_{const} = Cov.(W_n + W_p) \quad (2)$$

where W_n is the NMOS effective channel widths, respectively.

3.1 Delay Definition

Gate propagation delay (*Td*) is given by the difference between the time instants when the output and the input reach half the supply voltage (*Vdd*/2):

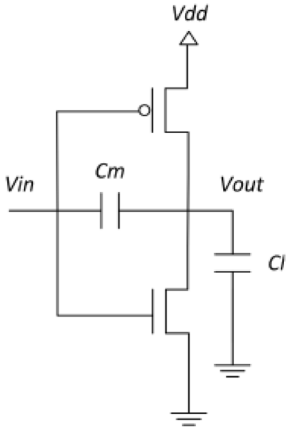


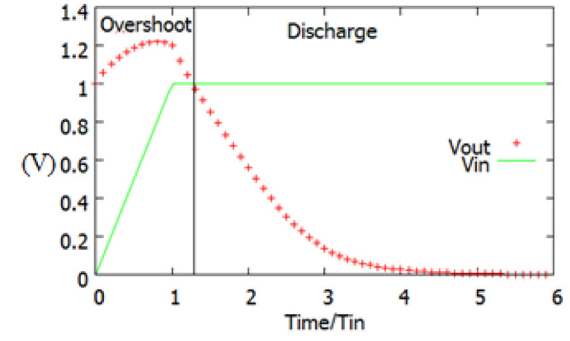
Figure 1. CMOS inverter schematic.

$$Td = Tout50 - Tin/2 \quad (3)$$

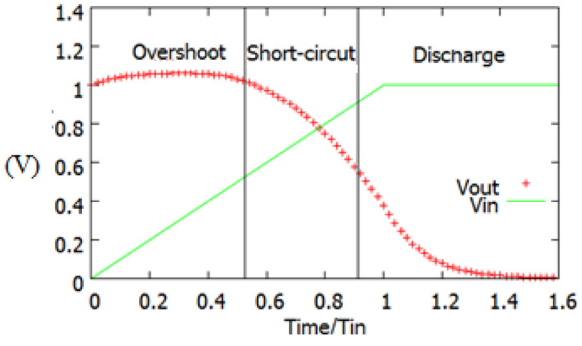
where *Tin* is the input transition time and *Tout50* is the time instant when the output reaches *Vdd*/2. Since *Tin* is either an input to the method or estimated from *Tout50* of the previous gate, prediction accuracy lies, mostly, on estimating *Tout50* [2,10].

3.2 Inverter Transient Behavior

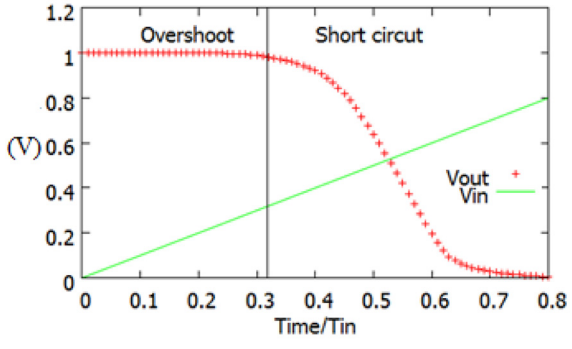
The inverter transient behavior can be divided into three main stages: the overshoot, the short circuit and the discharge [14], as depicted in Fig. 2. Only the overshoot stage is always observable in the output transition. The short circuit (discharge) stage is not identified for sufficiently fast (slow) inputs.



(a)



(b)



(c)

Figure 2. Inverter transient response: (a) no short circuit stage, (b) all three stages, and (c) no discharge stage.

1) Overshoot Stage

During the overshoot stage, the output voltage rises to a value higher than *Vdd* due to the I/O coupling capacitance. Therefore, the PMOS transistor operates in the linear region and under a reverse biasing, while the NMOS transistor enters saturation after the input reaches the NMOS threshold voltage. The reverse bias of the PMOS transistor prevents the existence of a SCC. Actually, the PMOS current tends to discharge the output node although such influence is small.

Even though the maximum overshoot voltage is usually a few percentage of the supply voltage value, it can be significant for fast inputs. Fig. 3 presents the maximum output voltage (normalized to *Vdd*) for different input transition times and PN ratios (W_p/W_n) to a fixed output load and supply voltage of 1.0 V, considering a bulk CMOS 32 nm predictive technology model [32].

2) Short Circuit Stage

After the overshoot, the short circuit stage occurs if the PMOS transistor is still ON. Therefore, this stage can be neglected for sufficiently fast inputs. During the short circuit stage, the current flowing through the NMOS transistor corresponds to the sum of SCC and discharge currents. For this reason, SCC can be seen either as a reduction on the NMOS current capability [9][15] or as an extra charge stored at the output node [7][16].

The influence of the short circuit stage increases if the input transition time rises or the output load decreases. For a sufficiently slow input or small load, the output capacitance is discharged while the input voltage is rising. In such cases, both the *Vgs* and *Vds* of the PMOS device can be large. In contrast, for a sufficiently fast input or large load, the output capacitance discharge is only significant after the input transition.

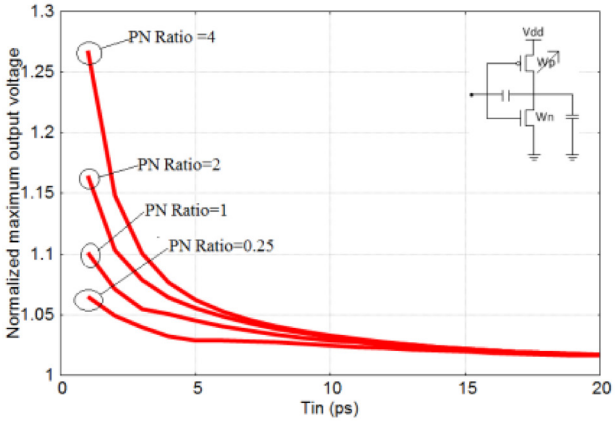


Figure 3. Maximum overshoot voltage for different input transition times.

3) Discharge Stage

Once the PMOS transistor is turned OFF, the entire current capability of the NMOS is used to discharge the output load. Therefore, the maximum discharge ratio is obtained in this stage. Nevertheless, for sufficiently slow input transitions, the output capacitance can discharge while the PMOS transistor conducts.

IV. PROPOSED CMOS INVERTER DELAY MODEL

In the following analysis, a rising ramp input is assumed. The falling input case is symmetrical. *Vin* is described as a function of time *t*, being:

$$Vin=\begin{cases}0 & t<0\\ \frac{Vdd}{Tin}t & 0\le t\le Tin\\ Vdd & t>Tin\end{cases} \quad (4)$$

Even though, in real circuits, signal waveforms are not actually ramps, such an input type is a common approximation [33][34]. Part of the success in using a ramp input is due to the possibility of determining an equivalent ramp for an exponential signal in such a way that the gate behavior is approximately the same for both exponential and ramp inputs [2][10][11][13].

The proposed delay estimation method applies the well-known α-power transistor model [5]:

$$Ids=\begin{cases}0 & OFF\\ KlinW(Vgs-Vth)^{\alpha/2}Vds & Linear\\ KsatW(Vgs-Vth)^{\alpha}(1+\lambda.Vds) & Saturatio\end{cases} \quad (5)$$

where *W* is the effective transistor channel width, *Klin* and *Ksat* are empirical constants, α is the velocity saturation index, λ is the channel length modulation parameters, and *Vgs* and *Vds* are the gate-to-source and drain-to-source voltages, respectively. The threshold voltage (*Vth*) can be expressed as follows:

$$Vth=Vth0-\eta.Vds \quad (6)$$

where *Vth*0 is the threshold voltage when no bias voltage is applied, and η is the DIBL coefficient. Since this work discusses the CMOS inverter behavior, body effect does not have to be considered. Indexes *n* and *p* are used to refer to NMOS and PMOS transistors, respectively.

4.1 Fast Input Domain

An input transition is considered fast if the input signal reaches the final value before the output reaches *Vdd/2*, *i.e.*, *Tin*<*Tout50*. The maximum output voltage (V_{max}) due to the overshoot is given by:

$$V_{max} = Vdd(1+\frac{Cm_{avg}}{Cm_{avg}+Cl}) \quad (7)$$

where *Cl* is the sum of the output capacitance and diffusion capacitances of both NMOS and PMOS transistors, and Cm_{avg} is an average value for the I/O coupling capacitance, given by:

$$Cm_{avg} = \frac{Cm_{low}(Vdd-|Vth0_p|)+Cov_{const}|Vth0_p|}{Vdd} \quad (8)$$

In (7), the average value Cm_{avg} is used instead of Cm_{low} from (1) because the PMOS transistor changes from linear region to OFF state or saturation. In both cases, *Cm* is initially near to Cm_{low} but diminishes as the input arises or the output is discharged. The utilization of $Cm_{avg,}$ instead of $Cm_{low,}$ is one of the main differences of the proposed model to previous related works that consider the coupling capacitance constant to Cm_{low} [13]-[22]. Even though, in many cases, both approaches give similar results, the difference between them can be important for small output loads.

Channel length modulation and DIBL effect are considered by estimating an average *Vds* (Vds_{avg}) for each transistor, as follows:

$$Vds_{avgn} = 0.5(V_{max}+Vdd/2) \quad (9a)$$

$$Vds_{avgp} = Vdd-0.5(V_{max}+Vdd/2) \quad (9b)$$

Tout50 is found by equating the total charge to be removed from the output node (*Qtot*) to the charge drained by NMOS transistor (*Qn*), as follows:

$$Qn = \int_0^{Tin} Id_{rise}.dt + \int_{Tin}^{Tout50} Id_{high}.dt \quad (10)$$

where Id_{rise} and Id_{high} are the NMOS drain-to-source currents when the input is rising and when *Vgs* equals *Vdd*, respectively. *Qtot* is the sum of the charge stored at *Cl* and the extra charge due to *Cm*. *Qtot*, Id_{high} and Id_{rise} can be expressed as:

$$Qtot = Cm_{avg}.Vdd + \frac{Vdd}{2}(Cm_{avg}+Cl) \quad (11)$$

$$Id_{high} = Ksat_n.W_n.(Vdd-Vth_n)^{\alpha_n}(1+\lambda_n.Vds_{avgn}) \quad (12)$$

$$Id_{rise} = Id_{high}.(\frac{\frac{Vdd}{Tin}t-Vth_n}{Vdd-Vth_n})^{\alpha_n} \quad (13)$$

where Vth_n is the NMOS threshold voltage with $Vds=Vds_{avgn}$. Equating (10) and (11), *Tout50* can be written for a fast input as:

$$Tout50 = \frac{Qtot}{Id_{high}} + \frac{Tin.(\alpha_n+\frac{Vth_n}{Vdd})}{\alpha_n+1} \quad (14)$$

The particular case when *Tin*=*Tout50* defines the boundary condition between fast and slow input transition domains. Such a specific input transition time (Tin_{ref}) is used to determine if an input is fast or slow. For any *Tin* smaller or equal to Tin_{ref} the input is fast. Otherwise, the input is slow. Tin_{ref} is given by:

$$Tin_{ref} = \frac{Qtot(\alpha_n+1)}{Id_{high}.(1-\frac{Vth_n}{Vdd})} \quad (15)$$

4.2 Slow Input Domain

The inverter delay modeling in the slow input domain requires additional considerations when compared to the fast input domain. The main challenge is to estimate the discharging current and SCC. The estimation of discharging current is a hard task because the input voltage is still rising when the output reaches *Vdd*/2. The prediction of SCC is also difficult because information about the output waveform is required to accurately estimate this current [28]-[31].

The impact analysis of SCC is essential for accurately estimate the CMOS inverter delay. In this work, SCC is seen as an extra charge to be discharged, as discussed in [7] and in [16].

As already mentioned, the short circuit stage does not occur whether the input transition is fast enough. Indeed, SCC can be neglected if PMOS transistor is already turned off when the overshoot time finishes. During the inverter output signal transition, PMOS transistor enters into the saturation region if the input is slow enough, and it can be considered that the maximum SCC (Isc_{max}) is obtained at this moment [7][16][31]. If the output is fast enough such that PMOS is turned off before entering the saturation, SCC is expected to present small impact on gate delay and can be ignored. In order to predict this current, it is necessary to determine the transistor gate voltage when PMOS saturates and the time interval when the short circuit occurs. The maximum short circuit duration (T_{SC}) can be estimated as follows:

$$Tsc = (Tin-Tin_{ref}).\frac{Vdd-Vth_n-|Vth_p|}{Vdd} \quad (16)$$

where Vth_p is the PMOS threshold voltage with $Vds_p=Vds_{avgp}$. For input transitions close to $Tin_{ref,}$ the value of *Tsc* is small. As the value for *Tin* increases, *Tsc*

approaches a maximum value which is equal to the time interval when both transistors are conducting. The average SCC (*Isc*) is estimated as:

$$Isc = Ksat_p.W_p.Vov_p^{\alpha_p}.(1+\lambda_p.Vds_{avgp})(1-Tin_{ref}/Tin) \quad (17)$$

where Vov_p is an effective overdrive voltage of the PMOS transistor, as follows:

$$Vov_p = (1-Tin_{ref}/Tin).\frac{Vdd-Vth_n-|Vth_p|}{2} \quad (18)$$

In both (17) and (18), the term Tin_{ref}/Tin reduces the *Isc* value for *Tin* values close to Tin_{ref}. The short circuit charge *Qsc* is simply the average current multiplied by the short circuit time:

$$Qsc = Tsc.Isc \quad (19)$$

Another important difference between fast and slow input domains is that, in the latter, only a fraction of the extra charge due to I/O coupling capacitance is transferred since *Vin* only reaches *Vdd* after *Tout50*. For this reason, a correction is proposed for this component in the slow input domain, as follows:

$$Qcm_{slow} = Vdd.Cm_{avg}.(\frac{Tin_{ref}}{Tin})^{1/(1+\alpha_n)} \quad (20)$$

Therefore, the total charge to be removed through NMOS device is given by:

$$Qtot = Qsc + Qcm_{slow} + Vdd\frac{Cm_{avg}+Cl}{2} \quad (21)$$

To estimate the discharge time for a slow input, it must be noticed that the current capacity of NMOS transistor does not reach the maximum value because the input is still rising when *Vout* reaches *Vdd/2*. Therefore, a different approach for fast input domain has to be applied.

In this work, it is assumed that the maximum NMOS current capacity during the output voltage swing to *Vdd*/2 is observed when the output reaches such a voltage level. That is a reasonable assumption because the NMOS transistor operates in saturation region. *Tout50* is found by calculating the total charge drained by NMOS transistor (*Qn*), from the beginning of input transition until *Tout50*:

$$Qn = \int_{Tvth_n}^{Tout50} Ids_{rise}.dt \quad (22)$$

where $Tvth_n$ is the time instant when the input reaches Vth_n. Solving (21), *Qn* is found as:

$$Qn = \frac{Id_{high}.Tin.(\frac{Vdd.Tout50}{Tin}-Vth_n)^{\alpha_n+1}}{Vdd.(Vdd-Vth_n)^{\alpha_n}(\alpha_n+1)} \quad (23)$$

Defining Δt as the time elapsed between $Tvth_n$ and *Tout50* (*i.e.*, Δt = *Tout50* - $Tvth_n$), and knowing that $Vth_n = (Tvth_n.Vdd)/Tin$, (23) can be written as follows:

$$Qn = \frac{Id_{high}.Vdd^{\alpha_n}.\Delta t^{\alpha_{n+1}}}{Tin^{\alpha_n}.(Vdd-Vth_n)^{\alpha_n}(\alpha_n+1)} \quad (24)$$

Since *Qtot* is equal to *Qn*, from (20) and (23), it follows:

$$\Delta t = (\frac{(\alpha_n+1).Tin^{\alpha_n}.Qtot(Vdd-Vth_n)^{\alpha_n}}{Id_{high}.Vdd^{\alpha_n}})^{1/(\alpha_n+1)} \quad (25)$$

Thus, the final value of *Tout50* in the slow input domain is given by:

$$Tout50 = (\frac{(\alpha_n+1)Tin^{\alpha_n}.Qtot(Vdd-Vth_n)^{\alpha_n}}{Id_{high}.Vdd^{\alpha_n}})^{\frac{1}{\alpha_n+1}} + Tin\frac{Vth_n}{Vdd} \quad (26)$$

V. SIMULATION RESULTS

The proposed method was validated using a 32 nm CMOS predictive transistor model (PTM32) for high performance (HP) [32] and a commercial 65 nm CMOS (C65) technology. For both technologies the transistor model is the BSIM4. Moreover, both single inverters and inverter chains are evaluated.

5.1 Single Inverter Evaluation

The proposed approach was compared to data extracted from HSPICE electrical simulations, based on BSIM4 transistor model, and to the state-of-the-art CMOS inverter delay models. In the first experiment performed, the PTM32 process parameters were applies. The PN ratio was varied from 0.25 to 8. The NMOS channel width value was fixed to 256 nm whereas PMOS width was modified to obtain a specific PN ratio. The channel length of both NMOS and PMOS transistors was kept constant and equal to 32 nm. The output load was normalized to the gate capacitance of NMOS device, remaining unchanged for a given technology since only PMOS transistor size is modified. The normalized output load varies from 0.25 to 64. Therefore, an output load equal to one is equivalent to the gate capacitance of a NMOS device. For each pair of PN ratio and output load, five hundred *Tin* values were applied. The input transition time was varied from 1 to 500 ps, by a step of 1 ps.

Table 2 presents the average (*AVG*) and the worst case (*WC*) relative errors obtained by applying the model proposed herein. Notice that, at this experiment, accuracy is being evaluated by considering the estimation of *Tou50*, instead of delay itself. This choice is made because the inverter delay tends to decrease for sufficiently large values of *Tin*, and may become zero. In this situation, any error becomes a large relative error even though the absolute error is insignificant. On the other hand, *Tout50* always increases with *Tin*, being a more appropiate metric for the single inverter analysis.

Table 3, Table 4 and Table 5 present the same data (also considering *Tout50* estimation) for the approaches proposed by Rossello and Segura [10], by Wang and Zwolinski [12] and by Huang *et al.* [22], respectively. Notice that two recent proposals presented in [25] and in [26] are not directly evaluated because these methods also neglect SCC, showing to similar errors as those observed in [12]. Consoli's approach, presented in [13] is also not directly evaluated since we found this model to be as accurate as Rossello's model [10]. The proposed model presents in the worst

Table 2. Relative error of proposed delay model, considering PTM32 parameters and *Vdd* equals to 1.0 V.

W_p/W_n	Normalized Output Load							
	1/4		1		4		64	
	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)
¼	0.20	1.90	1.50	1.90	2.80	3.50	2.90	3.60
½	2.70	3.30	1.20	1.60	1.00	1.75	1.95	2.00
1	2.35	3.00	2.30	3.25	2.55	2.60	1.50	1.65
2	2.20	2.40	1.05	1.75	1.15	1.70	2.25	3.00
4	2.05	2.80	1.25	2.30	0.75	2.40	2.00	3.10
8	2.25	4.25	2.60	4.40	3.50	5.90	4.10	7.30

Table 3. Relative error of delay model presented in [10], considering PTM32 parameters and *Vdd* equals to 1.0 V.

W_p/W_n	Normalized Output Load							
	1/4		1		4		64	
	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)
¼	1.75	11.5	2.15	11.5	3.10	11.0	4.00	9.80
½	2.40	9.10	2.55	9.45	2.90	9.85	3.20	9.60
1	2.50	5.10	2.40	6.10	2.25	7.80	2.20	8.65
2	2.60	5.30	2.40	4.50	1.90	4.40	1.10	7.30
4	2.90	8.00	2.70	7.35	2.20	5.50	1.50	3.90
8	3.50	13.5	3.45	13.5	3.00	9.10	2.40	5.75

Table 4. Relative error of delay model presented in [11], considering PTM32 parameters and *Vdd* equals to 1.0 V.

W_p/W_n	Normalized Output Load							
	1/4		1		4		64	
	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)
¼	4.20	10.0	4.05	9.90	3.75	9.60	3.50	7.60
½	6.00	15.5	5.50	15.0	4.90	14.5	4.05	10.0
1	7.50	20.0	7.20	19.5	6.75	19.0	4.70	13.0
2	8.30	22.5	8.00	22.0	7.75	22.0	5.55	16.0
4	8.40	23.5	8.15	23.5	7.50	23.0	6.00	18.0
8	8.00	24.0	7.75	24.0	7.30	23.5	6.20	18.5

Table 5. Relative error of delay model presented in [21], considering PTM32 parameters and *Vdd* equals to 1.0 V.

W_p/W_n	Normalized Output Load							
	1/4		1		4		64	
	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)
¼	3.75	20.0	5.10	20.0	7.20	13.0	5.00	10.0
½	6.15	23.0	5.65	19.0	4.33	13.0	4.37	11.0
1	8.00	16.0	6.00	21.0	4.50	18.0	4.15	15.0
2	9.00	29.0	9.00	27.0	8.00	26.0	9.00	22.0
4	12.0	34.0	12.0	35.0	11.0	36.0	10.0	37.0
8	11.0	37.0	12.0	38.0	11.5	39.0	11.0	39.0

case an error of 7.4%, whereas Rossello's [10], Wang's [12] and Huang's [22] approaches provide the worst case errors about 13%, 24% and 39%, respectively. Moreover, the average error of the proposed model is 2%, whereas the average errors provided in [10], [12] and [22] are approximately 3%, 6% and 7%, respectively.

Fig. 4 compares the probability density function (PDF) of the relative error for each model evaluated. $Tout50_{model}$ and $Tout50_{sim}$ stand for the *Tout50* values obtained from the proposed and from HSPICE electrical simulations. The error distribution is assumed to be normal which leads to a folded normal distribution. It is worth to notice that the proposed model has a smaller standard deviation when compared to related works. Even though the average error of Rossello's approach [10] is similar to the proposed model, the higher standard deviation results in a much larger worst case error, as previously stated. Hereafter, we consider Rossello's work as reference because it provides more accurate results than others related works.

Fig. 5 shows the relative errors as function of *Tin* for the proposed method and Rossello's approach, for two typical conditions: PN ratios of 1 and 2 with fanout approximately four. It is clear that the proposed model is more accurate for the majority of values of *Tin*. Moreover, the error of the proposed method appears to saturate as *Tin* grows, whereas Rossello's approach does not show the same behavior.

The proposed method was also evaluated considering a commercial 65 nm CMOS technology (C65). Table 6 and Table 7 present the average (*AVG*)

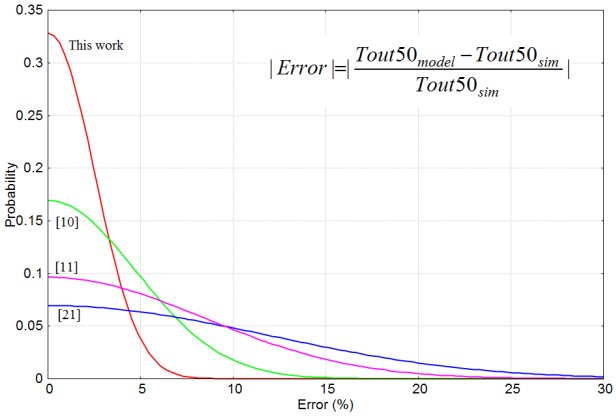


Figure 4. Probability density function for the error of each delay model evaluated.

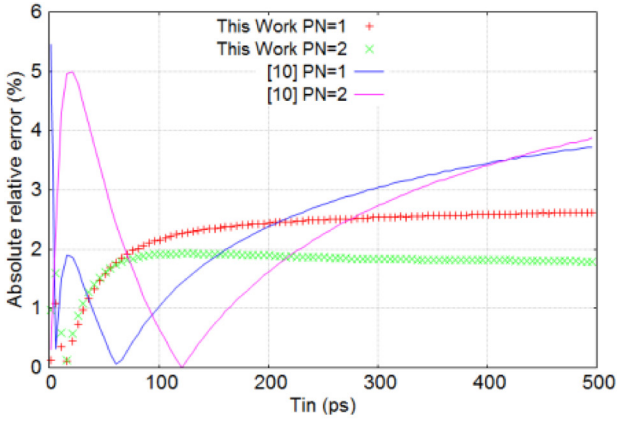


Figure 5. Comparison of the proposed method to the Rossello's approach, presented in [10], considering Tin variation.

Table 6. Relative error of proposed CMOS inverter delay method, considering C65 parameters and *Vdd* equals to 1.2 V.

W_p/W_n	Normalized Output Load							
	1/4		1		4		64	
	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)
¼	2.20	4.40	2.19	4.10	1.07	2.45	0.50	1.65
½	0.60	3.20	0.72	2.70	0.50	1.55	0.70	1.30
1	1.80	2.80	1.20	2.40	0.85	1.50	0.80	1.15
2	2.25	4.20	1.77	3.05	1.15	2.10	0.80	1.25
4	1.15	5.70	0.95	4.50	0.50	2.80	0.30	1.70
8	2.30	6.00	2.30	6.10	2.10	4.20	1.62	2.70

Table 7. Relative error of proposed CMOS inverter delay method, considering C65 parameters and *Vdd* equals to 0.96 V.

W_p/W_n	Normalized Output Load							
	1/4		1		4		64	
	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)	AVG (%)	WC (%)
¼	1.36	4.30	2.90	5.70	2.65	3.40	1.85	2.50
½	2.50	4.40	2.75	6.11	1.95	3.60	1.95	2.60
1	1.80	4.50	2.75	6.60	1.15	6.60	2.45	4.00
2	1.20	4.50	3.65	7.00	2.00	4.60	2.30	3.10
4	3.30	4.30	2.00	6.45	2.40	6.40	2.60	4.60
8	2.90	5.00	3.48	7.10	2.25	6.50	2.50	6.70

and the worst case (*WC*) relative errors for different single inverter configurations for C65 process parameters for *Vdd* equals to 1.2 V and 0.96 V, respectively. The simulation conditions are similar to those used in the PTM32 simulations, with the proper adjustments to take into account different design rules.

5.1 Inverter Chain Evaluation

In order to estimate the delay of an inverter chain, the effective output transition time, which determines the input for the next stage, must be estimate. Two common ways to obtain the effective slope of the output signal are: (a) the time required for the output voltage swing between two voltage values (*e.g.* 10% and 90% of *Vdd*) [5][7][13]; and (b) a percentage of the output derivative when the output reaches *Vdd/2* [2][10]. The second strategy was applied in this work.

Instead of the typical 70% of the output derivative, we chose the percentage considering the relationship between input and output transitions, similarly to [10]. In this work, the effective output transition time ($Tout_{eff}$) is given by:

$$Tout_{eff} = \frac{Vdd.(Cl + Cm_{avg})}{In_{50}.(1 - 0.3\frac{Tin_{ref}}{Tin})} \quad (27)$$

For a falling output, In_{50} is the NMOS current at *Tout50,* given by:

$$In_{50} = Id_{high}.(\frac{Vdd.Tout50/Tin - Vth_n}{Vdd - Vth_n})^{\alpha_n} \quad (28)$$

Two sets of 10-stage inverter chains with different configurations were evaluated, each containing 10,000 chains. The first set assumes usual inverter configurations whereas the second set is less restrict. In all cases, PTM32 parameters were used, with *Vdd*= 1.0 V. For the first set, the PN ratio of each stage is a random number between 1 and 2, while the maximum fanout is approximately 4. A comparison between the proposed method and Rossello's approach is given in Fig. 6. Clearly, the proposed model is more accurate, having a worst case error near to 3%.

For the second set of simulations, the PN ratio of each stage varied from 0.25 to 8, while the maximum fanout was approximately 32. A comparison between both methods is depicted in Fig. 7. Considering the proposed method, 99% of the cases present an error equal to or smaller than 7%, whereas only 37% of the cases considering the Rossello's method lie in the same range.

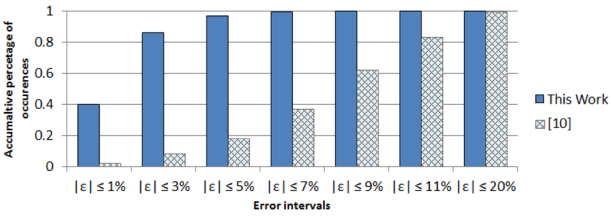


Figure 6. Comparison between the proposed method and Rossello's approach [10] when evaluating different 10-stages inverter chain circuits considering typical inverter configurations.

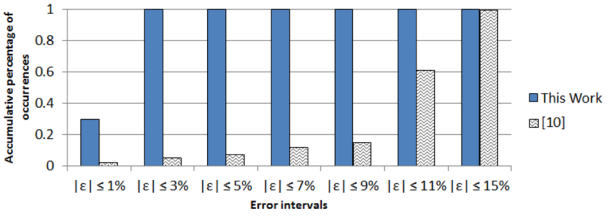


Figure 7. Comparison between the proposed method and Rossello's approach [10] when evaluating different 10-stages inverter chain circuits.

VI. CONCLUSIONS

A novel and more accurate delay model for CMOS inverter was proposed. The main advantage of this approach in comparison to previous related work is better inverter delay prediction due to more appropiate modeling of the most relevant second-order physical effects, for instance the influence of channel length modulation and DIBL, usually neglected by similar works. Furthermore, the proposed method improves the modeling of I/O coupling capacitance for delay estimation. Simulation data was obtained for single inverters and inverter chains with different configurations, taking into account different technology parameters. Results have shown an average error about 3%, and the worst case error smaller than 10%. Even though not discussed in this work, the proposed delay estimation method can be easily extended to consider the effects of variability on transistor performance since it relies solely on transistor parameters.

ACKNOWLEDGEMENTS

Research funded by the Brazilian funding agencies CAPES, CNPq and FAPERGS, under grant 11/2053-9 (Pronem).

REFERENCES

[1] J. R. Burns, "Switching response of complementary symmetry MOS transistor logic circuits," *RCA Review*, vol. 25, 1964, pp.627-661.

[2] N. Hedenstierna and K. O. Jeppson, "CMOS circuit speed and buffer optimization," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 6, no. 2, Mar. 1987, pp.270-281

[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. Solid-State Circuits,* vol. 29, no. 6, June 1994, pp.646-654.

[4] L. Bisdounis, S. Nikolaidis, and O. Loufopavlou, "Propagation delay and short-circuit power dissipation modeling of the CMOS inverter," *IEEE Trans. on Circuits and Systems I, Fundamental Theory and Application,* vol. 45, no. 3, Mar. 1998, pp.259-270.

[5] 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, no. 2, Apr. 1990, pp.584-594.

[6] T. Sakurai and A. R. Newton, "A simple MOSFET model for circuit analysis," *IEEE Trans. on Electron Devices,* vol. 38, no. 4, Apr. 1991, pp.887-894.

[7] N. Chandra, A. K. Yati, and A. B. Bhattacharyya, "Extended-Sakurai-Newton MOSFET model for ultra-deep-submicrometer CMOS digital design," in Proc. of *Int'l Conf. on VLSI Design*, 2009, pp.247-252.

[8] P. Cocchini, G. Piccinini, and M. Zamboni, "A comprehensive submicrometer MOST delay model and its application to CMOS buffers," *IEEE J. Solid-State Circuits,* vol. 32, no. 8, Aug. 1997, pp.1254-1262.

[9] 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, no. 2, Feb. 1998, pp.302-306.

[10] J. L. Rossello and J. Segura, "An analytical charge-based compact delay model for submicrometer CMOS inverters," *IEEE Trans. Circuits and Systems I, Regular Papers,* vol. 51, no. 7, July 2004, pp. 1301-1311.

[11] A. Chatzigeorgiou and S. Nikolaidis, "Efficient output waveform evaluation of a CMOS inverter based on short-circuit current prediction," *Int'l Journal of Circuit Theory and Applications,* vol. 30, no. 5, Sep./Oct. 2002, pp.547–566.

[12] Y. Wang and M. Zwolinski, "Analytical transient response and propagation delay model for nanoscale CMOS inverter," in Proc. of *Int'l Symp. on Circuits and Systems* (ISCAS), 2009, pp.2998-3001

[13] E. Consoli, G. Giustolisi, and G. Palumbo, "An accurate ultra-compact I-V model for nanometer MOS transistors with applications on digital circuits," *IEEE Trans. on Circuits and Systems I: Regular Papers*, vol. 59, no. 1, Jan. 2012, pp.159-169.

[14] D. Deschacht, M. Robert, and D. Auvergne, "Explicit formulation of delays in CMOS data paths," *IEEE J. Solid-State Circuits,* vol. 23, no. 5, Oct. 1988, pp.1257-1264.

[15] D. Auvergne, N. Azemard, D. Deschacht, and M. Robert, "Input waveform slope effects in CMOS delays," *IEEE J. Solid-State Circuits,* vol. 25, no. 6, Dec. 1990, pp.1588-1590.

[16] J. M. Daga and D. Auvergne, "A comprehensive delay macro modeling for submicrometer CMOS logics," *IEEE J. Solid-State Circuits,* vol. 34, no. 1, Jan. 1999, pp.42-55.

[17] S. H. K Embabi and R. Damodaran, "Delay models for CMOS, BiCMOS and BiNMOS circuits and their applications for timing simulations," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems,* vol. 13, no. 9, Sep. 1994, pp.1132-1142.

[18] A. A. Hamoui and N. C. Rumin, "An analytical model for current, delay, and power analysis of submicron CMOS logic circuits," *IEEE Trans. on Circuits and Systems II, Analog and Digital Signal Processing,* vol. 47, no. 10, Oct. 2000, pp.999-1007.

[19] S. Dutta, S. S. M. Shetti, and S.L. Lusky, "A comprehensive delay model for CMOS inverters," *IEEE J. Solid-State Circuits,* vol. 30, no. 8, Aug. 1995, pp.864-871.

[20] A. Kabbani, D. AlKhalili, and A. J. Al-Khalili, "Technology portable analytical model for DSM CMOS inverter delay estimation," *IEE Proc. Circuits, Devices and Systems,* vol. 152, no. 5, Oct. 2005, pp. 433-440.

[21] C. C. Wang and D. Markovic, "Delay estimation and sizing of CMOS logic using logical effort with slope correction," *IEEE Trans. on Circuits and Systems II, Express Briefs,* vol. 56, no. 8, Aug. 2009, pp.634-638.

[22] Z. Huang, A. Kurokawa, M. Hashimoto, T. Sato, J. Minglu, and Y. Inoue, "Modeling the overshooting effect for CMOS inverter delay analysis in nanometer technologies," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems,* vol. 29, no. 2, Feb. 2010, pp.250-260.

[23] W. C. Elmore, "The transient response of damped linear networks with particular regard to wideband amplifiers," *J. Applied Physics,* vol. 10, no. 1, Jan. 1948, pp 55-63.

[24] L. F. Uebel and S. Bampi, "A timing analysis tool for VLSI CMOS synchronous circuits," in Proc. of *Int'l Symp. on Circuits and Systems* (ISCAS), 1996, pp.516-519.

[25] M. Mehri, K.H. Kouhani, N. Masoumi, and R. Sarvari, "New approach to VLSI buffer modeling, considering overshooting effect," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 8, Aug. 2013, pp.1568-1572.

[26] N. Alam, B. Anand, and S. Dasgupta, "An analytical delay model for mechanical stress induced systematic variability analysis in nanoscale circuit design," *IEEE Trans. on Circuits and Systems I, Regular Papers*, vol. 61, no. 6, June 2014, pp.1714-1726.

[27] N. Weste and D. Harris, CMOS VLSI Design, 4th edition, Boston: Addison-Wesley, 2010.

[28] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, vol. 19, no. 4, Aug. 1984, pp.468-473.

[29] S. Turgis and D. Auvergne, "A novel macromodel for power estimation in CMOS structures," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 11, Nov. 1998, pp.1090-1098.

[30] J. L. Rossello and J. Segura, "Charge-based analytical model for the evaluation of power consumption in submicron CMOS buffers," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 4, Apr 2002, pp.433-448.

[31] C. C. Liu, J. Chang, and L.G. Johnson, "Energy model of CMOS gates using a piecewise linear model," in Proc. of *Int'l Symp. on Circuits and Systems* (ISCAS), 2010, pp.3829-3832.

[32] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm early design exploration," *IEEE Trans. on Electron Devices*, vol. 53, no. 11, Nov. 2006, pp.2816-2823. Available online at htpp://ptm.asu.edu.

[33] A. Hamid, "Automated cell characterization system," *U.S. Patent* US5655109 A Aug 1997.

[34] K. Tseng and K. Chou, “Systems and methods of efficient library characterization for integrated circuit cell libraries,” *U.S. Patent* US20110087478 A1 April 2011.

[35] B. S. Cherkauer and E. G. Friedman, “A unified design methodology for CMOS tapered buffers,” *IEEE Trans. on Very Large Scale Integration (VLSI) Systems,* vol. 3, no. 1, Mar. 1995; pp.99-111.

[36] V. Adler and E. G. Friedman, “Repeater design to reduce delay and power in resistive interconnect,” *IEEE Trans. on Circuits and Systems II, Analog and Digital Signal Processing*, vol. 45, no. 5, May 1998, pp.607-616.

[37] F. Frustaci, M. Alioto, and P. Corsonello, “Tapered-Vth approach for energy-efficient CMOS buffers,” *IEEE Trans. on Circuits and Systems I, Regular Papers*; vol. 58, no. 11, Nov. 2011, pp.2698-2707.

[38] A. Benfdila and F. Balestra, “On the drain current saturation in short channel MOSFETs,” *Microelectronics Journal,* vol. 37, no. 7, July 2006; pp.635-64