Variation Aware Timing Model of CMOS Inverter for an Efficient ECSM Characterization

Lomash Chandra Acharya¹, Arvind Kumar Sharma², Venkatraman Ramakrishnan³, Ajoy Mandal³, Sudeb Dasgupta¹, Anand Bulusu¹

¹Indian Institute of Technology, Roorkee, India, ²University of Minnesota, USA, ³Texas Instrument, Bengaluru, India
¹E-mail: anand.bulusu@ece.iitr.ac.in

Abstract
In static timing analysis (STA), delay estimation of CMOS standard cells is accomplished by the effective current source model (ECSM) characterization method. ECSM characterization stores the values of threshold crossing points (TCP’s) of output voltage in a look-up table for the combination of input transition time (T_R) and load capacitance (C_L). Due to variability, re-characterization of entire cell library is required to update the look-up tables of TCP’s, which is a tedious task. Due to temporal variability mechanisms such as negative bias temperature instability (NBTI) and hot carrier injection (HCI), non-critical paths may become critical. Therefore, a large coverage of T_R is required. In this work, an analytical timing model for delay estimation of CMOS inverter with large T_R coverage is proposed. This model matches well with HSPICE simulation with a maximum error of 3%. With this timing model, the ECSM characterization of CMOS inverter can be accomplished with very few HSPICE simulations compared to traditional approaches. The model coefficients are correlated with device/layout level parameters such as driving strength (W) or number of layout fingers (NF), supply voltage (V_dd), threshold voltage (V_th), and mobility (µ). Furthermore, with this model, re-characterization efforts can be reduced significantly in comparison to conventional approach and variability analysis could be done without exercising Monte-Carlo analysis.

Keywords
CMOS, Effective Current Source Model (ECSM), hot carrier injection (HCI), negative bias temperature instability (NBTI), STA

1. Introduction

With the miniaturization of CMOS technology (i.e., 65 nm) process in the nanometer regime, the logic gate delay is adversely affected by many variability issues. Due to variability, the small variation of delay in a logic gate results in a significant change in the total delay of a data path. In this context, an accurate timing model is required for standard cell characterization. Non-Linear delay model based conventional standard cell characterization approach does not provide an accurate estimation of delay information since it captures only a single value delay for a given C_L and T_R combination. This single value delay information is not sufficient for the amalgamation of the entire output node transition. The linear interpolation is used for estimation of delay at other C_L-T_R combinations, which is quite inaccurate below 65 nm technology node [1]. To address this issue, current source models (CSM) have been employed in the standard cell characterization process. Since the CSMs parameters are independent of waveform and type of load, these models ideally support arbitrary input waveform and load capacitance combination.EDA community employs two forms of CSMs: Effective CSM and composite CSM, which can mutually be derived from one another.

The ECSM stores the predetermined x% TCPs of output voltage transition in a LUT for a combination of input slew (T_D) and C_L while CCSM captures the output current for different TCPs [2]. Due to PVT variation, delay of standard cells changes, so re-characterization of a standard cell is required. Since characterization is an extensive computational task, for the accurate estimation of delay requires analytical timing models, which can also reduce the number of HSPICE simulations to simplify the re-characterization task.

(a) Representation of TCPs (b) LUT for captured TCPs

Figure 1: An overview of ECSM Characterization

Researchers have proposed different timing models for delay estimation of standard cells such as CMOS inverter, buffer, NAND, and NOR gate. Some of them do not consider the variation in T_R [3]. Some of them consider T_R variation, but they require their extraction separately for different standard cells. Out of proposed timing models, most models are derived based on the current model of the transistor, which does not have authenticity as we scale the technology node [4]-[6]. Due to this, these models are not extensively used in the modern characterization process. In [7], an analytical model is with an assumption that the value of output voltage discharge at T_R is greater than the value of output voltage discharge when it is equal to drain to source saturation voltage (V_out (T_R) > V_dd)) [Figure 1(a)]. The region of validity of this model is also proposed. This model covers relatively large T_R that are typical of an optimized critical path. However, due to reliability mechanisms such as NBTI and HCI, when a circuit is operated at several V_dd, non/near-critical paths may become critical. Therefore, there is a need to cover a large range of T_R in characterization,
and a new analytical timing model of CMOS inverter is required [8]. To extend the supported range of $T_R$ in [7], we propose a new analytical timing model of CMOS inverter which represents TCPs of output node transition as a function of $T_R$ and $C_{out}$ (a combination of $C_L$ and parasitic capacitance ($C_p$) looking from the output side). We derive relationships of model coefficients with $W$, $V_{dd}$, $V_{th}$, and $\mu$. The application of this novel timing model is the coverage of a wide $T_R$-$C_L$ range, which becomes critical in case of delay degradation caused by temporal variability mechanisms, reduction in re-characterization efforts, and variability analysis without exercising the Monte-Carlo analysis.

The paper is organized in the following manner: In section 2 we derive a newly proposed CMOS inverter timing models. In section 3 we validate the proposed timing model and verify that the model coefficient with device-level parameters is accomplished with HSPICE simulations. Then in section 4 we describe how this model can be used to reduce characterization efforts by the extraction of LUT and delay estimation of an inverter in a random-sized inverter chain.

2. CMOS Inverter Timing Model

To cover a larger range of $T_R$ as compared to [7], a timing model is derived with the assumption that $V_{dd} \geq V_{out}(T_R)$. Here, $V_{dd}$ is the drain to source saturation voltage at $t = T_R$, and $V_{out}(T_R)$ is the value of output voltage at time $t = T_R$. In the derivation of this timing model of CMOS inverter $T_{TCPs}$ (considering high to low transition of the output) are classified into three regions:

![Figure 2: Transient Response of CMOS Inverter](image)

- **Region I**: When $T_{TCP} \leq T_{sat}$
- **Region II**: When $T_{sat} < T_{TCP} \leq T_R$
- **Region III**: When $T_{TCP} > T_R$

Here $T_{sat}$ is the time when output voltage discharge ($V_{TCP}$) reaches $V_{dd}$, $T_R$ is the time when the input reaches $V_{dd}$, and $T_{TCP}$ is the time of $V_{TCP}$ that we want to calculate.

### A. Derivation of $T_{TCPs}$ in Region I

$T_{TCPs}$ within the region, $V_{out} = V_{dd}$ to $V_{out} = V_{dsat}$ lies in this region. The input to the gate terminal (gate to source signal) of CMOS inverter is assumed to be a piecewise linear (PWM) type rising ramp with slew $T_R$ is defined as:

$$V_{in}(t) = \frac{V_{dd}}{T_R} \times t \quad \text{for} \quad 0 \leq t \leq T_R$$

$$= V_{dd} \quad \text{for} \quad t > T_R$$

(1)

The value of drain current in saturation and the linear region is obtained using the alpha power model [4]:

$$I_{dsat} = W_n C_{OX} V_{sat} (1 + \lambda V_{dsat}) (V_{gs} - V_{th} - V_{dsat})$$

(2)

$$I_{linear} = \mu_n C_{OX} \left( \frac{W}{L} \right) (V_{gs} - V_{th}) V_{ds}$$

(3)

Here $V_{sat}$ and $(1 + \lambda V_{dsat})$ represent saturation velocity, and drain induced barrier lowering (DIBL) factor respectively.

To derive the expression for $T_{TCP}$ NMOS saturation current is integrated from 0 to $T_{sat}$ and equated to the total output discharge ($V_{dsat} - V_{dd}$) ($C_L + C_p$), where $C_L + C_p$ represents the total load capacitance $C_{out}$ looking from the output side.

$$\int (C_L + C_p) (V_{dsat} - V_{dd}) dt = \int \frac{WC_{OX} V_{sat} (1 + \lambda V_{dsat}) (V_{gs} (t) - V_{th} - V_{dsat})}{T_R} dt$$

(4)

$$\int (C_L + C_p) (V_{dsat} - V_{dd}) dt = \left[ \frac{V_{dsat} t^2}{2} - (V_{th} + V_{dsat}) t \right]_0^T$$

(5)

$$\int (C_L + C_p) (V_{dsat} - V_{dd}) dt = \frac{V_{dsat} T_R^2}{2} - (V_{th} + V_{dsat}) T_R$$

(6)

Solving (7) for $T_{sat}$ we get:

$$T_{sat} = K_1 T_R + \sqrt{K_1^2 T_R^2 + K_2 T_R^2 (C_L + C_p)}$$

(8)

Where

$$K_1 = \frac{(V_{th} + V_{dsat})}{V_{dsat}}$$

(9)

$$K_2 = \frac{2 (V_{dsat} - V_{dd})}{WC_{OX} V_{sat} (1 + \lambda V_{dsat})}$$

(10)

Here $K_1$ and $K_2$ are model coefficients, and values of these coefficients is extracted by HSPICE simulation. The observations from the derivation of (8) are as follow:

- **Observation1**: $K_1$ is independent of $W_n$.
- **Observation2**: $K_1$ increases linearly with $1/W_n$.
- **Observation3**: $K_1$ increases linearly with $V_{th}$ variation.
- **Observation4**: $K_2$ increases linearly with $V_{th}$ variation.
- **Observation5**: $K_1$ is independent of $\mu$ variation.
- **Observation6**: $K_2$ increases linearly with $\mu$ variation.

### B. Derivation of $T_{TCPs}$ in Region II

In this region, when output node voltage crosses $V_{dsat}$, NMOS begins to operate in the linear region. The rate of change of drain current for a value of $V_{out}$ is almost proportional to $T_R$. This is because of the following reason: The output discharge in Region II is linear with a time-constant RC [9], where $R$ is the drain to source resistance of the discharging NMOS device. If $T_R$ is increased, the average value of $R$ also reduces at almost an equal rate (i.e., slower by the same factor as the increase in $T_R$). Therefore, the arrival of $T_{TCP}$ with respect to $T_{sat}$ increases by an equal factor as $T_R$, and the average current between $T_{sat}$ and a $T_{TCP}$ remains constant with $T_R$. Therefore, the drain current in this region can be expressed as:

$$\frac{I_{dsat}}{I_{linear}} = \alpha_{TCP} (\text{const})$$

(11)
\[ I_{\text{linear}} = \frac{I_{\text{out}}}{\alpha_{TCP}} \quad (12) \]

We have verified (11) by HSPICE simulation, and is shown in figure 3.

![Figure 3: Variation of current ratio with \( T_R \)](image)

To calculate the \( T_{TCP} \)'s in this region, the NMOS drain current is integrated from \( T_{out} \) to \( T_{TCP} \) and equated to the total output discharge \( (V_{out} - V_{T}) \) \( (C_L + C_p) \).

\[ (C_L + C_p)(V_{out} - V_{TCP}) = \int_{T_{out}}^{T_{TCP}} I_{\text{out}} \, dt \quad (13) \]

Substituting the value from (2) and (12) into (16):

\[ (C_L + C_p)(V_{out} - V_{TCP}) = \frac{\alpha_{TCP}}{wC_{ot}V_{out}(1 + \beta_{\alpha_{TCP}})} \int_{T_{out}}^{T_{TCP}} (V_{out} - V_{Th} - V_{out}) \, dt \quad (18) \]

Substituting the value from (1) into (18): \[ \alpha_{TCP}(C_L + C_p)(V_{out} - V_{TCP}) = \frac{\alpha_{TCP}}{wC_{ot}V_{out}(1 + \beta_{\alpha_{TCP}})} \int_{T_{out}}^{T_{TCP}} (V_{out} - V_{Th} - V_{out}) \, dt \quad (19) \]

Solving (21) for \( T_{TCP} \), we get:

\[ T_{TCP} = k_s T_R + \left( K_3^2 T_R^2 + K_4 T_R \right) \quad (22) \]

Where \[ K_s = \frac{(V_{Th} + V_{out})}{V_{sd}} \quad (23) \]

\[ K_3 = \frac{2\alpha_{TCP}(V_{sd} - V_{TCP})}{wC_{ot}V_{out}(1 + \beta_{\alpha_{TCP}})} \quad (24) \]

Here \( K_3 \) and \( K_4 \) are model coefficients, and values of these coefficients are extracted by HSPICE simulation. The observations from the derivation of (24) are as follow:

- Observation7: \( K_3 \) is independent of \( W_n \).
- Observation8: \( K_3 \) increases linearly with \( 1/W_n \).
- Observation9: \( K_4 \) increases linearly with \( V_{th} \) variation.
- Observation10: \( K_3 \) increases linearly with \( V_{th} \) variation.
- Observation11: \( K_4 \) is independent of \( \mu \) variation.
- Observation12: \( K_4 \) increases linearly with \( \mu \) variation.

C. Derivation of \( T_{TCP} \)s in Region III

In this region, when \( T_{TCP} > T_R \), the input reaches \( V_{dis} \) and the NMOS starts operating in the linear region. To derive the expression for \( T_{TCP} \) first, we calculate the value of \( \frac{V_{out}}{R_{NMOS}} \) then equate, output current \( V_{out}/R_{NMOS} \) to output discharge current \( (C_L + C_p) \left( \frac{dV_{out}}{dt} \right) \).

\[ (C_L + C_p) \frac{dV_{out}}{dt} = \frac{V_{out}}{R_{NMOS}} \quad (25) \]

\[ (C_L + C_p) \int_{T_{out}}^{T_{TCP}} \frac{dV_{out}}{dt} = -\mu C_{ox} \left( \frac{W}{L} \right) \int_{T_{out}}^{T_{TCP}} \left( V_{sd} - V_{th} \right) \, dt \quad (28) \]

Integrating LHS from \( V_{out}(T_R) \) to \( V_{TCP} \) and RHS from \( T_R \) to \( T_{TCP} \), we get:

\[ (C_L + C_p) \int_{V_{out}(T_R)}^{V_{out}(T_{TCP})} \frac{dV_{out}}{V_{TCP}} = \mu C_{ox} \left( \frac{W}{L} \right) \left[ \frac{V_{sd}}{T_R} \right]_{T_R}^{T_{TCP}} \left( V_{sd} - V_{th} \right) \, dt \quad (29) \]

\[ (C_L + C_p) \left( \frac{V_{out}(T_R)}{V_{TCP}} \right) = \mu C_{ox} \left( \frac{W}{L} \right) \left[ \frac{V_{sd}}{T_R} \right]_{T_R}^{T_{TCP}} \left( V_{sd} - V_{th} \right) \, dt \quad (30) \]

Solving (30) for \( T_{TCP} \), we get:

\[ T_{TCP} = k_5 T_R + \left( K_5^2 T_R^2 + K_6 T_R \right) \quad (31) \]

Where \[ K_5 = \frac{\sqrt{V_{sd}^2 + V_{dis}^2}}{V_{sd}} \quad (32) \]

\[ K_5 = \frac{\sqrt{2V_{dis}^2 + 2V_{sd}^2}}{V_{sd}} \quad (33) \]

\[ K_6 = \frac{2 \mu C_{ox}}{V_{TCP}} \left( \frac{W}{L} \right) \quad (34) \]

3. Verification of the Proposed Timing Model

In this section, the validation of the results of section 2 is accomplished using HSPICE simulations. The model
coefficients are extracted using curve fitting of model expression on the simulation data. All the simulations are carried-out using ST Microelectronics 65 nm technology node. We start the simulation process with the appropriate selection of $W_p$ and $W_n$ for the CMOS inverter to get an equal value of the rise and fall time. The 20 to 80% variation of input is taken as the $T_R$ value. The variation of $TCP$s from different regions is verified with the variation of both $T_R$ and $C_L$. The figure 4 shows the variation of $TCP_{80}$ from region I with both $T_R$ and $C_L$ variation.

Figure 4: Variation of $TCP_{80}$ with $T_R$ and $C_L$

Figure 5 shows the variation of $TCP_{20}$ from region II with both $T_R$ and $C_L$ variation.

Figure 5: Variation of $TCP_{20}$ with $T_R$ and $C_L$

Figure 6 verify the observation (1, 2) for $TCP_{80}$ and observation (7, 8) for $TCP_{20}$ regarding the variation of TCP with width.

Figure 6: Variation of $TCP$s with $W_n$

Figure 7 verify the observation (3, 4) for $TCP_{80}$ and observation (9, 10) for $TCP_{20}$ regarding the variation of TCP with threshold voltage variation.

Figure 7: Variation of $TCP$s with $V_{th}$ variation

Figure 8 verify the observation (5, 6) for $TCP_{80}$ and observation (11, 12) for $TCP_{20}$ regarding the variation of TCP with mobility variation.

Figure 8: Variation of $TCP$s with mobility variation

Figure 9 represents comparision of proposed timing model (we call it as model 2) with state-of-the-art reported in [7] (we call it as model 1).
4. Efficient Approach for ECSM Characterization

The proposed timing model can be used to extract the ECSM LUTs for different TCPs with a single HSPICE simulation. The model coefficients are extracted by curve-fitting on HSPICE simulation data for a given PDK (in this case, 65 nm ST Microelectronics). ECSM LUTs for different TCPs are then determined using the models (8, 22, and 31) with these model coefficients. In Table 1, the LUT for TCP80 is shown for a combination of large $T_R$ and $C_L$. The accuracy of TCP80 has been verified with the HSPICE simulation. The maximum error is about 3%.

With this timing model, we can predict the delay of any random-sized inverter in an inverter chain. Figure 10 represents a random-sized inverter chain. Here the sizing of the first inverter is the same as the chosen value of $W_n$ and $W_p$ in section 2, and it is represented by 1X sized inverter, and all other inverters are sized in random fashion as a multiple of 1X sized inverter.

![Figure 10: Delay Estimation in a Random Size Inverter Chain](image)

For the DUT. Therefore, with a single HSPICE simulation, we can predict DUT delay in terms of TCPs for a $T_R$-$C_L$ combination.

![Figure 11: Validation of $T_{TCP}$ variation for DUT in an inverter chain](image)

5. Conclusion and Future Scope

We proposed a new timing model of CMOS inverter that supports a large range of $T_R$. This range may become crucial in the case of delay degradation due to NBTI and HCI. The model coefficients relation with the device-level parameter is derived. With this model, HSPICE simulations are reduced significantly. Therefore, the characterization process is further simplified. Due to model coefficient relation with device-level parameter variability analysis could be done without exercising Monte-Carlo analysis. This timing model works as a first step for the development of timing model for multistage and stacked logic as we can get information about intermediate node transition. This work is applicable to
planner devices, further it can be extended to FinFETs with some device level considerations. Finally, we may extend this work by updating the LUTs of TCPs with delay degradation offered by NBTI and HCI.

6. References


This work was supported in part by Semiconductor Research Corporation (SRC).