Variability aware gate delay model considering MIS for ultra-low power/energy CMOS circuits

Prasanjeet Das and Sandeep K. Gupta

Department of Electrical Engineering – Systems

University of Southern California

Los Angeles CA 90089

prasanjd@usc.edu, sandeep@usc.edu

Abstract—*Power is increasingly the primary design constraint for chip designers and one of the main techniques for addressing this concern is aggressive voltage scaling. Device variability increases with voltage scaling and significantly affects gate delays at low voltages. Although existing delay models for near- and sub-threshold circuits captures the effects of variability on gate delays, they do not capture advanced delay phenomenon such as multiple input switching (MIS; also known as near-simultaneous transitions) at inputs of a gate. As a result, most existing gate delay models often grossly underestimate worst case delays. In this paper we present a general approach for extending any delay model (pin-to-pin and beyond) to ensure that all minimum and maximum delay values computed are guaranteed to bound the corresponding delay values in silicon. We present extensive experimental results to demonstrate that MIS has significant impact (around 30-40%) on delays of near- and sub-threshold nominal gates. We develop our model empirically and show that it has practical run-time complexity and works equally well for super-, near- and sub-threshold circuits. In particular, via extensive experimentations we show that our model never underestimates the delay and tightly bounds the actual delays. (In contrast, in many of these experiments, existing delay models underestimate delays and always provide much looser bounds.)*

Keywords: multiple input switching, process variations, delay models, near-threshold circuits, sub-threshold circuits.

I. Introduction

Energy efficiency has become a ubiquitous design requirement for digital circuits and aggressive voltage scaling has emerged as the most effective way to reduce the energy use [1]. Traditionally, design optimization in the logic circuit community has always targeted the minimum-delay operational point (MDP), but as shown in Figure 1 the energy constraints have shifted the focus from traditional minimum delay operational region to ultralow-energy region around the minimum energy operational point (MEP) [3]. This shift in paradigm resulted in emergence of the family of circuits known as near-threshold circuits (NTVC) and sub-threshold circuits (STVC).

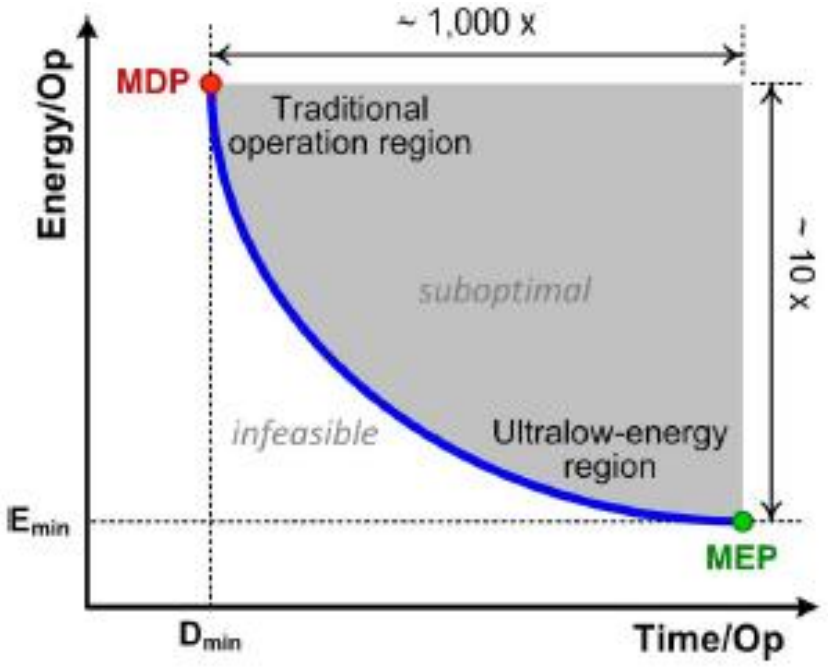

Figure 1: Energy delay trade-off in combinational logic [3]

Delay models are foundations of most of pre- and post-silicon timing related tasks [23]. It is imperative for a gate's delay model to accurately represent the logic as well as timing behavior of the gate. It is particularly important to ensure that a delay model never underestimates the actual delay. As the fabrication process moves into nano-scale, the importance of many delay phenomena [24][25] and levels of process variations [29] are growing.

A basic delay model considers basic delay determinants of the gate, such as input slew and output load. These models are extended to derive advanced delay models that capture additional phenomena associated with timing behavior. The timing behavior can be classified into three – single input switching (SIS), multiple inputs switching for to-controlling (MIS-TC) transitions, and multiple inputs switching for to-non-controlling (MIS-TNC). It is now increasingly common for delay models to also account for additional effects, such as crosstalk, ground bounce, and variability [23].

Given the wide feasible range of voltage scaling [2], it is important to analyze it's effect on delay (see Figure 2). In super-threshold regime ($V_{dd} > V_{th}$), circuit delay increases mostly quadratic with decreasing voltage. In near-threshold regime ($V_{dd} \sim V_{th}$), there is approximately 10X performance degradation compared to super-threshold region and in the sub-threshold regime ($V_{dd} < V_{th}$) delay increases exponentially with decrease in V_{dd}. This three-fold sensitivity of delay to voltage scaling, necessitates a delay model which can help accurately predicting the delay of circuits in all the three regions easily.

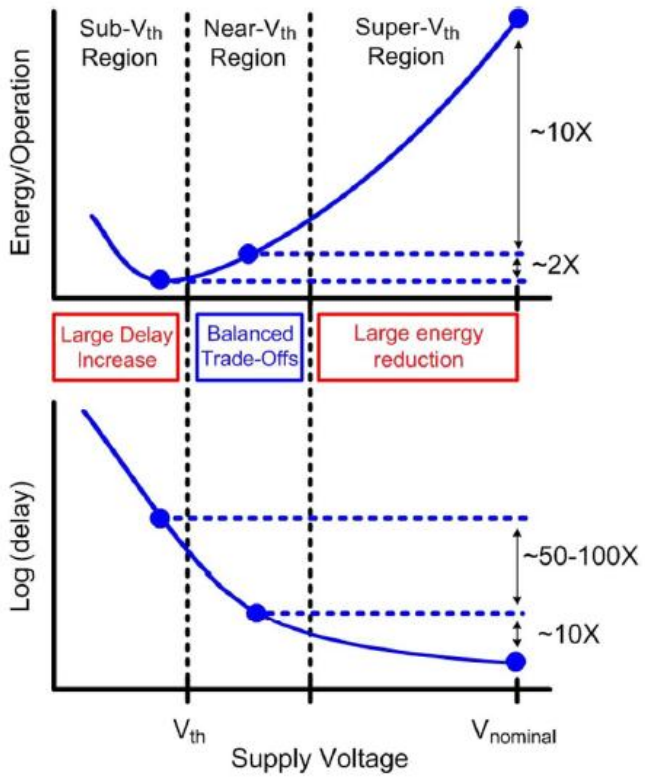

Figure 2: Energy and delay in different supply voltage operation regions [2]

One of the major barriers preventing NTVC and STVC to go mainstream is the increased delay variation [2][4]. Figure 3 shows the sensitivities of major delay defining parameters for the three operational regions. Since I_{on}/I_{off} ratio for near-threshold and super-threshold circuits are around the same order of magnitude, I_{off} plays a seemingly unimportant role in delay calculation and so does the sensitivity of I_{off} to variability. But, for subthreshold circuits I_{off} is significant (and so does the sensitivity of I_{off}) and hence the delay variation is much higher (exponential). This increases sensitivity of delay to voltage scaling and hence necessitates a variability aware delay model for NTVC [3][5] and STVC [8]-[12].

	Sub-V_{th}	*Near-V_{th}*	*Super-V_{th}*
V_{dd}	200 mV	400 mV	1 V
$V_{th,sat}$	270 mV	250 mV	180 mV
I_{on}	~20 μA/μm	~80 μA/μm	~1 mA/μm
Sensitivity of I_{on} to 100-mV V_{dd} reduction	18x	4.6x	1.20x
Sensitivity of I_{on} to 100-mV V_{th} increase	11x	3.7x	1.17x
Sensitivity of I_{off} to 100-mV V_{th} increase	16x	15x	12x
Sensitivity of $I_{on,n\text{-}FET}/I_{on,p\text{-}FET}$ ratio to 100-mV V_{th} mismatch	10x	3.7x	1.17x
I_{on}/I_{off} ratio	160x	3,150x	7,000x
I_{on}/I_{off} ratio vs. 100-mV V_{th} increase	1.44x	4.2x	11x

Figure 3: Comparison of key subthreshold, near-threshold and super-threshold sensitivities for 65nm [1]

It is widely known that MIS-TC transitions at the inputs of a primitive gate decrease the gate's delay due to activation of multiple charge/discharge paths [24]. On the other hand MIS-TNC transitions at inputs increase the gate's delay due to Miller effect. In this case, the gate's delay also depends on the initial state of the capacitances of internal nodes between series transistors, body effect, and impedance matching (history or stack effects) [25]. Though effect of MIS is widely acknowledged for super-threshold circuits [23], none of the existing methods for near- and sub-threshold circuits capture these (see next section). This is indeed a major gap in models for NTVC and STVC circuits.

In this paper we present a variability-aware gate delay model for low V_{dd} circuits (NTVC and STVC) that accounts for MIS, and tightly bounds the actual delays. We also show that our delay model, which is well-suited for both pre- and post-silicon timing related tasks, works equally well with super-,

near- and sub-threshold circuits and is much more accurate than the existing delay models for low V_{dd} circuits.

This paper is organized as follows. In Section II, the motivation and importance for our approach is presented. In Section III, our overall approach is outlined. The experimental setup and results are described in Section IV. Finally, conclusions are presented in Section V.

II. Motivation

Our motivation arises from the fact that even for near- and sub-threshold circuits MIS along with variability significantly affects CMOS gate delays, which becomes evident from the empirical observations in Table 1.

We also explored the effect of variability and MIS for all the three family of circuits at 65nm technology node using an industry standard library and circuit simulator (Spectre) and reported the results for the worst case delay of a 2-input NAND gate comprising of minimum size transistors in Table 1. We selected V_{dd} as 0.5V and 0.2V for NTVC and STVC respectively [1]. Waveforms with realistic values for input slew, skew between input transitions, and output load are applied using the characterization setup described in Section III.B, and results are reported in Table 1. Using the values of variations (from a foundry that fabricates chips in 65nm technology) for about 50 circuit parameters (including the major ones of V_{th}, L_{eff} etc), we perform Monte Carlo simulations to evaluate the effect of variability on gate delay. NV and FV stands for results with no variability and full variability; respectively. Full variability includes both global and local variabilities [29].

A. Effect of variability on gate delay

As evident from Table 1, in near-threshold and sub-threshold regions the effect of variability on gate delay is much more severe than in the super-threshold region. Results show that sub-threshold circuits can show up to 1100% delay variation compared to 50% delay variations observed on super-threshold circuits. This fortifies the notion that variability plays a major role in determining the gate delay for

ultra-low power CMOS circuits [6][8][9][10].

B. Effect of MIS on gate delay

Table 1 also shows that for a two input NAND gate comprising of minimum size transistors, effect of MIS in near- and sub-threshold circuits (though somewhat diminished compared to super-threshold circuits) is quite significant and can cause about 34% increase in delay (in case of NTVC) and hence cannot be ignored. The slight reduction in contribution of MIS to delay can be attributed to the mitigation of Miller effect (MIS-TNC – a major worst-case delay determinant for the results reported in Table 1) at reduced voltage levels.

C. Combined effect of MIS and variability on gate delay

It can be seen from Table 1 that MIS with variability further worsen the delay variation. For super-threshold circuits with full variability, MIS can increase the percentage delay variation from 52.82% for SIS (Max_FV (SIS)) to 116.36% for MIS (Max_FV (MIS)). Corresponding figures for near-threshold circuits and sub-threshold circuits are from 127% to 196% and from 1076% to 1125% respectively. Thus with or without variability, MIS plays a significant role in low-Vdd circuits as well and must be captured by delay models.

Table 1: Analysis of effect of variability and MIS on max delay of a 2-input NAND gate in 65nm

V_{dd}	Maximum delay				Delay deviation from SIS nominal (%)		
	Max_NV (SIS)	Max_NV (MIS)	Max_FV (SIS)	Max_FV (MIS)	Max_NV (MIS)	Max_FV (SIS)	Max_FV (MIS)
1.2V	113.4 ps	162.65 ps	173.3 ps	245.36 ps	43.69%	52.82%	116.36%
0.5V	9.397 ns	13.017 ns	21.39 ns	27.865 ns	38.52%	127.62%	196.53%
0.2V	5.54 µs	7.42 µs	65.2µs	67.9 µs	33.93%	1076.90%	1125.63%

D. Related work- existing near- and sub- threshold delay models

Existing delay models for NTVC are all empirical in nature where the on-current for a transistor is empirically determined is approximated by some form of the EKV model [21] and a fitting function [3][5][6][7] (see Table 2). Transistor delay is then approximated as a function of output load, on-current and

input voltage (input slew is accommodated in the fitting function). Subsequently, gate delay is represented as a function of transistor delay based on single input switching. Finally path delay is calculated as a simple weighted sum of gates along the path. The delay models of [3][5][7] besides being highly approximate, do not account for MIS and variability. The delay model of [6] (again approximate and SIS based) considers delay variability as a function of on current variability and represents gate delay as a normal distribution. Path delay distribution is then calculated using statistical analysis on normal distributions.

Table 2: Review of existing delay models for NTVC/STVC

Existing delay models for near-threshold circuits					
Reference #	**Analytical**	**Empirical**	**SIS**	**MIS**	**Variation-aware**
[3][5]		Y	Y	N	N
[6]		Y	Y	N	Y
[7]		Y	Y	N	N
Existing delay models for subthreshold circuits					
Reference #	**Analytical**	**Empirical**	**SIS**	**MIS**	**Variation-aware**
[1][19]		Y	Y	N	N
[8][9][10][11] [12][14][17][20]		Y	Y	N	Y
[13][15][16]	Y		Y	N	N
[18]	Y		Y	N	Y

Existing delay models for STVC are either empirical (similar to NTVC delay models) [3][5][8]-[12][14][17][19][20] or analytical (gate delay arrived at solving integrals of on-currents over time based on region of operation) [13][15][16][18].The analytical models though more accurate are too complex (even when ignoring MIS)for static/statistical timing analysis. Statistical delay models for STVC [8]-[12][14][17][18][20] represent gate delay distribution (based on current distribution) as a log-normal variable and path delay distribution is obtained using statistical operation on log-normal distributions [11].

All existing delay models for NTVC and STVC (see Table 2) are SIS based (ignores the effect of near-simultaneous transitions), vector unaware (unable to work with unspecified or partially-specified vectors) and deals with variation using distributions and not bounds. Since all timing related tasks (pre-silicon and post-silicon) require a delay model that is resilient and accurate, we decided to evaluate our resilient delay model [23] for NTVC and STVC circuits as well.

III. The approach – resilient delay model

In this section we will present our gate delay modeling approach (based on [23] for super-threshold circuits) for near-threshold circuits. Sub-threshold gate delay modeling can be performed identically.

A. Our overall approach

An accurate and resilient delay model for pre- and post-silicon tasks in near- and sub-threshold circuits must have the following characteristics:

- Captures known and emerging gate delay phenomena[23] as well as variability.
- Only uses bounding approximations to tackle unknowns and any simplifications necessary to make complexity manageable.
- Enables computation of tight timing ranges in an efficient manner (manageable complexity).

B. Characterization setup

The models in [24][25] only uses qualitative information, such as causality and other provable properties of underlying physics such as the effect of MIS and hence are used as the starting points.

The modeling approach in [24][25] consists of:

- Perform simulations by varying all input parameters (input slews, input skews, and the initial state of internal capacitances) over their typical ranges.
- Quantify the significance of associated delay phenomena for simultaneous transitions from the simulation data.
- Identify appropriate input waveforms that activate each phenomenon.
- Develop an empirical model that identifies input waveforms that excite these phenomena.

The characterization setup shown in Figure 4 can be used to feed the gate with different realistic waveforms with different attributes. The standard delay characterization methodology requires the delay calculation of a gate driving capacitive load to bedriven by a load less cell [23]. Thus, the parameters (C_1, S_{in}) of the driver circuit can be changed to generate waveforms with different transition times, skews (between pair of waveforms). The voltage at node N_0 is copied to node N_1 to ensure the load less driver

requirement of delay characterization. Note that S_{in} responsible for voltage N_0 is an ideal voltage source, but S_m resulting in N_1 is a controlled voltage source with voltage V (N_1) = V (N_0). Switch SW can be used to get the mirrored (SW=1) or inverted (SW=0) waveform. We also varied the internal capacitances (either precharged or pre-discharged) in our characterization step to account for Stack Effect [23]. We varied C_r to account for different capacitive loads on the gate's output.

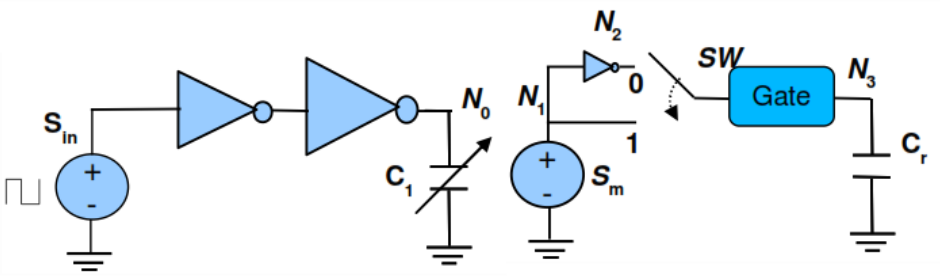

Figure 4: The characterization setup.

C. Basic delay model

In order to quantify the significance of associated delay phenomena for the effect of MIS on near-threshold gate delay, for a NAND gate in a 65nm CMOS technology we perform circuit-level simulations by varying input transition times (T_R) from 1 ns to 2ns, skew (δ) from -5ns to 5ns, and output load (C_L) from FO1 to FO4 at the supply voltage V_{dd} of 0.5V. (We used Spectre for all our simulations.) We arrived at the *delay vs. skew relationship* for near-simultaneous transitions (MIS) at the inputs of a 2-input NAND gate.

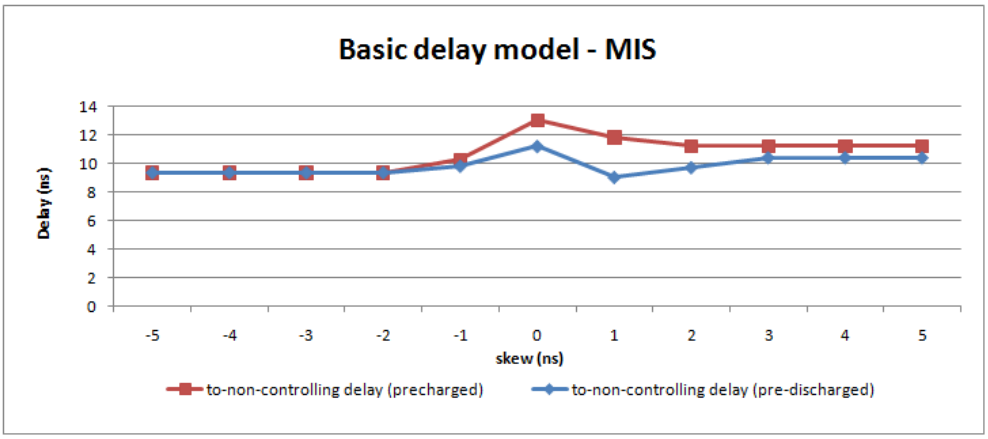

Figure 5: Delay vs skew curve for simultaneousto-non-controlling transitions (Basic delay model)

Figure 5 shows the delay vs. skew relationship (for T^X_R = 1 ns, T^Y_R = 1 ns, C_L = FO4) for MIS. It is clear from Figure 5 that multiple input switching can significantly affect delay – SIS underestimates the delay by about 38% (see Section II). Similarly, output transition time functions are derived.

The curves, such as in Figure 5, capture all the known basic delay phenomena associated with MIS. Simultaneous to-controlling transitions [23] at inputs of a primitive gate decrease gate delay due to activation of multiple charge/discharge paths. On the contrary, simultaneous to-controlling transitions at inputs of primitive gate increase the gate delaydue to a combination of various effects such as *short circuit current, initial state of internal capacitances (precharged and pre discharged), Miller effect, Body effect, Stopping early discharge and Impedance matching [23][25]*.

D. Timing functions

Given arrival times and transition times at a gate's inputs, and the initial state of internal capacitances, we compute timing functions for gate delays and output transition times [24][25]. The output arrival time and transition time is computed from the above data in our advanced timing analyzer - ETA [28]. Figure 6 shows the delay timing functions for a 2-input NAND gate where all inputs of the gate have either steady non-controlling values or transitions in one direction.Similarly output transition time (rise and fall) functions can be obtained. Here, N_c represents the number of internal capacitances which are precharged to $V_{dd} - V_{th}$ and $\delta = A_Y - A_X$ is the skew. Gate delay can now be represented by the following timing functions:

- MIS-TC (simultaneous to-controlling): Rise delay function d^Z_R (T^X_F, T^Y_F, $\delta^{Y,X}$).
- MIS-TNC (simultaneous to-non-controlling): Fall delay function d^Z_F (T^X_R, T^Y_R, $\delta^{Y,X}$, N_c).

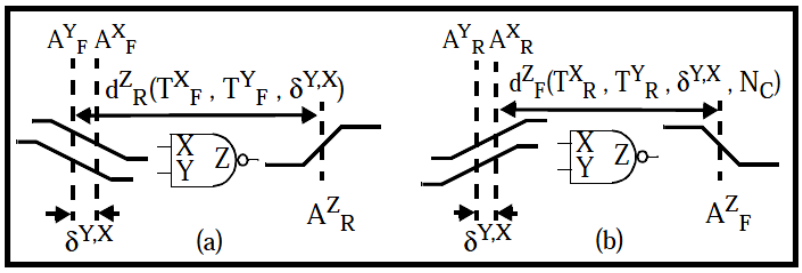

Figure 6: (a) Rise delay function and (b) Fall delay function

The timing functions can then be approximated into a piecewise linear model using empirical equations arrived at by using curve fitting as shown in Figure 7 where the accuracy of the model can be traded off with the complexity for development and the usage in the subsequent timing analysis.Note that any approximation we use to reduce complexity ensures that our model bounds actual delay in silicon.

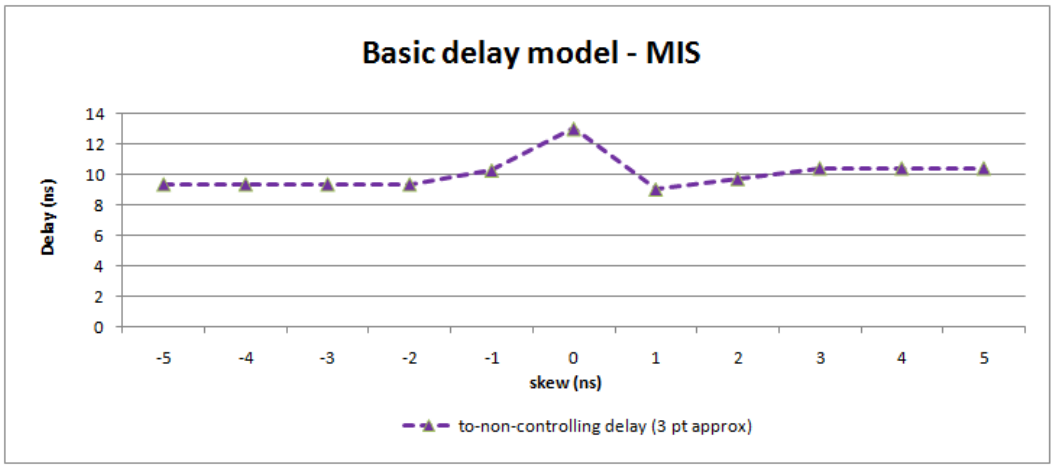

Figure 7: Piecewise linear approximation for simultaneous to-non-controlling transitions (Basic delay model)

E. Incorporating variability and bounding approximations

A single pair curve combining the two cases (shown in Figure 5) cannot capture all inaccuracies and variations. Hence, we capture the inaccuracies and variations in these delay parameters using bounding approximations. We consider process variations in terms of the parameters of the devices in the gates, such as V_{th}, L_{eff} etc. Using the values of variations (from a foundry that fabricates chips in 65nm technology) for about 50 circuit parameters (including the major ones of V_{th}, L_{eff} etc), we perform Monte Carlo simulations to obtain the two envelopes to bound the constellation of points representing the gate delay under variability, as represented by the two outer envelopes in Figure 8.

The relationship of Figure 8 can be easily represented as a pair of three/four point piecewise linear approximations (pair of curves in Figure 9).The two envelopes obtained, represent the bounded approximations for the resilient delay model. The tightness of the bound defines the cost-benefit of the subsequent steps of timing analysis, path selection, and vector generation. The relationship between output transition times and skew are derived in a similar manner. Also, for each timing function in [24][25] we

now have two sets of functions corresponding to the upper and lower bounds.

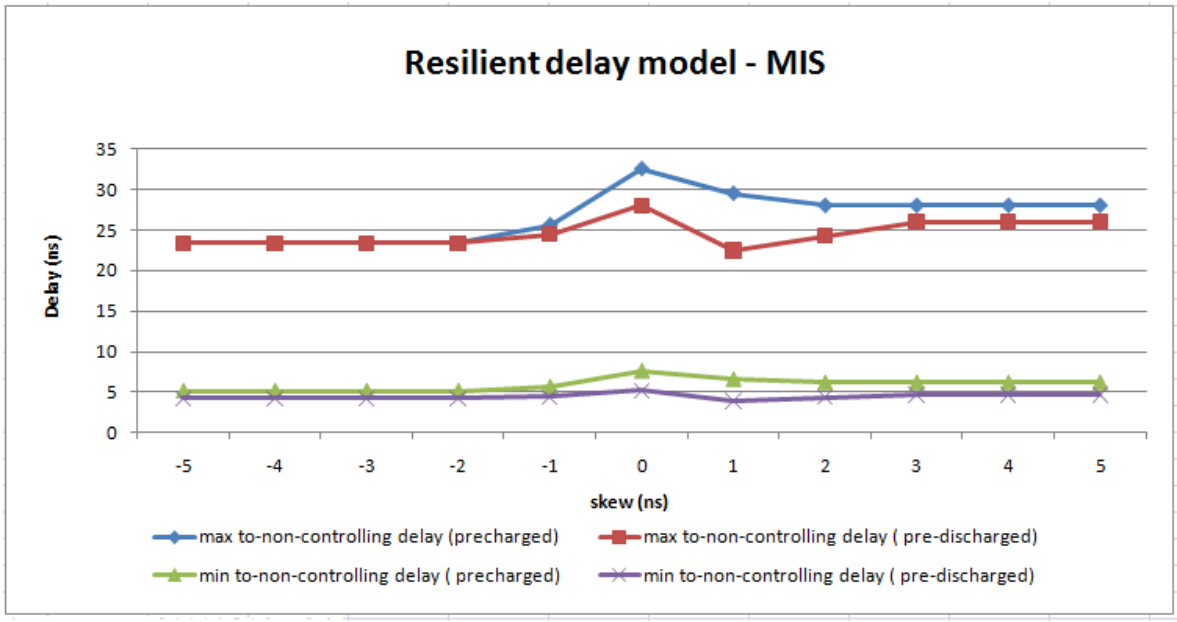

Figure 8: Delay vs skew curve for to-non-controlling transitions (Resilient delay model)

We would also like to bring to attention to the fact that though characterization of delay model is originally done for 11 skew points from -5 ns to +5 ns in steps of 1 ns, we only store the 3/4 skew points corresponding to the three/four point approximation and hence do not increase the characterization effort significantly.

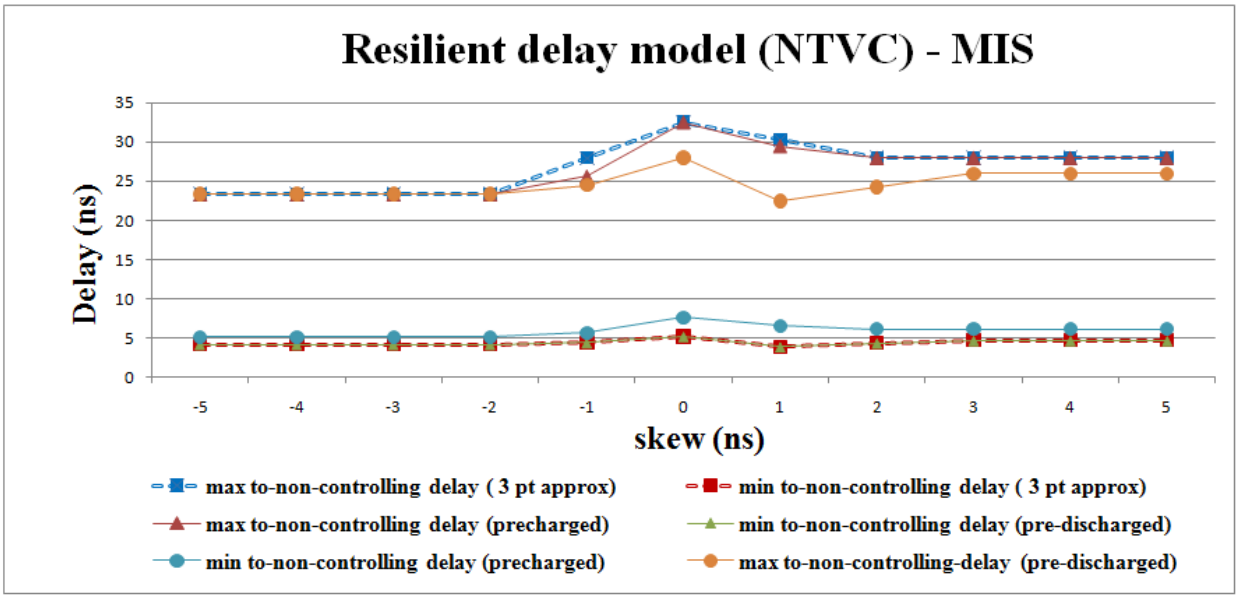

Figure 9: Resilient delay model for to-non-controlling transitions – 3,4 point linear bounding

The basic pin-to-pin delay model looks like a step function with a step at zero skew, where the size of step captures the difference between the pin-to-pin delays for the two inputs. Another way to bound the resilient delay model will be to use the simple pin-to-pin delay model to derive the bounds [23]. Please refer to [23] for more details about tightness of pin-to-pin bounding and associated characterization complexity (storage as well as runtime) trade-offs.

F. Extended Model

The proposed resilient delay model can handle different numbers of inputs, input positions and can be extended to handle more than two simultaneous transitions using the approaches from [25]. The extended delay model though more accurate, needs more cases to be enumerated and the corresponding characterization effort increases. Also, the timing analysis framework needs to consider more timing cases and thus the runtime complexity explodes. A workaround is to decompose the gates in the circuit as 2-input gates for delay calculation for post-silicon delay related tasks.

G. Application of delay model

The aforesaid delay model can be used for the following three pre-silicon tasks targeting pre- and post-silicon delay related activities (not central to this paper, so refer to [22][26][27] for more details).

1) Enhanced Timing Analysis

Given the arrival and transition times at a gate's inputs, we calculate the corresponding quantities at the gate's outputs. Figure 10 shows the possible input combinations for a rising/falling transition at output.

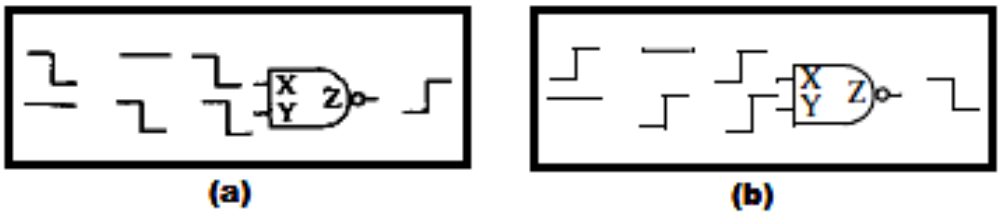

Figure 10: Possible input combinations for (a) output rising transition (b) output falling transition.

We enhanced the approach in [23][26][28] for both MIS, where by using the bitonic relationship between delay and input transition time and exploring the possible transition times within the min-max range in the input transition time vs. delay curves, one arrives at the equations for output arrival times (see Figure 11) and output transition times.

$$A_{RS}^{Z} = \min\left[A_{FS}^{X}, A_{FS}^{Y}\right] + min_{\beta\gamma \in \{S,L\}}\left[d_{Rmin}^{Z}\left(T_{F\beta}^{X}, T_{F\gamma}^{Y}, A_{FS}^{Y} - A_{FS}^{X}\right)\right]$$

$$A_{RL}^{Z} = \max\left[A_{FL}^{X} + d_{Rmax}^{Z,X}\left(T_{F}^{X*}\right), A_{FL}^{Y} + d_{Rmax}^{Z,Y}\left(T_{F}^{Y*}\right)\right]$$

$$A_{FS}^{Z} = \min\Big[A_{RS}^{X} + \min\left\{d_{Fmin}^{Z,X}\left(T_{RS}^{X}\right), d_{Fmin}^{Z,X}\left(T_{RL}^{X}\right)\right\}, A_{RS}^{Y} + \min\left\{d_{Fmin}^{Z,Y}\left(T_{RS}^{Y}\right), d_{Fmin}^{Z,Y}\left(T_{RL}^{Y}\right)\right\},$$
$$\max\left[A_{RS}^{X}, A_{RS}^{Y}\right] + min_{\beta\gamma \in \{S,L\}}\left[d_{Fmin}^{Z}\left(T_{R\beta}^{X}, T_{R\gamma}^{Y}, A_{RS}^{Y} - A_{RS}^{X}, N_c = 0\right)\right]\Big]$$

$$A_{FL}^{Z} = \max\left[A_{RL}^{X}, A_{RL}^{Y}\right] + max_{\beta\gamma \in \{S,L\}}\left[d_{Fmax}^{Z}\left(T_{R\beta}^{X}, T_{R\gamma}^{Y}, A_{RL}^{Y} - A_{RL}^{X}, N_c = 1\right)\right]$$

Figure 11: Calculation of output arrival times in ETA using resilient delay model.

2) Path selection

We enhanced the approach in [27] that identifies a set of paths that is guaranteed to include all paths that may potentially cause a timing error if the accumulated values of additional delays along circuit paths isupper bounded by a desired limit (Δ), works with upper and lower bounds given by our resilient delay model and also checks for both functional sensitization and high delay excitation.

3) Vector generation approach

In [22] a new approach to generate vectors for post silicon delay characterization using the proposed resilient delay model is presented. The method generates vectors that are guaranteed to excite the worst-case delays of fabricated chips without introducing any pessimism by intelligently dividing the delay model to various timing rangesand innovatively exploiting the effect of MIS on the gate delay in these timing ranges.

IV. Experimental results

We applied our approach to combinational parts of ISCAS89 benchmark circuits using an Intel Core 2 Duo 2.2 GHz machine. All gates in the benchmark circuits are assumed to use minimum-size

transistors, and a 65nm CMOS technology is used. Our experiments used our new resilient delay modelfor both to-controlling as well as to-non-controlling MIS transitions. We selected V_{dd} as 0.5V and 0.2V for NTVC and STVC respectively [1]. Using the approach described earlier we characterize all the basic gates (NAND, NOR, AND, OR, NOT and BUF).

Table 3: Comparison of NTVC/STVC delay models for basic circuits

NTVC Analysis					
Basic circuit	**Max delay (ns)**			**Max delay error (%)**	
	Spectre	**[3][5]**	**Ours**	**[3][5]**	**Ours**
INV	1.283	1.260	1.285	-1.8%	0.15%
NAND	13.017	9.397	13.032	-27.9%	0.115%
INV-INV-INV	3.684	3.629	3.69	-1.5%	0.16%
NAND-NAND-NAND	27.707	20.267	27.752	-26.85%	0.162%
STVC Analysis					
Basic circuit	**Max delay (µs)**			**Max delay error (%)**	
	Spectre	**[8]-[12]**	**Ours**	**[8]-[12]**	**Ours**
INV	0.45	0.45	0.45	0%	0%
NAND	7.42	5.55	7.43	-25.2%	0.13%
INV-INV-INV	1.45	1.45	1.455	0%	0.34%
NAND-NAND-NAND	15.33	11.4	15.5	-25.63%	1.13%

In Table 3 we compare the delay model of [3][5] for NTVCwith our resilient delay model for basic circuits. As evident, even for a chain of 3 NAND gates, the SIS based delay model of [3][5] will report a delay with an error of about -27% whereas our resilient model reports delay within an error of 0.2%. The inaccuracy of [3][5] can be attributed to factors such as ignoring MIS, approximations such as path delay being sum of gate delays. Table 3 also shows the results for a similar comparison of our resilient delay model with the delay modelof [8]-[12]. Our resilient delay model gives much better results (1.1% compared to -25.63%) even for a simple circuit such as a chain of 3 NAND gates. Moreover, we would like to draw attention to the fact that existing delay models for NTVC/STVC grossly underestimates the actual delay (as evident from the negative sign), and are unable to bound the worst case gate delay in any meaningful way.

For medium size ISCAS benchmarks to evaluate the effect of ignoring MIS, we simulated randomly generated 10,000 vectors using Spectre and ETS (our timing simulator) and compared them to the result calculated by our ETA (Table 4).

Table 4: Comparison of NTVC/STVC delay models for ISCAS benchmark circuits

NTVC Analysis									
Benchmark	**Max delay reported (ns)**					**Accuracy (%)**			
	Random simulations			**Analysis**		**Simulations**		**Analysis**	
	Spectre	**ETS (ours)**	**ETS [3][5]**	**ETA (ours)**	**ETA [3][5]**	**ETS ours**	**ETS [3][5]**	**ETA (ours)**	**ETA [3][5]**
s298	42.83	43.15	33.25	44.48	37.12	0.73%	-22.37%	3.7%	-13.3%
s444	59.71	61.04	45.86	63.66	51.67	2.22%	-23.20%	6.2%	-13.47%
s953	58.72	61.57	46.09	63.97	52.49	4.84%	-21.51%	8.2%	-10.61%
s1196	122.06	125.4	92.23	137.92	101.4	2.74%	-24.44%	11.5%	-16.93%
STVC Analysis									
Benchmark	**Max delay reported (µs)**					**Accuracy (%)**			
	Random simulations			**Analysis**		**Simulations**		**Analysis**	
	Spectre	**ETS (ours)**	**ETS [8]-[12]**	**ETA (ours)**	**ETA [8]-[12]**	**ETS (ours)**	**ETS [8]-[12]**	**ETA (ours)**	**ETA ([8]-[12])**
s298	22.02	22.72	16.5	23.21	18.1	0.83%	-25.09%	5.1%	-17.82%
s444	36.76	38.2	24.3	40.04	28.4	3.92%	-33.89%	8.2%	-22.74%
s953	33.78	34.52	26.8	37.66	29.2	2.18%	-20.67%	10.3%	-13.55%
s1196	61.72	63.05	48.4	71.36	50.5	2.14%	-21.58%	13.5%	-18.18%

In Table 4 for ETS and ETA we report the results with our resilient delay model and existing delay models for NTVC ([3][5])/STVC([8]-[12]) as well. As expected, our resilient delay model based ETS gives much better results than the delay models for NTVC [3][5] and STVC [8]-[12]. For s1196 the error in delay estimate by our simulation based approach for NTVC and STVC are 2.74% and 2.14% respectively. Corresponding figures for the approach of [3][5] (for NTVC) and [8]-[12] (for STVC) are much higher (about -24% and -21%) respectively due to approximations and ignoring MIS. Again, the negative sign indicates the inability of existing SIS based approaches to establish meaningful bounds on the worst case gate delay. Even the static approach (ETA) based on existing delay models for NTVC ([3][5]) and STVC ([8]-[12])(that do not capture MIS) underestimates the actual delay by about -18% for s1196 operated in sub-threshold.

Monte Carlo simulations using circuit simulators for these medium and large size circuits are very time consuming, and given the number of process parameters varied for the 65nm industrial library provided to us, such simulations will take days. Hence for experiments with variability on these

benchmarks we report the max delay obtained our approach for generating vectors for validation to identify vectors [22] and compare them to the result calculated by our ETA (Table 5).

Table 5: ETA accuracy for NTVC/STVC (ISCAS benchmarks)

Benchmark	ETA Accuracy – Tightness of bounds (max) (%)			
	No variability		Full variability	
	NTVC	STVC	NTVC	STVC
s298	2.9%	3.5%	3.3%	6.2%
s444	3.5%	7.3%	6.8%	9.3%
s953	5.3%	8.2%	7.6%	12.5%
s1196	7.2%	9.5%	9.7%	14.3%
s5378	7.5%	11%	14.5%	23%
s9234	12%	15%	21%	33%

The inaccuracy of ETA increases for the experiments with variability (from 7.2% to 9.7% for NTVC and 9.5% to 14.3% for STVC in s1196 w.r.t. MDS simulations [22] that guarantee to include the worst case delay invoking vector). The results clearly demonstrate that our static approach is indeed tight and the inaccuracy can be attributed to Monte Carlo runs on the full circuit not covering the complete process space that is covered during characterization of gates. Results for STVC are more pessimistic because effect of variability on STVC is much more severe than NTVC.Note that for a medium sized benchmark circuit such as s1196 our ETA based static approach gives reasonably tight accuracy at a much lower complexity (ETA takes about 300 CPU clocks whereas the MDS based approach takes a much larger 150,000 CPU clocks).

We performed similar experiments with bigger benchmarks s5378 and s9234.Even for s9234 the looseness of ETA for the nominal case is around 12% and 21% (compared to MDS simulations [22]) for NTVC and STVC respectively, which is a reasonable accuracy considering the fact that ETA ran only once in a vector unaware manner to achieve this. Also, our results with variability tends to become much more looser as the size of circuit increases which can be attributed to looseness in bounding approximations, uncertainty in vector space and smaller process coverage during full circuit Monte Carlo simulation compared to gate characterization.

Since the proposed delay model will be used primarily for post-silicon delay related tasks and not design analysis, the delay will be measured on the fabricated chip and not estimated by the loose ETA. The

bounding approximation will definitely contribute to the looseness of the estimation but the delay measured by applying vectors generated by our approach [22] will be the actual delay with *zero* pessimism.

V. Conclusion

Experimental results demonstrate that our new resilient delay model for low V_{dd} circuits captures the effect of MIS and variability, is much more accurate than existing low Vdd delay models and generates tight bounds at low complexity. It can also tighten these bounds using available logic values at any circuit lines and thus is suitable even for post-silicon tasks.

REFERENCES

[1] S. Hanson et al., "Ultralow-voltage, minimum-energy CMOS", In *IBM Journal of Research and Development*, Vol. 50, Issue 4,2006, pp. 469-490.

[2] R. G. Dreslinski et al., "Near-Threshold Computing: Reclaiming Moore's Law through Energy Efficient Integrated Circuits", In *Proc. IEEE,* Vol. 98, No. 2, 2010, pp. 253-266.

[3] D. Markovic et al., "Ultralow-Power Design in Near-Threshold Region", In *Proc. IEEE,* Vol. 98, No. 2, 2010, pp. 237-252

[4] B. H. Calhoun, and D. Brooks, "Can Subthreshold and Near-Threshold Circuits Go Mainstream", In *IEEE MICRO*, Vol. 30, Issue 4, 2010, pp. 80-85.

[5] D. M. Harris, B. Kellar, J. Karl, and S. Kellar, "A Transregional Model for Near-Threshold Circuits with Application to Minimum-Energy Operation", In *International Conf. on Microelectronics*, 2010, pp. 64-67.

[6] M. Slimani, F. Silveira, and P. Matherat, "Variability-speed-consumption trade-off in near threshold operation", In*Proc. PATMOS*, 2011, pp. 308-316.

[7] S. Fisher et al., "An Improved Model for Delay/Energy Estimation in Near-Threshold Flip-Flops",In *International Symp. On Circuits and Systems*, 2011, pp. 1065-1068.

[8] B. H. Calhoun, A. Wang, and A. P. Chandrakasan, "Modeling and Sizing for Minimum Energy Operation in Subthreshold Circuits," In *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 9, 2005, pp. 1778-1786.

[9] J. Kwong, and A. P. Chandrakasan, "Variation-Driven Device Sizing for Minimum Energy Sub-threshold Circuits", In *Proc. International Symp. On Low Power Electronics and Design*, 2006, pp. 8-13.

[10] N. Verma, J. Kwong, and A. P. Chandrakasan, "Nanometer MOSFET Variation in Minimum Energy Subthreshold Circuits", In *IEEE Trans. on Electronic Devices*, vol. 55, No. 1, 2008, pp.163-174.

[11] S. H. Chou, "Computationally efficient characterization of standard cells for statistical static timing analysis"– ME Thesis, Dept. of Electrical Engineering and Computer Science, MIT

[12] B. Zhai, S. Hanson, D. Blauuw, and D. Sylvester, "Analysis and Mitigation of Variability in Subthreshold Design", In *Proc. International Symp. On Low Power Electronics and Design*, 2005, pp. 20-25.

[13] A. Raychowdhury, B. C. Paul, S. Bhunia, and K. Roy, "Computing With Subthreshold Leakage: Device/Circuit/Architecture Co-Design for Ultralow-Power Subthreshold Operation", In *IEEE Trans. On VLSI Systems*, Vol. 13, No. 11, 2005, pp. 1213-1224.

[14] Y. Lin and V. D. Agrawal, "Statistical Leakage and Timing Optimization for Submicron Process Variation", In *International Conf. on VLSI Design*, 2007, pp. 439-444.

[15] J. R. Tolbert, and S. Mukhopadhyay, "Accurate Buffer Modeling with Skew Propagation in Subthreshold Circuits", In *International Symp. On Quality Electronic Design*, 2009, pp. 91-96.

[16] A. Valentian et al., "Modeling Subthreshold SOI Logic for Static Timing Analysis", In *IEEE Trans. on VLSI Systems*, Vol. 12, No. 6, 2004, pp. 662-668.

[17] T. Lin et al., "Analytical Delay Variation Modeling for Evaluating Sub-threshold Synchronous/Asynchronous Designs", In *IEEE NEWCAS Conf.*, 2009, pp. 69-72.

[18] F. Frustaci, P. Corsonello, and S. Perri, "Analytical Delay Model Considering Variability Effects in Subthreshold Domain", In *IEEE Trans. on Circuits and Systems II: Express Briefs*, Vol. 59, No. 3, 2012, pp. 168-172.

[19] Y. Osaki et al., "Delay-Compensation Techniques for Ultra-Low-Power Subthreshold CMOS Digital LSIs," In *IEEE International Midwest Symp. on Circuits and Systems*, 2009, pp. 503-506.

[20] N. Lotze, J. Goppert, and Y. Manoli, "Timing Modeling for Digital Sub-threshold Circuits",In *Proc. Design Automation and Test in Europe Conf.*, 2010, pp. 299-302.

[21] C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An Analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications", In *Journal of Analog Integrated Circuits and Signal Processing*, Vol. 8, No. 1, 1995, pp. 83-114.

[22] P. Das and S. K. Gupta, "On generating vectors for accurate post-silicon delay characterization", In *Proc. Asian Test Symp.*, 2011, pp. 251 - 260.

[23] P. Das and S. K. Gupta, "Capturing Variability in advanced gate delay models," USC technical report No.CENG-2010-3, 2010.

[24] L. C. Chen, S. K. Gupta, and M. A. Breuer, "A New Gate Delay Model for Simultaneous Switching and Its Applications", In *Proc. Design Automation Conf.*, 2001, pp. 289-294.

[25] L. C, Chen, S. K. Gupta, and M. A. Breuer, "Gate Delay Modeling for Multiple to non-controlling stimuli" – Unpublished.

[26] L. C. Chen, S. K. Gupta, and M. A. Breuer, "A New Framework for Static Timing Analysis, Incremental Timing Refinement, and Timing Simulation", In *Asian Test Symp.*, 2000, pp. 102-107.

[27] I.D. Huang and S. K. Gupta, "Selection of Paths for Delay Testing", In *Proc. Asian Test Symp.*, 2005, pp. 208 - 215.

[28] I. D. Huang and S. K. Gupta, "On Generating Vectors That Invoke High Circuit Delays - Delay Testing and Dynamic Timing Analysis", In *Proc. Asian Test Symp.* , 2007, pp.485-492.

[29] K. Bernstein et al., "High- Performance CMOS variability in the 65-nm regime and beyond", In *IBM Journal of Research and Development*, Vol. 50, Issue 4.5, 2006, pp. 433-449.