Center for Embedded Computer Systems
University of California, Irvine

_____

Variability-Aware Static Latch Modeling

Il-Joon Kim, Amin Khajeh, Fadi J. Kurdahi, and Ahmed M. Eltawil

Center for Embedded Computer Systems
University of California, Irvine
Irvine, CA 92697-2620, USA

ijkim@uci.edu

Variability-Aware Static Latch Modeling

Abstract

In this paper we study the impact of variability on the transmission gate based latch. The threshold voltage (Vt) fluctuation due to Random Dopant Fluctuation (RDF) and Process, Voltage, and Temperature (PVT) effects to propagation delay, as well as subthreshold leakage and probability of failure are discussed. We propose a modeling methodology which is not tied to a specific topology such as Monte Carlo simulation. To generate failure analysis, we sampled the probability domain and reconstructed the probability density function.

Keywords

Latch, Process Variation, RDF, Error Tolerance, Propagation Delay

1. Introduction

The desire to improve device performance has resulted in aggressive scaling of technology to below 45nm. One of the side effects of scaling is an increase in process parameter variation. Due to this effect, there is an increased probability of failure due to excessive delay and leakage. To address the effects of process variation, designers are forced to add margins to their designs to guarantee correct functionality under process variation. One of the most popular methods of assessing designs under process, temperature and voltage variations is Monte Carlo simulation. Monte Carlo simulations rely on repeated random sampling to compute results and are used to model phenomena with uncertainty in inputs, or design parameters. The main drawback of Monte Carlo simulations is runtime and scalability. As the number of the random variables in the design (in this case transistor parameters) or extent of variability (number of sigmas to consider as limits of correct operation) increase, the number of needed Monte Carlo Simulations points increases exponentially and at some point it becomes impossible to draw meaningful conclusions from the those simulations. To address this, new sampling methods are introduced such as Latin Hypercube or Sobol Sampling [1][2]. However all these methods rely on generating sets of samples is simulation them which again is a temporary solution since they lose their benefits as the design size increases. To address this for our failure analysis we sampled the probability domain instead of the design parameters. We started by identifying the most sensitive parameters in the design. Then we uniformly sampled the threshold voltage of each sensitive device and measured the delay. Once we know the relationship between the threshold voltage shift and delay, we can reconstruct the probability density

function (PDF) of the delay [3]. Having the PDF of the delay, we can find the probability of error regardless of number of sigmas that we want to consider. More details are given in the simulation setup section.

2. Simulation Setup

2.1. Transmission gate based latch simulation scheme

The transmission gate based latch is presented in Figure 1. This structure of the latch is most widely used part in processors and DSPs. It has three input signals, D, CLK, and CLKbar. The D input is selected when the clock signal (CLK) is high, and the Q output is held (using feedback) when the CLK is low [4]. Hence this latch is positive latch.
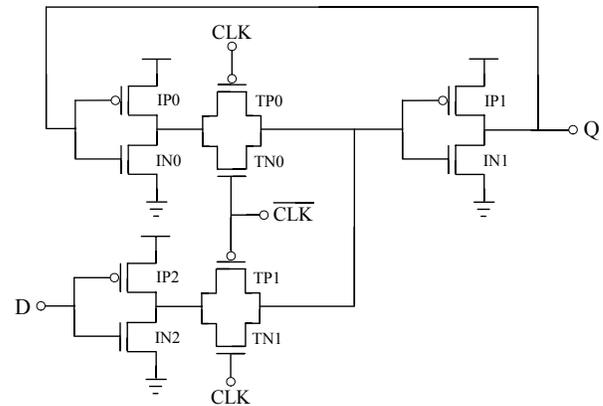
Figure 1: Transmission gate based latch

A chain of inverters scheme is used as realistic waveform generator and it is given in Figure 2. Input signals (D, CLK, and CLKbar) come from each chain of inverters. So it is possible to make their waveforms closer to actual silicon and scale with Process, Voltage and Temperature (PVT). With initial condition of simulation, rising time is changed from 50ps to 92.75ps for CLK and from 80ps to 127.7ps for D, respectively.
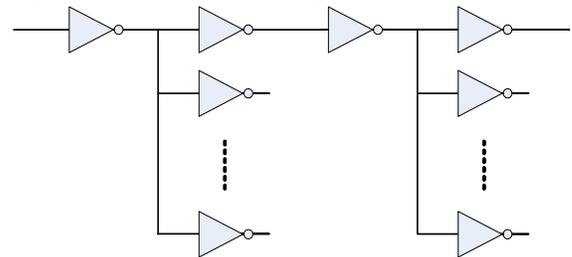
Figure 2: Chain of inverters scheme for input signals

Figure 3 illustrates the simulation flow chart. The simulation is performed two stages. At the first stage, input signals are generated using inverter chains with specific PVT condition. The realistic input signal waveforms are dumped out and saved. Using the dumped waveforms, a set of simulations is performed during the second phase with uniform sampling which is in probability domain. By separating simulation into two stages, simulation time can be decreased. And by using uniform sampling simulation instead of Monte Carlo simulation, simulation is more accurate and faster. To generate failure analysis, the probability domain is sampled and the probability density function is reconstructed as explained later.
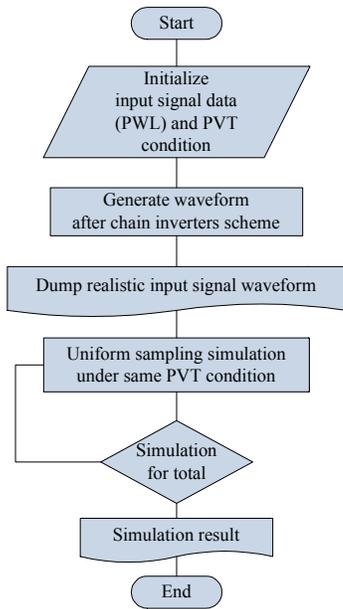
Figure 3: Sampling simulation flow chart

There are 4 different states of Propagation Delay (PD) which is pictorially represented in Figure 4.

- DQ_Delay_RR : When CLK is logic high and Q is logic low, rising edge of D change the Q to rise
- DQ_Delay_FF : When CLK is logic high and Q is
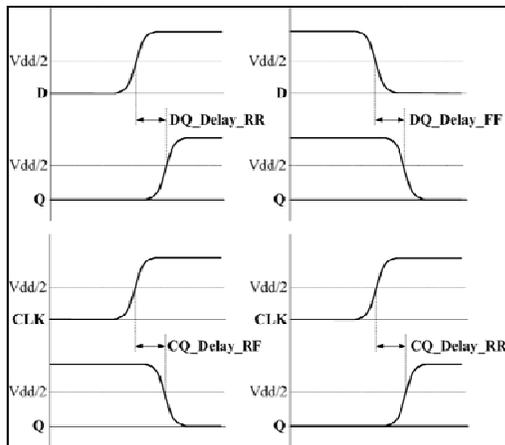
Figure 4: Naming of propagation delay

logic high, falling edge of D change the Q to fall
- CQ_Delay_RF : When D is logic low and Q is logic high, rising edge of the CLK change the Q to fall
- CQ_Delay_RR : When D is logic high and Q is logic low, rising edge of the CLK change the Q to rise

PD is measured from time of triggered signal when it is on half of Vdd to target signal when it is on half of Vdd.

The paralleled connected inverter is used at output Q. It increases capacitance value at output node to make more realistic simulation condition. Compare with that there is no load inverters at output Q, PD is changed about 20ps~40ps.

2.2. Vt Fluctuations due to RDF

The Vt fluctuations due to Random Dopant Fluctuation (RDF) can be considered as zero-mean Gaussian random variables [5]. Under RDF, Vt of transistors have independent random variations (δVt) with mean=0. The standard deviation of the δVt which can be denoted in σ_{Vt} is given by [6]:

$$\sigma_{Vt} = \sigma_{Vt0}\sqrt{\frac{L_{min}}{L}\frac{W_{min}}{W}}$$

where σ_{Vt0} is the σ_{Vt} for minimum sized transistor and it is given by [7]:

$$\sigma_{Vt0} = \frac{qT_{ox}}{\varepsilon_{ox}}\sqrt{\frac{N_aW_d}{3L_{min}W_{min}}}$$

where N_a is the effective channel doping, W_d is the depletion region width, T_{ox} is the oxide thickness, and L_{min} and W_{min} are the minimum channel length and width, respectively.

2.3. Identifying the Sensitive Transistors

In order to identify the most sensitive transistor in transmission gated based latch, a set of simulation is performed at $\pm 6\sigma_{Vt}$ for each transistor separately. This simulation is done using 45nm low power Predictive Technology Model (PTM) at room temperature and using nominal Vdd (1.1V) condition [8]. The result which is given in Figure 5 shows the change of PD between -6σ_{Vt} and +6σ_{Vt} in each single transistor. According to the result, the most sensitive transistor for DQ_Delay_RR is IN2 (from
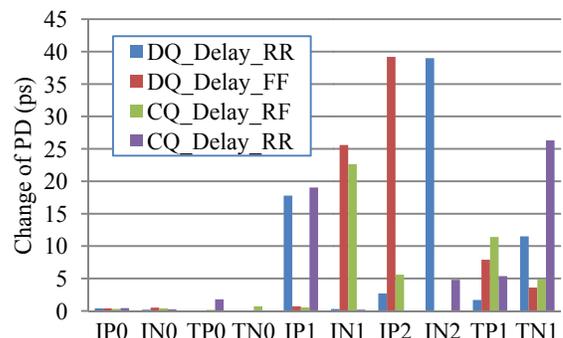
Figure 5: Change of PD from -6$\sigma_{Vt}\sigma$ to +6σ_{Vt}

92.01ps to 131ps) and second most sensitive transistor is IP1 (from 101.2ps to 119ps). IP2/IN1, IN1/TP1 and TN1/IP1 are selected for most sensitive transistors to DQ_Delay_FF, CQ_Delay_RF and CQ_Delay_RR, respectively.

2.4. PD Simulation with Two Most Sensitive Transistors

From previous simulation for indentifying most sensitive transistor on PD, we can select two most sensitive devices which are affected by Vt variations. Table I shows the result of each PDs when two most sensitive transistors have Vt variation at the same time. The simulation is performed with 45nm low power PTM.

Table I: PD simulation with two most sensitive transistors

	DQ_Delay_RR	DQ_Delay_FF	CQ_Delay_RF	CQ_Delay_RR
No Variation	108.8ps	122.5ps	79.66ps	84.88ps
Worst Case	142ps	159.9ps	100.7ps	108.9ps
Best Case	84.7ps	94.42ps	66.67ps	63.4ps
Worst(%)	30.5%	30.5%	26.4%	28.3%
Best(%)	-22.2%	-22.9%	-16.3%	-25.3%
Best to Worst	57.3ps	65.48ps	34.03ps	45.5ps

2.5. Predictive Technology Model (PTM) Modeling

To analyze the effect of technology scaling on the transmission gate based latch, a set of simulation was performed. Both low power model and high performance model for 16nm, 22nm, 32nm and 45nm are used to this simulation. The result is given in Table II. PD under no Vt fluctuation is simulated in this Table. It shows better delay change on high performance model and lower technology model. Also Head Room is calculated for each PTMs:

$$\text{Head Room} = V_{dd_{Nominal}} - [\,Vthn0 + \delta\,(3\sigma_{Vtn})\,]\,(V)$$

At all times, the Head Room of high performance PTM is relatively high compared with the low power PTM. And the Head Room of earlier technology PTM is relatively high compare with the more advanced technology PTM. Therefore the high performance PTM and higher technology PTM shows narrow distribution under Vt fluctuation. Figure 6 illustrates distribution of PD under Vt fluctuation from -$6\sigma_{Vt}$ to +$6\sigma_{Vt}$ for 32nm PTM and 45nm PTM.

3. Simulation Result

In this section propagation delay (PD) simulation was performed to analyze probability of failure. And also the impact of PVT to PD and subthreshold leakage current was discussed. 45nm low power PTM was selected for simulations.
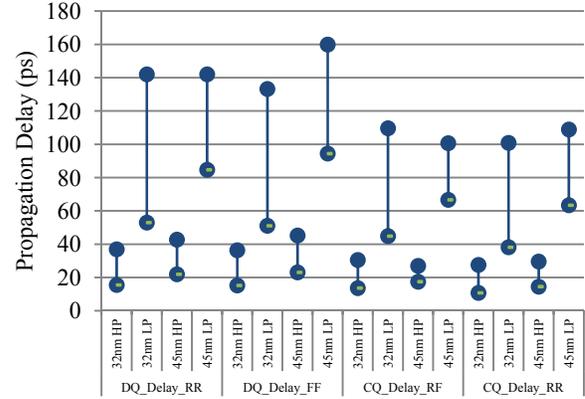
Figure 6: Distribution of PD under Vt fluctuation with 32nm and 45nm PTM

3.1. Uniform sampling

The first step in finding the probability of failure under voltage scaling is to find the most sensitive device(s) for each operation (read 1/0 and write 1/0). Once we identified the most sensitive devices, we uniformly sample from -$6\sigma_{Vt}$ and +$6\sigma_{Vt}$ for each device and measure the delay in HSPICE. Figure 7 shows the PD (DQ_Delay_RR) as a function of changes in the threshold voltages of the two most sensitive devices. Using the measured delay we can reconstruct the CDF for the delay give that the ΔVth for each device has a Gaussian distribution [9]. Let's call the two sensitive devices T1 and T2 we will have

$$\Delta Vth_{T1} \sim N(0, \sigma_{T1})$$
$$\Delta Vth_{T2} \sim N(0, \sigma_{T2})$$

$$P_e = P[T_{delay} > T_{Max}]$$
$$= \iint\limits_{R_{failure}} G(Vth_{T1}, Vth_{T2})dVth_{T1}dVth_{T2} \quad (1)$$

TABLE II: The effect of each predictive technology model on propagation delay

	16nm HP	16nm LP	22nm HP	22nm LP	32nm HP	32nm LP	45nm HP	45nm LP
Vthn0 (V)	0.47965	0.68191	0.50308	0.68858	0.49396	0.63	0.46893	0.62261
Vthp0 (V)	-0.43121	-0.6862	-0.4606	-0.63745	-0.49155	-0.5808	-0.49158	-0.587
Nominal Vdd (V)	0.7	0.9	0.8	0.95	0.9	1.0	1.0	1.1
Vthn0 + $\delta(3\sigma_{Vtn})$ (V)	0.65645	0.85871	0.63163	0.81713	0.58236	0.7184	0.53178	0.68546
Head Room (V)	**0.04355**	**0.04129**	**0.16837**	**0.13287**	**0.31764**	**0.2816**	**0.46822**	**0.41454**
DQ_Delay_RR (ps)	18.55	85.22	20.91	88.47	22.67	77.08	30.97	108.8
DQ_Delay_FF (ps)	19.71	98.06	20.25	89.09	21.5	75.21	33.01	122.5
CQ_Delay_RF (ps)	17.27	85.3	18.14	76.15	19.49	63.99	21.61	79.66
CQ_Delay_RR (ps)	14.22	63.86	15.59	66.92	16.92	58.15	21.43	84.88

where $R_{failure}$ is the region in the plane (Figure 7) in which $T_{delay}(\Delta Vth_{T1}, \Delta Vth_{T2}) > T_{Max}$ and is given by equation 2 and $G(Vth_{T1}, Vth_{T2})$ is given by equation 3 which is a two-dimensional Gaussian function with zero mean and standard deviations of $\sigma_{vth\,T1}$ and $\sigma_{vth\,T2}$.
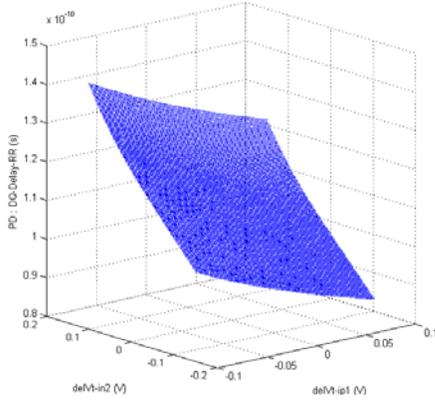
Figure 7: Uniform distribution for PD (DQ_Delay_RR) on Vt variations

$$R_{failure} = \{\Delta Vth_{T1}, \Delta Vth_{T2} | T_{delay}(\Delta Vth_{T1}, \Delta Vth_{T2}) > T_{Max} \quad (2)$$

and

$$G(Vth_{T1}, Vth_{T2}) = \frac{1}{2\pi\sigma_{VthT1}\sigma_{VthT2}} exp\left(-\left(\frac{\Delta Vth_{T1}^2}{2\sigma_{VthT1}^2} + \frac{\Delta Vth_{T2}^2}{2\sigma_{VthT2}^2}\right)\right) (3)$$

For a given T_{Max} and supply voltage based on equation 1, one can find probability of failure for each operation. It is important to note that this approach does not rely on approximating the delay distribution nor limited to the number of sigmas considered for each transistor. Figure 8 shows the probability of failure for DQ_Delay_RR for three different voltages. As it is shown in this figure, at more aggressive frequency targets, probability of failure is very sensitive to change in supply voltage due to tighter negative margins.
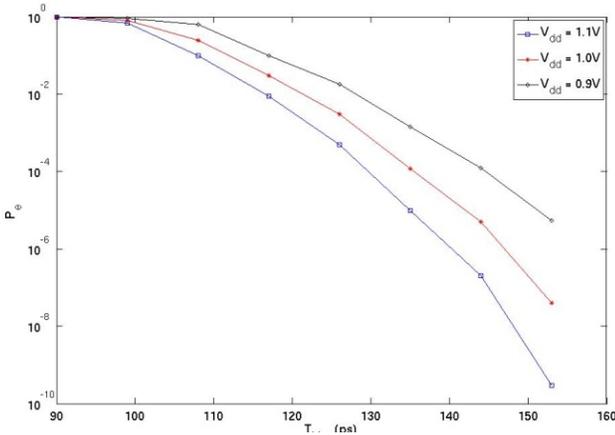
Figure 8: Probability of failure for different Vdd

Cumulative Distribution Function (CDF) is also illustrates in Figure 9. This plot shows the distribution of four different reconstructed PDs. According to CDF plot, Data-to-Q (DQ) PD has larger distribution compare with CLK-to-Q PD.
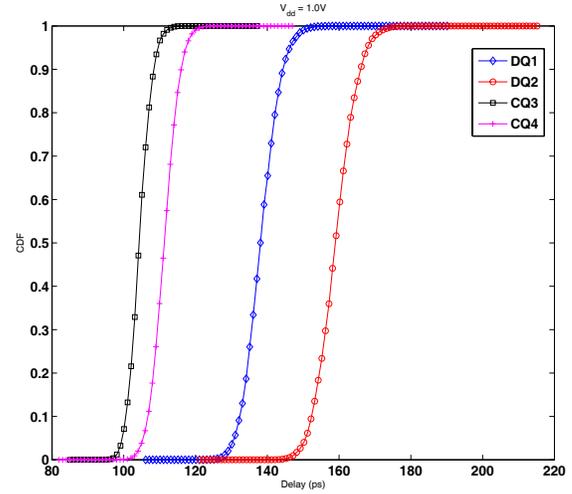
Figure 9: Cumulative distribution function of PD

3.2. Impact of temperature variation on PD

The values of PD taking into account temperature is simulated. The PDs increase over 3 times with temperature ranging from -30°C to 130°C.

Figure 10 shows the impact of Vt variation (±6σ_Vt) on PD. Each bar is calculated with summation of percentages which are at best case and at worst case. For example at 70°C, DQ_Delay_FF is 171.9ps with default Vt. It changes to 230.6ps (+34.2%) when it has +6σ_Vt fluctuation. And It change to 128.2ps (-25.4%) when it has -6σ_Vt fluctuation. The plot shows, at higher temperature, Vt variation can effect to PD more [10][11].
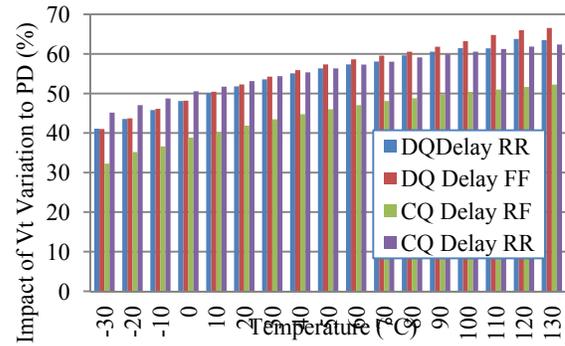
Figure 10: The impact of Vt variation on each temperature to PD

3.3. Impact of Vdd scaling on PD distribution

To analyze the impact of supply voltage on PD, a set of simulation is performed with Vdd set at 0.9V, 1.0V, and 1.1V. The PD is decreasing exponentially when supply

voltage is increasing. Figure 11 illustrates Probability Density Function (PDF) of DQ_Delay_RR which shows the impact of Vt variation ($\pm 6\sigma_{Vt}$) on PD. As Vdd increases, both the PD mean and standard deviation decrease, rendering the latch less sensitive to process variations.
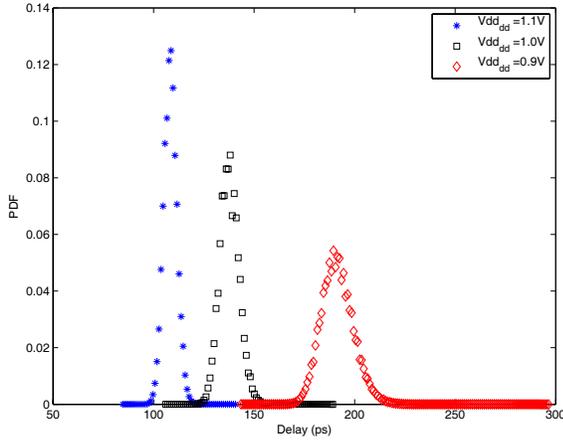
Figure 11: Probability density function of DQ_Delay_RR

3.4. Subthreshold leakage current

The Subthreshold leakage current (I_{SUB}) is the current flowing between drain and source of a device under "off" state [12].

$$I_{SUB} = K_1 \cdot W \cdot e^{\frac{-Vt}{nV_\theta}} \cdot [1 - e^{\frac{-Vdd}{V_\theta}}]$$

where K_1 and n are the experimentally derived constants. W is the gate width, Vt is the threshold voltage. Vdd is the supply voltage, and V_θ is the thermal voltage which increases linearly as the temperature is increased.

When D signal is steady state as logic high or logic low, the transmission gate based latch is standby (off) state. If D is logic low, there are three subthreshold leakage currents through 3 different transistors (IN0, IN2 and IP1). And if D is logic high, subthreshold leakage currents flow through transistor IP0, IP2 and IN1.

According to the simulation, total subthreshold leakage current can be changed on variance of temperature and supply voltage, respectively. At room temperature (25°C) and 1.1V for supply voltage, total I_{SUB} is 20.96pA when D is logic low, and 24.42pA when D is logic high. Total I_{SUB} is increasing exponentially when temperature and supply voltage are increasing. For example, at 1.1V for supply voltage and 130°C for temperature total I_{SUB} is 423.13pA when D is logic low, and 480.23pA when D is logic high. And, at 1.5V for supply voltage and 25°C for temperature total I_{SUB} is 47.13pA when D is logic low, and 60pA when D is logic high.

4. Conclusions

We have presented a study of the impact of variability to the transmission gate based latch. The study shows that Vt fluctuation due to RDF and PVT are important sources of variability. For a specific performance target, the designer can tradeoff error tolerance versus variability. To improve performance and reliability in the presence of variability, it is important to have accurate and fast modeling. The proposed modeling in this paper can be used to decide on the acceptable level of error tolerance which is allowable by the application and the system design.

5. References

[1] Yin Jinhang, Lu Wenxi, Xin Xin, Zhang Lei, "Application of Monte Carlo Sampling and Latin Hypercube Sampling Methods in Pumping Schedule Design during Establishing Surrogate Model", ISWREP, 2011.

[2] Eduardo Saliby, Flavio Pacheco, "An Empirical Evaluation of Sampling Methods in Risk Analysis Simulation: Quasi-Monte Carlo, Descriptive Sampling, and Latin Hypercube Sampling", Proceedings of the 2002 Winter Simulation Conference, 2002.

[3] Zarko P. Barbaric, Miroslav D. Lutovac, Ivan D. Dokic, "Analyses of Probability Density Function of Displacement Signal for Laser Seeker Systems", TELSIKS, 2011.

[4] Jan M. Rabaey, Anantha Chandrakasan, Borivoje Nikolic, Digital Integrated Circuits, Prentice Hall, 2003.

[5] Amin Khajeh Djahromi, Ahmed M. Eltawil, Fadi J. Kurdahi, Rouwaida Kanj, "Cross Layer Error Exploitation for Aggressive Voltage Scaling", ISQED, 2007.

[6] Y. Taur, T. H. Jayatissa, Fundamentals of Modern VLSI Devices, Cambridge University Press, 1998.

[7] S. Mukhopadhyay, H. Mahmoodi, K. Roy, "Statistical Design and Optimization of SRAM Cell for Yield Enhancement", ICCAD, 2004.

[8] http://ptm.asu.edu/

[9] Saibal Mukhopadhyay, Hamid Mahmoodi, Kaushik Roy, "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS", IEEE Trnasactions on Computer-Aided Design of Integrated Circuits and Systems, VOL. 24, NO. 12, 2005.

[10] A. Golda, A. Kos, "Temperature Influence on Power Consumption and Time Delay", Digital System Design, 2003.

[11] Leon Chang, Khoa Vo, John Berg, "A Simplified Model to Predict the Linear Temperature Coefficient of a CMOS Inverter's Delay Time", IEEE Transactions on Electron Devices, Vol. ED-34, No. 8, 1987.

[12] Mohammed Aftab Alam, Diganta Das, Michael H. Azarian, "Influence of Molding Compound on Leakage Current in MOS Transistors", Components, Packaging and Manufacturing Technology, 2011.