Variability-Aware Modeling of Pulse Latches

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

CECS Technical Report 13-15
Oct 29, 2013
Abstract—In this paper we study the impact of variability on the pulse latch. Pulse (or pulsed) latch is crucial element of the current SoC designs for both delay and power consumption. The threshold voltage (Vt) fluctuations due to Random Dopant Fluctuation (RDF) and Process, Voltage, and Temperature (PVT) effects on probability of failure are discussed. We propose a modeling methodology which is not tied to a specific topology and scaleable in contrast to techniques such as Monte Carlo simulation. To estimate the probability of failure, we uniformly sample the variation space and reconstruct the Probability Density Function (PDF) for each delay path. Then we statistically calculate the lower bound of failure.

I. INTRODUCTION

Flip-flops and latches are fundamental building blocks of sequential digital circuits. Chip-wide, the total number of transistors in these circuits can reach hundreds of millions of operating in certain clock speed range. As a result, the contribution of these components to chip area, performance and power can be significant. Pulse latches have been proposed as means of reducing power consumption [5]. On the other hand, deep sub-micron technologies exhibit significant inter and intra die process variations. Given the relatively short pulse width of pulse latches, there is increased probability of failure due to excessive delay and/or leakage. Due to the effects of process variation, designers need to guarantee both correct functionality as well as timing checks on the fabricated chip. Among the different sources of random intra-die variations, the most significant one is the threshold voltage (Vt) variation due to RDF. One of the most popular methods of assessing designs under PVT variations is Monte Carlo simulation. Monte Carlo simulations are a class of computational algorithms that rely on repeated random sampling to compute their 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. Latin Hypercube or Sobol Sampling are used to make up these weakness [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 PDF of the delay [3]. Having the PDF of the delay, we can find a relationship between the probability of error and PVT regardless of number of sigmas that we want to consider. More details are given in the simulation setup section.

II. SIMULATION SETUP

A. Explicit Pulsed Latch Simulation Scheme

To narrow the transparency window of the latch, pulsed latch is designed. It is clocked with short pulses generated locally from the global clock signal [4]. Intel’s version of Explicit Pulsed Latch is shown in Figure 1 [5]. Additional benefit of this design is low power consumption due to the

![Figure 1. Intel’s explicit pulsed latch [5]](image-url)
common clock signal generator (pulse generator) and a simple structure of the latch. This power can be traded off for speed and/or reliability. The pulse generator used in this design employs the principle of re-convergent fan-out with non-equal parity of inversion in order to obtain the desired short clock pulse. Due to its simplicity, this latch is generally faster and more energy-efficient than traditional latches.

In our simulation, we explored a range of initial condition as follows: rising time is swept from 50ps to 92.75ps for CLK and from 80ps to 127.7ps for D.

Figure 2 illustrates the simulation flow chart. The simulation is performed in three stages. At the first stage, the input signal is generated using chain inverters with certain PVT condition. The realistic input signal waveforms are dumped out and saved. Using the dumped waveforms, a set of simulation is performed with uniform sampling which is in probability domain. Then we reconstruct the probability density function (PDF). Having the PDF, we can find the probability of error regardless of number of sigmas that we want to consider. By separating simulation into three stages, simulation time can be decreased. And by using uniform sampling simulation instead of Monte Carlo simulation, simulation is more accurate and faster.

There are three measurement parameters which are pictorially represented in Figure 3.

- PW : Pulse Width
- WTF(Write Time Falling) : When D and Qbar are logic high, rising edge of PULSE change the Qbar to fall
- WTR(Write Time Rising) : When D and Qbar are logic low, rising edge of PULSE change the Qbar to rise

As a reasonable tradeoff between power and reliability, we set the pulse width to about 1/5 of clock width. When we use 500MHz clock cycle, for example, the target pulse width is therefore set at 200ps.

B. Vt Fluctuations due to RDF

The Vt fluctuations due to Random Dopant Fluctuation (RDF) can be considered as zero-mean Gaussian random variables [6]. Vt fluctuations in transistors in pulse latch are considered as independent Gaussian random variations (δVt) with mean=0. The standard deviation of the δVt which can be denoted in σVt is given by [7]:

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

where σVt0 is the σVt for minimum sized transistor and it is given by [8]:

$$\sigma_{Vt0} = \frac{q T_{ox}}{\varepsilon_{ox}} \sqrt{\frac{N_d W_d}{3 L_{min} W_{min}}}$$

where N_d 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.

C. Identifying the Sensitive Transistors

In order to identify the most sensitive transistor in pulse latch, a set of simulation is performed at ±6σVt for each transistor separately. This simulation is using 45nm low power Predictive Technology Model (PTM) at room temperature and using nominal Vdd (1.1V) condition [9]. According to the...
simulation, there are two sensitive transistors which can effect to the change of pulse width. Transistor IN0 and IN2 shows the change from 162ps to 452ps and from 182.5ps to 353.9ps respectively. The other transistor can be effected much less than these two sensitive transistors relatively. For write time, the result which is given in Figure 4 shows the change of Clock-to-Qbar delay between -6σVt and +6σVt in each single transistor. According to the result, the most sensitive transistor for WTF is IN4 (from 114.6ps to 59.95ps) and second most sensitive transistor is TP0 (from 92.92ps to 110.1ps). And TN0/IP4 are selected for most sensitive transistors to WTR.

D. Calculating the Headroom

To analyze the effect of technology scaling on the pulse latch, a set of simulation was performed. Both low power model and high performance model for 32nm and 45nm are used to this simulation. The result is given in Table I. Pulse Width and Write Time regardless 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 technology node:

\[
\text{Head Room} = V_{dd\text{Nominal}} - [V_{thn0} + \delta (3\sigma_{V_{thn}})] (V)
\]

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 5 illustrates distribution of Write Time under Vt fluctuation from -6σVt to +6σVt for various PTMs.

<table>
<thead>
<tr>
<th></th>
<th>32nm HP</th>
<th>32nm LP</th>
<th>45nm HP</th>
<th>45nm LP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vthn0 (V)</td>
<td>0.49396</td>
<td>0.63</td>
<td>0.46893</td>
<td>0.62261</td>
</tr>
<tr>
<td>Vthp0 (V)</td>
<td>-0.49155</td>
<td>-0.5808</td>
<td>-0.49158</td>
<td>-0.587</td>
</tr>
<tr>
<td>Nominal Vdd (V)</td>
<td>0.9</td>
<td>1.0</td>
<td>1.0</td>
<td>1.1</td>
</tr>
<tr>
<td>Vthn + δ(3σVthn) (V)</td>
<td>0.577815</td>
<td>0.713855</td>
<td>0.528555</td>
<td>0.682235</td>
</tr>
<tr>
<td>Head Room (V)</td>
<td>0.322185</td>
<td>0.286145</td>
<td>0.471445</td>
<td>0.417765</td>
</tr>
<tr>
<td>PW (ps)</td>
<td>57.2</td>
<td>203</td>
<td>64.52</td>
<td>219</td>
</tr>
<tr>
<td>WTF (ps)</td>
<td>26.27</td>
<td>95.42</td>
<td>134.7</td>
<td>601.7</td>
</tr>
<tr>
<td>WTR (ps)</td>
<td>25.51</td>
<td>95.44</td>
<td>39.73</td>
<td>129.2</td>
</tr>
</tbody>
</table>

Table I. The head room and effect of each PTM on delay

III. Simulation Result

A. 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 -6σVt and +6σVt for each device and measure the delay in HSPICE. Figure 6(a) shows the PW 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 given that the ΔVth for each device has a Gaussian distribution [10]. Let’s call the two sensitive devices T1 and T2 we will have

\[
\begin{align*}
\Delta V_{th,T1} &\sim N(0, \sigma_{T1}) \\
\Delta V_{th,T2} &\sim N(0, \sigma_{T2}) \\
\end{align*}
\]

Where \(R_{failure}\) is the region in the plane (Figure 6(b)) in which \(T_{PW}(\Delta V_{th,T1}, \Delta V_{th,T2}) \leq T_{WT}(\Delta V_{th,T1}, \Delta V_{th,T2})\) and is given by equation 2 and \(G(V_{th,T1}, V_{th,T2})\) is given by equation 3 which is a two-dimensional Gaussian function with zero mean and standard deviations of \(\sigma_{vth,T1}\) and \(\sigma_{vth,T2}\).

\[
R_{failure} = (\Delta V_{th,T1}, \Delta V_{th,T2}) | (\Delta V_{th,T1}, \Delta V_{th,T2}) \leq T_{WT}(\Delta V_{th,T1}, \Delta V_{th,T2}) \quad (2)
\]

and

\[
G(V_{th,T1}, V_{th,T2}) = T_{PW}(V_{th,T1}, V_{th,T2}) - T_{WT}(V_{th,T1}, V_{th,T2}) \quad (3)
\]

For a given \(T_{WT}\) 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 5](image1.png)

Figure 5. Distribution of write time under Vt fluctuation with 32nm and 45nm PTM

![Figure 6](image2.png)

Figure 6. (a) Uniform distribution for pulse width on Vt variations  
(b) Region of failure on the distribution of delay
B. Impact of Vdd Scaling on PW and WT Distributions

To analyze the impact of supply voltage on delay, a set of simulation is performed with Vdd set to 0.9V, 1.0V, and 1.1V. As expected, the delay is decreasing exponentially when supply voltage is increasing. Figure 7 and figure 8 illustrate Probability Density Function (PDF) of pulse width and write time respectively. As Vdd increases, both the mean and standard deviation decrease, rendering the latch less sensitive to process variations.

Table II shows the probability of error (Pe) for the pulse latch on three different supply voltages. Our target was to obtain lower bounds on the failure rates of the pulse latch at each voltage level. To do so, we first simulated the extreme conditions of the pulser’s inverter chain, under which that chain’s delay is maximum between $-6\sigma_{VT}$ and $+6\sigma_{VT}$. The clock width is then set to a value larger than that delay by a sufficient margin to ensure that PW is always positive. It is important to note that since we are varying two parameters in the $\pm 6\sigma$, the smallest probability of failure that we can calculate in this experiment space will be $\sim 10^{-18}$ and any Pe less that this value will be assumed zero. If Pe smaller than $10^{-18}$ is of interest, one can increase the experiment space by considering higher number of sigma. At 1.1V which is nominal voltage of 45nm low power technology, Pe is 0 given the above discussion. At lower supply voltage, Pe is very sensitive to change, and is exponentially increased as lower supply voltage. It is caused by tighter negative margins of voltage as quantified in Table I.

As discussed above, these failure rates are pseudo-independent of the clock width and can be thought of in a manner similar to hold time violation in traditional sequential logic. Given that, it is clear from the table that scaling Vdd down to 0.9 V in this case renders the pulse latch unusable unless significant effort is expended in correcting the failures at runtime (e.g. Razor [11]).

IV. CONCLUSIONS

As technology scales down, operation at lower power supplies with less power consumption has become the utmost

| Table II. Probability of error with supply voltage scaling |
|---------------------------------|-------------------|-------------------|
| Vdd (V) | Pe | |
| 1.1V | 0 ($\leq 10^{-18}$) | 27.059e-12 | 3.205e-04 |

Figure 7. Probability density function of pulse width

Figure 8. Probability density function of write time

priority. However lower power can address increased chance to error of system. We have presented a study of the impact of variability to the pulse latch. The study shows that Vt fluctuation due to RDF and scaling of supply voltage are important sources of variability. To have more refined model, the designer can tradeoff probability of failure versus variability. In our proposed modeling experiments, we can assume a lower bound of failure in certain specifications. It can be used to decide on acceptable level of error tolerance which is allowable by the application and the system design.

REFERENCES


