12. LECTURE 12: CIRCUIT IMPLICATIONS OF NBTI

12.1 Outline

In the previous lectures we observe the temporal degradation of a PMOS device that is attributed to NBTI. Then we developed physical models to explain the experimental data and we explained why the Nit vs. time slope is 1/4 or 1/6. The scope of this lecture is to understand the implications of PMOS degradation due to NBTI. We study the effects for both logic and memory circuits and finally, we present some solutions to overcome NBTI degradation during or after design time.

12.2 NBTI Implications on Logic Gates

For logic gates, delay is one of the most important parameters. In this section, we derive the delay expression for logic gates based on Sakurai’s model that is evidently a function of $V_{th}$. Then we study how the delay behaves due to NBTI ($V_{th}$ shift).

12.2.1 Transistor Model (Sakurai’s Model)

Firstly, we have to derive a formula for the delay of a gate. In this report we use a-Power model [] and we apply it for an inverter. For nanoscaled transistors the use of quadratic transistor model is not accurate. Hence the use of Sakurai’s model matches better the $I-V$ characteristics of a transistor and henceforth the delay of a circuit. The equations that describe the transistor for different operation modes are given below

\[ I_D = 0, \ (V_{GS} < V_t) \]

12.1

a) Cutoff Region

b) Linear Region
\[ I_D = I_{D,sat} \left( 2 - \frac{V_{DS}}{V_{D,sat}} \right) \frac{V_{DS}}{V_{D,sat}}, \quad (V_{DS} < V_{D,sat}) \]  \hspace{1cm} (12.2)

c) Saturation region

\[ I_D = I_{D,sat} = \left( \frac{W_{eff}}{L_{eff}} \right) B (V_{GS} - V_t)^n \quad (V_{DS} \geq V_{D,sat}) \]  \hspace{1cm} (12.3)

\[ V_t = V_{t0} - \gamma V_{BS} \quad \text{and} \quad V_{D,sat} = k (V_{GS} - V_t)^m \]  \hspace{1cm} (12.4)

Where and \( n, m, K \) and \( B \) are constants which describe the short channel effects in an empirical manner. Note that if \( n=2, m=1, K=1 \) and \( B=1/2 \beta \) then we have long channel \( I_{DS}-V_{DS} \) characteristics. In Fig. 12.1 we show how well this model matches with experimental results. It is worthwhile to observe that linear increase of \( V_{GS} \) increases linearly the \( I_{DS} \) current. This is a characteristic of short channel effects (SCE) or more specifically velocity saturation.

![Fig. 12.1 Fig.1: Matching experimental I-V data with the Sakurai’s model.](image)
12.2.2 Inverter Propagation Delay

In this subsection we show the connection between the device and circuit level that is, through which parameter NBTI, affects vital circuit parameters. In Fig. 12.2 we show the equivalent circuits that are useful for the derivation of rising and falling propagation delay ($t_{pLH}$, $t_{pHL}$). Next, we will calculate the falling gate delay only but the procedure for rising propagation delay is similar.

In digital circuits there are many ways to define the propagation delay but the conventional way is to measure the time between 50% input to 50% output (see Fig. 12.2). Generally, $t_{pLH}$ and $t_{pHL}$ are different because the mobility of a NMOS is different than that of a PMOS.

To begin with, we divide the input into three regions because the transistor current that charges the capacitor depends on input ($V_{GS}$). In addition, we have normalized the variables in order to obtain more compact equations. For each region we apply Kirchhoff’s current equation and thus three differential equations are obtained. The initial condition of each differential equation is the final solution of the previous region.

Fig. 12.2: Rising-Falling gate delay and the corresponding equivalent circuits for delay analysis.
Region 1. Starting with the first region the gate voltage increases linearly with time (small \(V_{GS}\)) and \(V_{DS}\) almost constant and high value, the transistor operates in saturation region. Hence the equation that describes the circuit during this region is:

\[
\frac{dV_0}{dt} = -I_{D,sat} = \left(\frac{t/t_T - v_T}{1 - v_T}\right)^n \left(\frac{1 + \lambda v_0}{1 + \lambda}\right) (v_0(t = t_T = v_T) \quad 12.5
\]

Region 2. For that region the transistor still operates in saturation but the input signal is constant. The differential equation that describe the circuit for this region is given by

\[
\frac{dV_0}{dt} = -I_{D,sat} = \left(\frac{1 + \lambda v_0}{1 + \lambda}\right) (v_0(t = t_T = v_T) = v_1) \quad 12.6
\]

Region 3. For the last region the transistor operates in linear region because the input signal is very high but \(V_{DS}\) is very low. \(V_{DS}\) decreases because the capacitor discharges. The differential equation that describe the circuit for this region is given by

\[
\frac{dV_0}{dt} = -I_{D,lin} = -\left(2 - \frac{v_0}{v_{D0}}\right) \frac{v_0}{v_{D0}} \left(\frac{1 + \lambda v_0}{1 + \lambda}\right) (v_0(t = t_{D0}) = v_{D0}) \quad 12.7
\]

Solution. All the above differential equations are first order and can be easily solved. Note that the initial condition of the next equation is the final point of the previous one. This is necessary condition to validate that the solution will be continuous. The delay of an inverter is obtain from the solution of all the above equations is given by (Note that the is a function of \(V_{th}\) and this is the connecting ling with the theory we have developed in the previous lectures)

\[
t_d = t_{pHL} = t_r \left(\frac{1}{n+1} - \frac{1 - u_T}{n+1}\right) \left(\frac{u_v - u_T}{n+1}(1 - u_T)^n\right) + \frac{C_0V_{DD}}{I_{D0}} f(V_{DD}, V_t) \quad 12.8
\]
12.2.3 Delay in Standard Cells

From the previous analysis it is become evident that this methodology can be applied to any other digital gate even to a whole library (set of standard digital gates). In Table 1 we show the degradation in delay that some cells of a library suffer. It worthwhile to mention that some cells degrades more the 25%. Before leave this subsection about the delay of standard cells, we have to note that the inputs to logic gates in a real circuit are not constants. They have their switching activity (duty cycle) that depends on the switching activity of the output of previous gates. Hence, we should remember that the delay degradation is also a function of duty cycle as it is shown in Fig. 12.3.

One important caveat regarding the numbers tabulated above. In the first generation Alam-Roy SPICE model, relaxation was not accounted for, neither did they periodic vs. non-periodic stress and variation in duty cycle during circuit operation. Therefore, the results were somewhat more pessimistic that one would obtain in practice. Since then SPICE models have improved considerably. Groups lead by Kevin Cao and Sachin Sapatnekar now used SPICE model that captures some aspect of relaxation in the SPICE model. In effect, the capture the results from numerical simulation in effective model with technology-specific relaxation parameters. For details, see IEEE TDMR, 7(4), 2007. P. 509. And IEEE Trans. VLSI, 18(2), p. 173, 2010.

<table>
<thead>
<tr>
<th>Logic Cell</th>
<th>fanin</th>
<th>Delay (ps)</th>
<th>Degradation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>t=0</td>
<td>3 years</td>
</tr>
<tr>
<td>INV</td>
<td>1</td>
<td>13.77</td>
<td>16.77</td>
</tr>
<tr>
<td>NAND</td>
<td>2</td>
<td>16.86</td>
<td>19.88</td>
</tr>
<tr>
<td>NAND</td>
<td>3</td>
<td>19.57</td>
<td>22.45</td>
</tr>
<tr>
<td>NOR</td>
<td>2</td>
<td>17.26</td>
<td>21.89</td>
</tr>
<tr>
<td>NOR</td>
<td>3</td>
<td>23.80</td>
<td>30.19</td>
</tr>
</tbody>
</table>
12.3 NBTI-aware Logic Design

Having developed the connecting bridge between devices and simple digital standard cells, the next question is how to design large circuits with many gates that will work properly for many years. One simple but rather not efficient way to overcome this obstacle is to make the circuit faster (see Fig. 12.4) but that demands the use of larger transistors. This overdesign waste not only area but also power. Hence, the exigency for a design methodology is essential.

Since area is very important then we can write an optimization problem with objective function the area for which we demand minimization and constrains the delay of each gate and the lower and upper limits for the transistor size (these are given by the technology).

This optimization problem is given below and it is based on the Lagrangian relaxation:

\[
\begin{align*}
\text{Minimize} & \quad \sum_{i=1}^{n} \alpha_{N,i} x_{N,i} + \alpha_{P,i} x_{P,i} \\
\text{Subject to} & \quad f_p \leq A_0, r_p \leq A_0, \quad j \in P \quad (\text{all outputs}) f_j + D_{r,j} \\
& \quad \leq r_i, r_j + D_{f,i,j} \leq f_i, i = 1, ... n
\end{align*}
\]

In Fig. 12.5 we show the flowchart with the algorithm that solves the problem above.
Fig. 12.5  (a) IDS-VDS current curves for NMOS and PMOS transistors of an inverter, (b) Voltage Transfer Characteristic (VTC) of an inverter, (c) circuit schematic diagram of an SRAM cell, (d) Butterflies diagrams of an SRAM cell operating in hold mode
12.4 **NBTI Implications on SRAM cells**

Memories and more specifically SRAM cells, comprise the second large category of digital circuits that affected by NBTI. In this subsection we will analyze how NBTI affects the stability of a six transistor (6T) cell (Fig. 12.5 c). Stability is the most important characteristic for a cell because due to process variations and NBTI many cells of a memory array fail to operate properly. The stability of a cell is measured by the quantity static noise margin (SNM). SNM can be estimated by the diagonal of the inscribed square between the butterflies (Fig. 12.5 d). The butterflies can be formed from inverter’s VTC (see Fig. 12.5 a,b)

An SRAM cell has three operational modes: (a) read, (b) write and (c) hold. In the next subsections we will analyze the effect of NBTI to each of these modes.

Due to process variations and NBTI some cells do not operate properly to one or more of these modes. Hence, the probability to fail a cell is given by

\[
\frac{dN_{IT}}{dt} = k_F(N_0 - N_{IT}) - k_RN_H(0)N_{IT}
\]

12.10
12.4.1 **Read Failure Probability**

During read mode we try to read the stored data in the cell. NBTI is a serious problem because during the previous mode (hold) one of the two PMOS transistors is degraded. Note that the hold mode lasts a lot of time. Due to this degradation one of the two VTCs shifts and if it shifts enough then the cell loses its stability (see Fig. 12.7(a)-(d)).

In addition, it is evident that the degradation depends on the switching activity that is, how often we try to access for read a specific cell.
Fig. 12.7  Butterflies during read mode (a) t=0sec (b) and (c) SNM degradation due to
NBTI without read failure, (d) read failure, read failure probabilities (e) due to process
variations and (f) due to switching activity.

12.4.2  Write Failure Probability

In the previous subsection we explained that the probability to fail a cell during
read mode increases with time. In contrast, the write margin improves that is, the
probability to fail a cell during write mode decreases (Fig. 12.8). This is attributed to the
fact that due to NBTI the PMOS transistor becomes weaker (increased $V_{th}$) and less able
to keep the values that is stored in the cell. Note that we expect this behavior intuitively
because for an SRAM cell read and write operations are controversial and hence the
improvement of one implies the melioration of the other.

![Write Failure Probability Graph]

Fig. 12.8  Write failure probability as a function of time. We observe that the write margin
improves with time.

12.4.3  SRAM Array Failure

An SRAM array is organized in rows and columns of cells as it is depicted in
Fig. 12.9. If one cell of a column fails (read, write or hold failure, see above) then this
column does not operate properly. Lets define the failure probability of a column

$$P_{col} = 1 - (1 - P_F)^N$$  \hspace{1cm} 12.11
where $P_F$ is defined above and it is the probability to fail a cell from the column. In addition the probability to fail the memory array, which is called also as yield, is given by

$$P_{MEM} = \sum_{i=N_{RC}+1}^{N_{COL}} \binom{N_{COL}}{i} P_{COL}^i (1 - P_{COL})^{N_{COL} - i}$$  \hspace{1cm} 12.12

Fig. 12.9 Organization of a memory array. Redundant columns are added to replace the columns that fail to operate properly.
12.4.4 Solutions

In this subsection we present solutions that help the designer to improve the yield of an SRAM array. The first solution is the use of additional columns that replace the columns that fail to operate properly (Fig. 12.9). The second solution is the use of body bias. Body bias gives us the ability to control the threshold voltage (dynamically). Hence, reverse bias reduces read and hold failures but forward bias reduces write failures (Fig. 12.10).

The last solution is called self repair and is based on the measurements of a leakage monitor which measure the NBTI degradation and calibrate the body bias dynamically. The circuit diagram of this technique is shown in Fig. 12.11.
12.5 Post-Silicon Techniques

Odometer is a sensor that scales and integrates the total miles traveled by a vehicle. Invented by the Mormon Pioneers as they traveled to Utah, the odometer provides information regarding distance traveled and thereby – as the buyers of used cars know – indirectly providing information regarding the wear/tear of the car. A similar sensor that measures the wear/tear of an IC would allow preventive measures either by reducing the clock speed or alerting for replacement. These new sets of sensors are called silicon odometers and will be discussed next.

12.5.1 Silicon Odometer: Temporal $I_{DDQ}$

$I_{DDQ}$ is one of the most important post-silicon techniques. This technique is based on monitoring the leakage current. An example is given in Fig. 12.12. In Fig. 12.13 (a) we present an inverter chain implementation and how we measure the leakage current. It is important to mention that we should flip the state of the chain in order to measure the leakage current. Initially we degrade the PMOS transistor and after flipping we measure their leakage which is a signature of NBTI degradation (see Fig. 12.13(b)).

![NOR2 Leakage](image)

Fig. 12.12 (a) Input vector dependent leakage component, (b) PMOS $V_{th}$ and NOR2 leakage $I_{DDQ}$ reduction over time.
12.5.2 PLL-Sensors for Degradation

Checking for IDDQ requires additional circuits: wouldn’t it be better if we could use the existing circuits within the IC to do the checking? Inside each IC, we have a clock that synchronizes the activities of the circuit. This on-chip clock is not very reliable because the frequency could change depending on the temperature or variability of parameters in the transistor making up the clock. To correct for rather deplorable situation where the ‘drill-sergent’ himself is not very reliable, one needs a off-chip clock quartz clock to keep the ‘drill-sergent’. Fortunately, the on-chip clock can be made to go faster or slower by tuning the voltage up or down. The difference of where the system-clock wants to be (based on temperature, its own parameters, degradation, etc.) generates a control signal, V_cnt (see Fig. xx) and the voltage grows until the two clocks are brought back in sync. Therefore, just like Iddq, V_cnt is an indicator for the state of the degradation of the system. In addition, you get the information for free. Anything that is this good cannot be fully true! If you think about it, V_cnt only knows about the clock and it has no idea what the rest of the circuit is doing. So long the situation of the on-clock chip reflects the stress seen by the rest of the circuit – generally true for ambient temperature, inter-chip variability – etc. we are pretty safe, otherwise Iddq will give better results.

Fig. 12.13 (a) Inverter chain implementation for $I_{DDQ}$ measurement, (b) Leakage current for the inverter chain. Note which transistors are degraded by NBTI and which leakage we measure.
12.5.3 Clock-Stretching

If neither Iddq or PLL seems reasonable, there are other options. In one option, the Stanford group suggests giving the circuit a guard-band good only for a short period of time (say a week) under worst case degradation condition and then check periodically to see if the circuit is beginning to need this extra band. When they do, additional delay budgets are given by slowing down the circuit, or by reducing the voltage. The real trick is design the circuit that checks if the IC needs the extra budget. Moreover, it is important that the checker circuit itself is not subjected to significant degradation and NOR based logic that has more NMOS transistors than PMOS transistors are preferable. Of course, in circuits with high-k gate dielectrics with NBTI and PBTI concerns, NOR does not offer such solace either and the checker circuit might be susceptible to degradation as well!
12.6 System Design

In this section we give an example of a CAD tool (Computer Aided Design) proposed by Texas Instruments. The flowchart of this tool is shown in Fig.14. According to this methodology, in the first level (device level) degradation data are extracted. Then a device is chosen and this is used for the next level (circuit level). In the next step the circuit is designed using iterative trials until to meet the specifications. In order to meet the specifications the circuit is verified using SPICE simulations. Note that the models used for these simulations incorporate NBTI degradation.

![Flowchart for system design considering device degradation (NBTI).](image)

Recently, there have been considerable progress in understanding the impact of NBTI in truly complex circuits. For example, B. Vaidyanathan and his colleagues (p. 164, 2009 IEEE/ACM ICCAD Proc.) shows an impressive analysis of Pipeline performance analysis and tuning in combined presence of statistical time-dependent NBTI degradation as well as time-zero process induced variability.
12.7  **Homework**

**NBTI-Aware Pipeline Performance**

In general, the system level problem require detailed numerical simulation using SPICE equivalent circuits. A recent paper by B. Vaidyanathan, A. Oates, and Y. Xie, Proc. of 2009 IEEE/ACM Int. Conference on Computer-Aided Design, p. 164, has an instructive simple example that is worth working out (see p. 169).

**Question.** For a $N$-stage pipeline, show that the failure probability due to NBTI degradation is linearly proportional to the number of stages, with a power-exponent a simple multiple of NBTI power-exponent.

**Ans.** They authors proceed with the weakest link model to calculate the failure probability of $N$ series connected system, to the failure probability of individual pipeline.

\[
1 - F_p = 1 - \prod (1 - f_i) = 1 - \prod (1 - A_i t^{kn}) \sim 1 - N \times A \times t^{3n}
\]

where $k \sim 3$ from numerical simulation of single stage inverter, NAND, NOR gates, etc. For detailed derivation, you may wish to consult the reference above.