Abstract

Many components of variability become larger percentage design factors with decreasing feature size. Hence, the small transistors in SRAM cells are particularly sensitive to these variations. The SRAM cell transistors in sub-100 nm designs may contain fewer than 100 channel dopant atoms. To achieve a robust design with such variability, one must enhance the normal static-noise-margin and write-trip-point analysis, often with Monte Carlo simulations using statistical transistor models including the process and mismatch fluctuations. Similar challenges exist for the sense amplifiers normally used with SRAM arrays. Except with very low speed designs, yield to speed can be substantially reduced by variations between nominally matched sense amplifier transistors as well as by the variability resulting in a very worst memory cell low read current. This also increases the hazards of delay timing with dummy paths and dummy cells and increases the need for at-speed testing prior to repair.

I. Introduction

SRAM memory cells have always been designed to occupy the minimum amount of silicon area consistent with the performance and reliability required. The reliability is usually measured by static noise margin, SNM [1], and write trip point simulations and measurements. As the lithography shrinks, the device variations are becoming an ever increasing concern. With the sub-100 nm processes, statistical variations must be included in the SRAM cell and SRAM block analysis to achieve circuit designs with sufficient yield and performance within a limited area and power budget. The following sections will examine the memory cell, sense amplifier, and timing circuitry variations found with sub-100 nm designs and what must be evaluated to predict circuit yield due to process variations.

II. Major Sources of Variation

The major source of device variations has been the $V_t$ variations due to the fluctuations in the small number of dopant atoms in the channel. This is modeled as a $1/(\text{Area})^{0.5}$ factor being primarily a function of the transistor channel area [2]. An example of how the uncertainty in $V_t$ is increasing with process generations for small devices is shown in the Table.

<table>
<thead>
<tr>
<th>$L$ (nm)</th>
<th>250</th>
<th>180</th>
<th>130</th>
<th>90</th>
<th>65</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_t$ (mV)</td>
<td>450</td>
<td>400</td>
<td>330</td>
<td>300</td>
<td>280</td>
<td>200</td>
</tr>
<tr>
<td>$\sigma$($V_t$/mV)</td>
<td>21</td>
<td>23</td>
<td>27</td>
<td>28</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>$\sigma$($V_t$/Vt)</td>
<td>4.7%</td>
<td>5.8%</td>
<td>8.2%</td>
<td>9.3%</td>
<td>10.7%</td>
<td>16%</td>
</tr>
</tbody>
</table>

From ITRS Roadmap 2001 [4]

Table: Threshold voltage variability in high performance processes.

Memory cells are most sensitive to device variations causing device mismatch for several reasons. First and foremost, the memory cell uses the smallest viable transistors for a given process in order to minimize the array area. In most cases the devices used in the smallest memory cell for a given process are smaller than the devices allowed elsewhere in the design. In addition, the memory cell read timing is usually such that full logic levels are not achieved. So the device differences become output level differences at the edge of the memory array. Also, the memory cells must be able to keep the stored state when accessed for reading but quickly change state when accessed for writing. These conflicting needs are even more difficult to achieve with the process variations of sub-100 nm processes.

III. Static Noise Margin

If a memory cell cannot hold state, it is not usable. For a symmetric six transistor memory cell as shown in Figure 1, the cell is least stable when it is just beginning to be accessed for reading. The read operation of this cell begins with the word line being raised to the power supply voltage, $V_{cc}$, with the Bit and Bit lines initially high. This causes the low node within the cell to rise due to the source follower action of the bit line coupling transistor, ‘xsfr.’ If this node voltage becomes close to the $V_t$ of the pull down devices, process variations combined with noise coupling may flip the state of the cell. The margin between the read state and the unstable cell is the cell static noise margin and is usually measured as described by
Seevinck, et. al. [1] producing the well known butterfly curve shown in Figure 2. The static noise margin is defined in this reference as the edge of the largest square which can be constructed between the two transfer curves. A larger static noise margin implies a memory cell is less likely to loose state during a read operation. Note that for Figure 2 and all other simulations and waveforms, the U.C. Berkeley BSIM4 65 nm model [5] is used. The waveforms in Figure 2 are voltage transfer curves for the two inverters within the memory cell with the transfer gates included as source follower loads. By overlaying the two curves, one can see the static noise margin of the circuit. For well marched devices, the curves are identical and the two SNM maximums are the same. However, when device variations are included, the SNM maximums become unequal and the minimum of the two must be used.

Once it was sufficient to require the ratio of the drive transistor’s W/L to the W/L of the transfer transistor to be greater than 2 to assure good stability. But sub-100 nm processes cannot meet this goal and maintain the desired small cell area. Monte Carlo analysis using statistical models are now used to achieve sufficient SNM with the minimum cell area. The expectation is that the SNM will be most sensitive to the ‘drive’ and ‘xsfr’ transistors and relatively insensitive to the PMOS ‘up’ transistor. The memory cell variations due to die-to-die process shifts are best analyzed separate from within the die mismatch variations as the influence on the SRAM array yield is different.

IV. Die Average Static Noise Margin Variations

Process variations between die, from wafer to wafer, and from lot to lot are limited by the foundry practices and standards. Often any wafer which is not within 2-3 sigma of the process expectations is rejected and not included in the finished wafers to be tested. Designers examine this expected process range by using “worst case” corners to verify that circuits will operate correctly over the expected spread of process variations. However, some of these corners involve impossible situations such as P-fast with the minimum oxide thickness and N-slow with the maximum oxide thickness which obviously cannot exist at the same time. Monte Carlo simulation with a statistical die average model gives much more realistic results of how the circuits and especially how a memory cell will operate over the expected die average process variations. Figure 3 shows the results of a few Monte Carlo loops where the $V_{tn}$ and $V_{tp}$ of all devices are allowed to vary together as with a die average statistical analysis. To quantify the results of such a simulation, one can apply standard statistical formulas as long as the results are in the range where the results are approximately gaussian. For example, a
die average statistical process model might result in a
SNM(mean) = 123.1 mV and a sigma = 15.6 mV. If the pro-
cess control rejects wafers beyond the 3 sigma limit, the min-
imum SNM(die-ave) which the circuit design must
accommodate is:
SNM_{min}(die-ave) = SNM(mean) - N_{σ} \cdot σ(die-ave)
or
SNM_{min} = 123.1 - 3 \cdot 15.6 = 76.3 mV.

VI. Write Trip Point and Cell Read Current
While SRAM yield requires a good SNM, it is equally
important to have a reasonable write trip point. The write trip
point is primarily determined by the $V_{t}$ of the NMOS ‘xsfr’
transistor connected to the bit line being pulled low (see Fig-
ure 1). It is required that the trip voltage be far enough from
either supply that no combination of offsets and noise can
cause a write failure or a write when a read is intended. Usu-
ally a symmetric 6T memory cell has a trip point slightly less
than the half power supply voltage, that is: $V_{trip} < V_{cc}/2$. A
typical simulation result is shown in Figure 5. In this figure,
the bit line being pulled low is shown as a ramp from $V_{cc}$ on
the left to $V_{ss}$ at the right. The cell internal nodes (A and B of
Figure 1) remain either high or low until the write trip volt-
age is approached. Then the high node starts dropping and
finally the low node rises suddenly as the high node falls and
the cell switches state. The multiple curves are for a few dif-
ferent $V_{t}$ points in a set of statistical simulations. As with
SNM, the process variation range is fixed by the fabrication
specification while mismatch variation tolerance is deter-
mined by the number of cells or circuits used and the accept-
able yield loss increase. A formula for this trip point margin
due to mismatch relative to $V_{ss}$ can be expressed as:
$V_{trip}(mean) - N_{σ} \cdot σ(trip-mis) > ΔV_{ss}$. An exam-
ple: $ΔV_{ss} = 150 mV, V_{trip}(mean) = 410 mV$ and $σ(trip-mis) = 44 mV$, then $N_{σ}(trip-mis) = 5.91 (1 in 580 million)$.
With the SNM and Write-trip-voltage under control, one then tries to achieve a low standby current and an acceptable read current for the weakest cell. The cell read current, $I_{\text{read}}$, is a major component in determining array access time. Hence, the overall chip speed may be limited by the $I_{\text{read}}$ of the weakest cell. The symmetric 6T cell read current is set by the series stack of NMOS transfer transistor and NMOS pull down transistor. The PMOS transistor is only a leakage path (source-drain and gate). Hence, the statistical variations in $I_{\text{read}}$ are almost completely determined by the NMOS characteristics. After determining $I_{\text{read}}(\text{mean})$ and $\sigma_{I_{\text{read}}}(\text{mis})$ from simulations, one can apply:

$$I_{\text{read}}(\text{min-mis}) = I_{\text{read}}(\text{mean}) - N \sigma_{I_{\text{read}}}(\text{mis})$$

For example: $I_{\text{read}}(\text{mean}) = 100 \mu\text{A}$, $\sigma_{I_{\text{read}}}(\text{mis}) = 6 \mu\text{A}$, and $I_{\text{read}}(\text{min}) = 70 \mu\text{A}$ to meet timing. This is satisfied for $N = 5$ which implies that one cell in 3.5 million will have insufficient $I_{\text{read}}$ to meet the timing goal.

For this set of read current parameters, a reasonable yield of a 1MB array would require redundancy for replacing low $I_{\text{read}}$ cells as well as non-functioning cells. The yield to a minimum $I_{\text{read}}$ is shown in Figure 6 for 0 to 5 repairs. This figure shows that repair of the slowest 3-5 cells can significantly improve the yield to a fixed read current and hence a fixed access time. But note that beyond a few cells the improvement is much more limited. Of course, the test/repair procedures must be capable of detecting and replacing these weak cells to achieve the added yield to speed.

![Figure 6. Yield of 1MB RAM to a minimum $I_{\text{read}}$ with 0 to 5 repairs.](image)

**VII. SRAM Sense Amplifier**

The sense amplifier, SA, in a SRAM design converts the small signal output of a memory cell (perhaps < 100 mV) to logic levels of 1 V or more. Most sense amplifiers involve carefully matched transistors designed to minimize $I_{ds}$ and $V_t$ mismatches. Variations with a sub-100 nm process can even very well matched sense amplifiers to become significant limitations to the circuit yield. With a differential voltage sense amplifier, the careful matching of the NMOS transistors of the differential pair is usually most critical. Any layout or $V_t$ mismatch is critical so all the best historic layout matching practices feasible are used including current flow direction. For a voltage mode cross coupled sense amplifier, such as the one shown in Figure 7, the layout of all cross coupled devices and especially the differential NMOS devices should be exactly the same including the physical current flow direction. Best practices include a 50% split current flow. All other devices are matched as exact as practical and checks are made with 1/2 design rule offset of each layer in all directions. Current mode sense amplifiers are less sensitive to $V_t$ mismatch but are more sensitive to $I_{ds}$ mismatch.

Layout care similar to that used in the voltage sense amplifier is needed to minimize the variations due to layout differences. A skewed inverter sense amplifier is not as robust as it would be as part of a logical block. The memory cell current gives a very slow voltage slew rate. Hence, the inverter trip point is much more sensitive than when used for standard logic and all parameters determining the trip point must be included in a statistical analysis. In general, all sense amplifier types need statistical simulation with a bit line fully populated with statistically varying cells to determine the effective SA input offset. This is critical in order to verify that the circuit can overcome all the offsets and meet timing with a statistically weak cell. As an example, examine the drain fed differential voltage sense amplifier shown in Figure 7. The critical mismatch is between the NMOS differential pair transistors as they are just turning on when $V_{\text{in}} > V_t(\text{mismatch})$ [7]. Monte Carlo simulation is used with the fail condition measured as a function of $\Delta V_{\text{in}}$ at this sense amplifier turn on time, $\tau_{sae}$. The fail conditions may be defined either as the wrong output latching or the correct output latching but with a glitch on the high output greater than the maximum allowed. This inherent noise glitch, shown in figure 8, is reduced by relaxed timing. But high speed SRAM operation requires that the time between the wordline turn on and $\tau_{sae}$ be minimized consistent with 100% correct sensing.

![Figure 7. Drain fed differential sense amplifier](image)
To get a rough idea of the mismatch which $\Delta V_{in}$ must overcome, assume the circuit electrical yield is limited by the high side bump with $V_{bump} < 200$ mV required. The mismatch simulation is done with the lowest fan-out to obtain the worst case bump. For this example, the sense amplifier is for a large SRAM block requiring 20,000 identical sense amplifier copies. Assume $\Delta V_{in} = 3$ mV, $\sigma(SA-mis) = 12.5$ mV, and add the requirement that no more than 1% of these SRAM blocks contain a sense amplifier with $V_{bump} > 200$ mV. This is 1 fail in 2 million, which is satisfied by $N \sigma(SA-mis) = 4.9$. Then $\Delta V_{in}(\text{min}) = 3 + 12.5 \times 4.9 = 64.25$ mV. The huge difference between the expected sense amplifier minimum input voltage without mismatch considerations and with mismatch shows the difficulty of sensing small differences with compact designs in a sub-100 nm process.

**VIII. Critical Path Timing**

To verify that the statistically weak cell meets timing with this offset, the statistical simulation should be run with the memory cells and the sense amplifier varying. As this can be a long involved simulation study, a quick slightly pessimistic estimate can be done first by combining the results of the separate sense amplifier and memory cell evaluations. Since the minimum cell read current must generate 64.25 mV as was found above, it follows that the typical cell will generate a larger differential which is approximately the scaling of the average to minimum cell read currents. Recall that: $I_{\text{read}}(\text{min-mis}) = I_{\text{read}}(\text{mean}) - 5 \times \sigma(I_{\text{read-mis}})$. With $I_{\text{read}}(\text{mean}) = 100 \mu A$ and $\sigma(I) = 6 \mu A$, then $I_{\text{read}}(\text{min-mis}) = 70 \mu A$. Applying this to the $\Delta V_{in}(\text{min})'$ for a typical cell to account for the statistically weak cell gives: $\Delta V_{in}(\text{min})' = \Delta V_{in}(\text{min}) \times I_{\text{read}}(\text{mean}) / I_{\text{read}}(\text{min-mis})$ or $64.25 \times 100 / 70 = 91.8$ mV. Of course, this is a worst case analysis with statistical limiting. The combined statistical analysis is best but adds substantial simulation time.

**IX. SRAM Timing Circuits**

Many SRAM blocks use replica memory cells, rows, and columns as references to determine the appropriate timing within a large SRAM block. However, these reference memory cells, rows, and columns used for timing control are subject to the same statistical variations as the normal circuitry. The simple SRAM block shown in Figure 9 might be designed with Word-m, Memory cell-mn and Column-n used as timing references for the entire SRAM block. However, the probability that a single reference cell is more than $+/-2\sigma$ away from the expected average is 4.5%. Hence, large devices or many copies are needed to achieve a replica which is a true mean of the replicated circuitry. For a cell read current replica, a large number of cells must be used to achieve at good statistical mean. For a timing reference, the loading and exact structure the of delay elements are critical and must include the correct metal to device gate plus diffusion area ratio for good tracking. Good references will be even more difficult to construct with the higher percentage statistical variations expected in the 65 and 45 nm processes.

![Figure 8. Sense amplifier control signals and node waveforms with inherent noise and $\tau_{sae}$ identified.](image8.png)

![Figure 9. Block diagram of a simple SRAM block.](image9.png)
Hence, a non-clock edge is used and either the sense amplifier sampling time or the bit line recovery time does not scale with frequency. Obviously, this edge must be very carefully simulated and appropriate margins must be checked to guarantee operation in these sub-100 nm processes. Still, clock based timing is usually a more reliable choice than dummy rows, columns, and memory cells at these process nodes.

**X. Concluding Remarks**

The variation in device characteristics is becoming an acute problem for sub-100 nm SRAM designs. Circuit wide process windows of 2 to 3 times the typical value occur often and will not decrease with future technologies. Within the die parametric variations and adjacent device parametric variations are not decreasing as quickly as the parameters themselves. Yet the number of SRAM cells and other circuit blocks which must work on a chip is growing rapidly. Traditional mean and corner simulations must be supplemented with a number of statistical simulations to assure circuits will work and yield at the expected speed and reliability. Repair of slow cells must be made possible by design.

**XI. Acknowledgements**

The authors acknowledge the Device Group at UC Berkeley for the Berkeley Predictive Technology Model, BPTM, used in all simulations. http://www-device.eecs.berkeley.edu/~ptm [5].

**XII. References**


**XIII. Related Reading**


