Abstract

Reduced device dimensions and operating voltages that accompany technology scaling have led to increased design challenges with each successive technology node. Large scale 6T SRAM arrays beyond 65nm will increasingly rely on assist methods to overcome the functional limitations imposed by increased variation, reduced overdrive and the inherent read stability/write margin trade off. Factors such as reliability, leakage and data retention establish the boundary conditions for the maximum voltage bias permitted for a given circuit assist approach. These constraints set an upper limit on the potential yield improvement that can be obtained for a given assist method and limit the minimum operation voltage (V\text{min}). By application of this set of constraints, we show that the read assist limit contour (ALC) in the margin/delay space can provide insight into the ultimate limits for the nano-scale CMOS 6T SRAM.

1. Introduction

Increased device variability and reduced overdrive associated with lower operating voltages have reduced the functional yield margins in VLSI circuits. This is particularly true for the 6T SRAM, which continues to play a dominant role in future technology generations because of its combination of density, performance, and compatibility with logic processing. The successful commercial scaling of the 6T SRAM driven by strong industry competition is expected to continue beyond the 32nm node. The continued trend in area reduction is accompanied by the well known consequence of increased variability associated with the reduced channel area. Although technology options such as high κ with metal gate may provide some relief in variability, diminished functional margins coupled with the growth in bit count pose a serious technical challenge beyond the 28/32nm generation.

Because of the commercial success of the 6T SRAM, methods to address the failure mechanisms of large memory arrays will extend the life of the 6T SRAM in VLSI circuits. Fail types for SRAM arrays may be divided into two distinct categories: “hard fails”, i.e., those attributable to defects, and “soft fails”. Soft fails defined in this context are those voltage, temperature and timing dependent fails resulting from one of the following four modes: (1) failure to write, (2) failure to read (insufficient signal developed on the BL), (3) stability upset during a read or half-select condition, and (4) data retention failure. These four failure modes each first occur at the distribution tail stemming from global and local variation sources.

The use of bias based circuit assist methods has become increasingly common, primarily to address soft fail modes 1 and 3 and to preserve the 6T cell functionality as the variation continues to increase and both read and write margins decrease with scaling. Although numerous recent articles have discussed bias based assist for SRAM [1-21], limitations exist for all of these techniques. This limit may be reliability, performance, leakage, energy, or other factors which ultimately bound the extent to which the assist method compensates for the reduced functional margins.

The objective of this paper is to explore the boundaries of bias based assist methods to understand the impact on the minimum operation voltage (V\text{min}) and the effectiveness of the assist methods for future generations of 6T SRAM. Based on the relationship between performance and functional margin with the applied bias constraints, we define the assist limit contour (ALC) for read assist and demonstrate it across four technology generations. For write assist methods, besides the constraint from reliability, the read stability of half-selected cells limits the permissible assist bias. The remainder of the paper is organized as follows. Section 2 provides a brief background of various bias based assist methods. Section 3 provides the methodology we followed in this analysis. The results and analysis for read assist and write assist simulations across four LP-PTM technology generations are contained in sections 4 and 5. We then summarize and draw our final conclusions in section 6.

2. Bias based assist methods

Read assist methods refer to the set of circuit options that are used to either reduce the read noise source or improve the cell stability so that the cell remains stable during a read access. The methods which weaken read disturbance include reduced word line gate voltage [1][2][3][4][5]. Increased pass gate threshold voltage through body bias [6][7], and reduced bit line charge by lowering the voltage or capacitance [8][9][10][11]. The methods that are intended to improve the resilience of the latch are increased array VDD (VDDc) [2][12][13][14][15], decreased array VSS (VSSc) [3], and strengthening the pull up PMOS device by NWELL bias [6].

Those bias conditions which improve the write margin of the SRAM cell are referred to as write assist methods. These include boosted word line gate voltage [2][12][13][16] or reduced bit line voltage [3][5][17] to increase the V\text{GS} of the pass gate device. Those publications that address improving write margin by means of reducing the latch strength, include reducing the array supply voltage (VDDc) [1][2][4][7][14][18], raising the array VSS (VSSc) [9][19][20] or reducing the strength the pull up PMOS by NWELL bias [6][21]. Table 1 provides a summary of bias based SRAM assist methods. In addition to the methods listed in Table 1, raised global VDD, simultaneously improves both read and write margin.
Figure 1: Static noise margin as function of (a) global VDD, (b) negative array VSS, (c) array VDD boost, and (d) WL droop.

Figure 2: Write margin as function of (a) negative BL, (b) raised array VSS, (c) array VDD droop, and (d) WL boost.

Table 1: Summary of bias based assist methods

<table>
<thead>
<tr>
<th>Assist  Method refs</th>
<th>Assist Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>WL voltage ↓ 1,2,3,4,5</td>
<td>Read Assist</td>
</tr>
<tr>
<td>Pass Gate NMOS V_T ↑ 6,7</td>
<td></td>
</tr>
<tr>
<td>Pull up PMOS IV ↑ 6</td>
<td></td>
</tr>
<tr>
<td>BL voltage ↓ 8,9,10,11</td>
<td></td>
</tr>
<tr>
<td>Array VDD (VDDc)↑ 2,12,13,14,15</td>
<td></td>
</tr>
<tr>
<td>negative VSS (VSSc)↓ 3</td>
<td></td>
</tr>
<tr>
<td>WL voltage ↑ 2,12,13,16</td>
<td></td>
</tr>
<tr>
<td>negative BL ↓ 3,5,17</td>
<td></td>
</tr>
<tr>
<td>array VDD (VDDc)↓ 1,2,4,7,14,18</td>
<td></td>
</tr>
<tr>
<td>Array VSS (VSSc)↑ 9,19,20</td>
<td></td>
</tr>
<tr>
<td>Pull up PMOS IV ↑ 6,21</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Custom low power technology models 65nm-22nm

3. Quantifying assist limits

To compare the margin sensitivity of the specific assist methods discussed in this paper, quantitative analysis is performed using custom low power predictive technology models (LP-PTMs), Table 2. The models were adapted from published PTM values [22] to more closely match industry published LP technologies [23][24]. The bit cell device dimensions emulate the dense industry cell and scale with each generation to approximate published industry values [18] [25] [24]. Using the custom LP-PTM models combined with the scaled 6T device dimensions we quantify the specific functional challenge for 6T SRAM across four technology generations (65nm to 22nm). The detailed analysis of the specific simulations performed will be illustrated using the 45nm LP-PTM technology. We then extend the simulations across the range of technologies to better understand ultimate limitations of 6T SRAM assist methods with scaling.

Figure 1 shows the read static noise margin (SNM) response to modulated voltage bias across the four technology nodes (65nm-22nm) considered in this work. A negative slope corresponds to those methods requiring the terminal voltage to decrease below the nominal value. The change in write margin (WM) with terminal bias is shown in Figure 2 for four different bias assist methods. The read SNM values in this paper are extracted from the butterfly curve simulation while the write margin is measured using the WL voltage sweep approach [26]. The bias sensitivity provides a quantitative means of characterizing the assist methods and is defined as:

\[
\text{Sensitivity} = \frac{\partial (\text{Margin})}{\partial V}
\]

Margin may refer to either SNM or WM while V refers to the modulated voltage.

Table 2: Custom low power technology models 65nm-22nm

Mann, Limits of Bias Based Assist Methods in…
We will use a margin/delay analysis to compare the assist methods. For read assist, we are specifically concerned with the relationship between the read static noise margin (RSNM) and the read delay. For write assist, the relationship is write margin (WM) versus write delay. The effect of global and local variation is implicitly addressed through the establishment of the worst case margin/delay requirements.

Figure 3: Margin/delay relationship for write (a) and read (b) for the 45nm LP-PTM. Points radiating from the origin as identified by the arrow are modulated in 100mV increments.

Using the 45nm LP-PTM technology, we illustrate the margin/delay concept for write and read in Figure 3(a) and (b), respectively. The optimal assist method will reduce the delay while increasing the margin. Figure 3 shows a simulation of the condition in which the global VDD is drooped 20% below the nominal technology voltage. Application of the bias based assist methods as shown produce a change in the margin/delay space which may be advantageous. As shown in Figure 3(a), the WL boost and negative BL result in a stronger net write delay/margin response over the collapsed supplies across the latch as represented by reduced VDDc or increased VSSc. For the read delay/margin space, as shown in Figure 3(b), the negative VSSc improves the read delay but is significantly less effective in improving the margin. The drooped WL improves the RSNM but quickly degrades the read performance.

We characterize this margin/delay relationship schematically in Figure 4 to capture the response for both read and write assist. Following the diagram as depicted in Figure 4, at nominal VDD, the normal variation in margin and delay are contained within the required window. However, as the VDD is reduced, both the performance and margin values move outside the desired window (thick arrow). Applying an assist bias (a thin arrow) may improve margin and/or delay but the result may not be optimal.

The functional window boundaries, shown schematically in Figure 4, are established by four factors, (1) array size, (2) variation, (3) the delay requirement, and the (4) target soft fail yield requirement for the array. The lower horizontal SNM boundary is delineated by the array size, the SNM distribution, and the soft fail yield requirement at the worst case voltage corner (Vwc). The vertical functional boundary is determined by the worst case delay requirement and variation for a given array size. For the case shown in Figure 4, this is illustrated by the right hand boundary of the shaded box. For the purposes of this illustration, this window represents the worst case margin/delay value required for a given array size including variation.

For a given margin requirement, several assist methods may suffice. With the additional performance requirement, the set of assist options and required bias conditions necessary to fulfill both margin limited yield and performance is reduced.

Figure 4: Schematic diagram of read/write margin versus read/write delay and desired functional window based on margin limited yield and performance requirements for application.

The magnitude and trajectory values are influenced by the sensitivity as given in expression (1) and the applied voltage bias of the assist technique. Because the magnitude of the applied assist bias must note exceed a specified maximum value as defined by the technology reliability requirements, this establishes an upper margin boundary which will ultimately limit the array soft fail limited yield.
By application of the constraint limitations we may then begin to map the maximum assist margin values permissible. We begin by examining the maximum permissible assist bias based on the reliability constraints defined for each technology. The reliability limit may be due to several factors such as time dependent dielectric breakdown, hot carrier, NBTI or a combination of the known mechanisms with sufficient voltage acceleration. The maximum assist bias offset \( V_{assist} \) that may be applied for any given assist method based on the reliability (Vmax) constraint is expressed as:

\[
V_{assist} = (V_{\text{max}} - V_{\text{nom}}) + V_{\text{droop}}
\]  

(2)

\( V_{\text{nom}} \) and \( V_{\text{max}} \) refer to the nominal and maximum operation voltage as specified by the technology developers (Vnom values provided in Table 2 for this work). \( V_{\text{droop}} \) refers to the difference between \( V_{\text{nom}} \) and the instantaneous operation voltage. To illustrate this concept briefly, for a technology in which the Vnom/Vmax is 1.2V/1.32V respectively, if the array VDD is dropped from 1.2V to 1V, the maximum assist bias is 0.32V. Any bias exceeding 0.32V would exceed Vmax for the transistor, violating the reliability constraint. For the same reason, a maximum negative assist bias of 0.32V may be applied provided all VDD supply terminals associated with the array are maintained at 1V. Additional constraints may apply, but this single constraint provides a defined boundary that we will discuss in section 4.

In addition to the technology defined Vmax constraint, other assist bias constraints for read assist bias include; forward bias diode turn-on (VFwd) when VSSc is intentionally driven below ground, and cell upset by writing a zero when both bit lines are drooped sufficiently low (Write 0). For write assist, the constraints are again reliability (Vmax) as well as data retention (DR) for non-accessed cells sharing the intentionally modulated common supply, forward biased diode turn-on (VFwd) when the ‘write zero’ bit line is driven below ground, and cell stability for the half-selected cells on the asserted word line. The primary bias constraints are summarized in Table 3 for the bias based assist methods evaluated in this paper. Vmax is a valid constraint for all cases. This is less obvious for the two write assist options that involve collapsed supply across the latch. Vmax remains a constraint for the maximum write margin because it still limits the maximum WL voltage.

For the purposes of this work, Vmax will be defined as 10% above the nominal operation voltage. Because Vmax is a limiting factor in all bias based assist methods, we will exploit this fact to explore the limits of the assist methods across the scaled technologies. This approach allows us to effectively define the upper envelope of assist bias conditions permissible for a given technology. By mapping the assist methods across the margin/delay space, the functional window may then be used to illuminate the practical voltage bias boundaries.

### 4. Results

To examine the maximum soft fail limited yield boundaries that can be achieved for a given assist method, we will first describe the relationship with VDD and then apply the assist bias using the Vmax constraint. The read static noise margin as a function of VDD is shown in Figure 5 for the 45nm LP PTM technology. The three cases shown are with array VDD (VDDc) boost, array VSS (VSSc) reduced, and with no assist. It is clear from Figure 5(a) that the use of the maximum assist bias, consistent with relationship (2), can significantly improve the otherwise reduced static noise margin when the word line is asserted. The SNM improves beyond the nominal value when the VDDc assist method is invoked because the noise source is being reduced with VDD reduction, and the latch strength is increasing with the boosted VDDc.

Competing mechanisms produce a different result with negative VSSc. In this case, although the net latch strength is improved over the non-assist case, the noise source through the pass gate NFEF is becoming stronger due to the body effect producing a reduction in pass gate \( V_T \) on the side of the cell storing a zero. Additionally, the \( V_T \) is reduced for the pull down NMOS device with drain storing a one. This results in an earlier turn of this pull down NMOS and further reduces the SNM. The read delay for the cell is improved due the body effect which strengthens both the pull down and pass gate series devices on the side of the latch storing a zero.

While the cell stability compensation is larger for VDDc assist, the improvement in performance (read delay) may not be sufficient depending on the functional window as discussed earlier. Boosting the VDD at the cell (VDDc) has a small impact on the read delay, consequently the read delay continues to degrade as VDD is reduced. The alternate read assist method (VSSc) shown in Figure 5 improves SNM to some degree but more significantly improves the read performance. This is because the body effect associated with reduced VSSc causes both the pull down and pass gate NFET device \( V_T \) to be reduced, boosting the read current.

We next examine the margin/delay relationship for the assist methods with maximum assist bias. The effect of maximum assist bias on both SNM and delay based on the modulation of single and multiple terminals is shown in Figure 6 for the 45nm LP-PTM technology. Each of the read assist bias conditions given in Tables 1 and 3 except those involving well bias \( V_T \) modulation were employed.

| Table 3: Summary of constraints for bias based assists |
|---|---|---|
| Assist | Method | Bias Constraint |
| Read Assist | WL voltage ↓ | Vmax |
| Pass Gate Vt ↑ | Vmax |
| BL voltage ↓ | Vmax and Write 0 |
| Array VDD ↑ | Vmax |
| negative VSS ↓ | Vmax, VFwd |
| PMOS IVtl ↓ | Vmax |
| Write Assist | WL voltage ↑ | Vmax, RSNM (½ Select) |
| negative BL ↓ | Vmax and VFwd |
| array VDD ↓ | Vmax, DR (with shared VDDc) |
| Array VSS ↑ | Vmax, DR (with shared VSSc) |
| PMOS IVtl ↑ | Vmax, RSNM (½ Select) |
The margin/delay analysis reveals the limits of the bias based assist methods across the relevant design space. This boundary further defines a contour, as shown by the solid continuous line (demonstrated using VDDc and VSSc assist bias following the Vmax constraint). We will refer to this boundary as the assist limit contour (ALC). It establishes the effective limit in SNM and corresponding relationship to read performance for a given technology and bit cell. This boundary or ALC mapped by the assist methods therefore provides a means of assessing the functional limits of the 6T SRAM.

Figure 5: Change in RSNM with reduced VDD (a) and effect of VDD on read delay (b) with the maximum allowable assist bias at each VDD. Data based on 45nm LP PTM.

For the cases we studied, the read ALC as defined by the latch supply voltages were found to provide a reasonable approximation of the full multi-terminal Vmax read assist contour. Because drooping the WL provides a degree of freedom that is not limited by the Vmax constraint, those combinations of negative VSSc combined with WL voltage reduction where found to produce a slightly improved margin/delay response for the LP-PTM technologies.

Because the primary goal of this work is to identify and delineate the bias based assist limitations of the scaled 6T SRAM cell, the delay required in developing the bias conditions is not included in this analysis. A complete SRAM macro design would need to include the overhead delay associated with the specific implementation and circuit choice.

Figure 6: Multiple read assist options involving both single and multiple terminals with Vmax constraint preserved.

Figure 7: Write margin decreases as VDD is reduced when no assist is used. With assist at Vmax, the write margin is increased with reduced VDD.

For write assist, the margin/delay analysis leads to a different result. In this case, there is no inherent trade off between write margin and write delay. The relationship is shown in Figure 7 for two assist methods (negative BL and boosted WL) showing improved margin and delay as array VDD is drooped. For boosted WL assist, the array VDD, high bit line and NWELL voltages are reduced, while the WL line is boosted to Vmax limited by the reliability constraint between the WL voltage and the low (write zero) bit line at 0. For the negative BL case, as the VDD is reduced on the word line, high bit line, and array VDD, while the low bit line is drooped by the same amount to preserve the Vmax constraint.

With the word line boosted to Vmax, the write margin continues to increase with corresponding VDD reduction. Similarly, with the (write zero) bit line driven below ground by a value equivalent to the VDD reduction (preserving the Vmax constraint), the write margin continues to increase. In
addition to Vmax, the maximum write assist bias may become limited by other constraints, such as the read margin for the half-selected bits, shown in Table 3.

Figure 8: The impact of WL boost on the WM of the selected bits and the stability (RSNM) of the half-selected bits.

Figure 9: The impact of negative BL on the WM of the selected bits and the stability (RSNM) of the half-selected bits.

For the WL-boost write assist, Figure 8 shows that as the array supply voltage is reduced, boosting the WL while preserving Vmax reduces the stability (RSNM) of the half-select bits on the same WL. Therefore, the limiter for the WL boost quickly becomes the reduced SNM on the half-selected bits. For the negative BL assist, the BL bias does not directly impact the half-select bits. However, because the amount of bias between the BL voltage and the global VDD is limited to Vmax, a larger negative bias on the BL implies a lower global VDD. Thus the RSNM of the half-selected bits consequently decreases, permitting a larger negative BL bias, as shown in Figure 9. By comparison, the degradation in RSNM for the half-selected bits using the negative BL assist, Figure 9, results in less degradation for the half-selected RSNM. This is an advantage of the negative BL assist over the WL boost. However, as the array supply droops, the negative BL bias can eventually become limited by leakage to the substrate as the forward bias diode begins turning on. To overcome the problematic stability concern for the half-selected bits during a write, a read assist such as VDDc boost may be applied to the non-selected columns. Alternatively, the array architecture can be designed so that the half-select is avoided and all bits on the asserted WL are latched during a write operation.

5. Discussion

We further examine the characteristic features of the margin/delay plot for read assist. Figure 10 reveals that the read assist limit contour (ALC) asymptotically approaches the hold SNM (HSNM) limit with increased delay. This is anticipated as the latch strength is increased relative to the noise source, the SNM upper limit will approach the HSNM with VDD=Vmax. For the case where the NWELL potential is tied to the array VDD (VDDc), the upper limit will be equal to the hold SNM. This can be more clearly seen from the butterfly curves. Figure 11 plots the butterfly curves when the VDDc assist method is used with increased assist bias. The characteristic shape evolves with increased assist bias, becoming more similar to the hold SNM shape. Because the performance implications of achieving this limit are in most cases not practical, the more relevant portion of the ALC is across the intersection of the functional window as shown schematically as a shaded region in Figure 10.

For a given set of technology bias constraints, a contour line defining the upper most noise margin at a given read delay for the technology may be derived. Figure 12 plots the read ALC mapped across margin/delay space for 45nm LP PTM 6T SRAM.

For designs requiring both read and write assist, the yield limiting condition, if not otherwise addressed, may then be the stability upset half-selected bits during a write operation.
As shown in Figures 8 and 9, the half-selected bits may be adversely affected by the choice of write assist method. This challenge may be addressed by adding an additional bias based assist solution or by array architecture changes which in effect avoid the half-select condition.

6. Conclusions

Continued scaling of the planar 6T SRAM will necessitate increased reliance on assist methods to overcome reduced functional yield margins. Because added assist features will incur costs in design complexity, area, and in most cases power, these factors must be balanced against the potential improvement in soft fail limited yield margin and performance. For bias based assist methods, bias constraints ultimately limit the margin improvements that can be obtained. The applied voltage bias associated with a given assist method must conform to the existing technology bias constraints.

For write assist, in addition to the Vmax constraint, other combined factors also limit the attainable margins. For read assist, by imposing the Vmax constraint a contour is observed in the margin delay space that reflects the relevant attainable limits of a given assist method. The intersection of the ALC with the functional window requirement provides a means to establish bias based assist limitations for a given technology and bit cell. By accounting for these factors, we map the Vmax constrained read ALC across four technology generations to gain additional insight into the extent to which assist methods may continue to compensate for the reduced functional margins with continued scaling of the planar 6T SRAM.

7. References


