Solid-State Electronics 54 (2010) 1398–1407

Contents lists available at ScienceDirect

Solid-State Electronics

journal homepage: www.elsevier.com/locate/sse

Impact of circuit assist methods on margin and performance in 6T SRAM

Randy W. Mann [a,*], Jiajing Wang [a], Satyanand Nalam [a], Sudhanshu Khanna [a], Geordie Braceras [b], Harold Pilo [b], Benton H. Calhoun [a]

[a] Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904, USA
[b] IBM Microelectronics, Essex Junction, VT 05452, USA

ARTICLE INFO

Article history:
Received 17 February 2010
Received in revised form 1 June 2010
Accepted 10 June 2010
The review of this paper was arranged by Prof. A. Zaslavsky

Keywords:
SRAM
SNM
Write margin
Read assist
Write assist
Vmin
Scaling
Process variation
Yield

ABSTRACT

Large scale 6T SRAM beyond 65 nm will increasingly rely on assist methods to overcome the functional limitations associated with scaling and the inherent read stability/write margin trade off. The primary focus of the circuit assist methods has been improved read or write margin with less attention given to the implications for performance. In this work, we introduce margin sensitivity and margin/delay analysis tools for assessing the functional effectiveness of the bias based assist methods and show the direct implications on voltage sensitive yield. A margin/delay analysis of bias based circuit assist methods is presented, highlighting the assist impact on the functional metrics, margin and performance. A means of categorizing the assist methods is developed to provide a first order understanding of the underlying mechanisms. The analysis spans four generations of low power technologies to show the trends and long term effectiveness of the circuit assist techniques in future low power bulk technologies.

1. Introduction

The 6T SRAM cell design has been successfully scaled in both bulk and SOI down to the 32/28 nm node and has remained for more than a decade the dominant technology development vehicle for advanced CMOS technologies. Reduced device dimensions and operating voltages that accompany technology scaling have led to increased design challenges with each successive technology node. This is especially true for the 6T SRAM cell that often uses minimum device dimensions and requires many waived design rules to achieve its aggressive density targets. Despite these challenges, the 6T SRAM is expected to continue to play a dominant role in future technology generations because of its combination of density, performance, and compatibility with the CMOS logic process. The successful commercial scaling of the 6T SRAM driven by strong industry competition has followed a well defined linear shrink factor of 0.7X over multiple generations, which results in a fairly predictable 2X reduction in cell area per generation. This continued trend in area reduction is accompanied by the well known consequence of increased variance associated with the reduced channel area. Although technology options such as high-*k* with metal gate have provided some relief in variation, the level of integration and functional margins beyond the 28/32 nm generation pose a serious technical challenge.

A unique feature of the 6T SRAM is an inherent trade off between stability when holding data during a read or non-column selected write access and the ability of the cell to be written. This fact means that the device dimensions and threshold voltage targets established for the SRAM devices are a compromise by design. The ability to read and write will be characterized in terms of margins to assess the functional implications. These margins, which we will refer to as write margin (WM), and read static noise margin (RSNM) or static noise margin (SNM), tend to decrease with scaling. Reduced functional margins combined with the growth in bit count and increased variation with each successive generation, lead to a mounting concern for the viability of the 6T SRAM in future generations.

Circuit assist techniques will become increasingly necessary to preserve the 6T cell functional window of operation as scaling con-

* Corresponding author.
E-mail addresses: rwm3p@virginia.edu (R.W. Mann), jjwang@virginia.edu (J. Wang), svn2u@virginia.edu (S. Nalam), sk4fs@virginia.edu (S. Khanna), geordie@us.ibm.com (G. Braceras), hpilo@us.ibm.com (H. Pilo), bcalhoun@virginia.edu (B.H. Calhoun).

0038-1101/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sse.2010.06.009

tinues. A range of SRAM functional assist methods have been proposed and discussed [1–23], however there remains no clear agreement in the industry as to which method or combination of methods will emerge as the more optimal solution. While different works compare the assist features in varied settings of technology node and technology type, often little detail is given on the trade offs involved in the selection process. Although power and cost are clearly important factors in determining the optimal assist method, it is first necessary to determine if an assist method will meet the functional margin and delay requirements. Once the assist methods which meet the functional requirements are established, the power and implementation costs can by weighed. The goal of this paper is to provide an approach for assessing the functional effectiveness of the assist methods. A second objective is to explore the impact of CMOS scaling trends on the robustness of various assist methods. The specific contributions of this paper are:

- A margin/delay analysis method is developed for the evaluation of the functional effectiveness of circuit assist methods in 6T SRAM.
- A concurrent analysis across four technology nodes to explore the potential impacts of scaling in low power bulk CMOS technologies.
- A concise overview, and method for categorizing the 6T SRAM assist options.

2. Assist categories

A categorization of the assist methods is introduced to establish a systematic means of characterizing the range of circuit assist techniques used in this discussion. For a given foundry cell design, there are three distinct circuit types or categories to address the reduced window of functionality for the 6T SRAM:

1. Altering noise source amplitude or duration through the access transistor.
2. Modification of the latch strength or voltage transfer characteristics of the latch inverters.
3. Avoidance or masking by design or architecture methods.

While category 3 is included for thoroughness and encompasses a range of approaches including ECC masking or prohibiting the half select issue during a write operation [1], the scope of this work will focus on the bias based methods as defined by type 1 and 2. A categorized summary of the bias based circuit assist methods is shown in Table 1. The assist type given in Table 1 provides the primary mechanistic explanation for the assist method effectiveness.

Table 1
Summary of SRAM circuit assist methods with predominant assist type.

Read assist	Type	Write assist	Type	Terminal(s)
Raise VDD	2	Raise VDD	1	global[a]
Raise VDD at cell	2	Reduce VDD at cell	2	VDDc
Reduce VSS at cell	2	Raise VSS at cell	2	VSSc
WL droop	1	WL boost	1	WL
Reduce Q on BLs[b]	1	Increase (BL–BLB)	1	BL & or BLB
Weaken pass gate NMOS	1	Strengthen pass gate NMOS	1	array PWELL[c]
Strengthen pull-up PMOS	2	Weaken pull-up PMOS	2	array NWELL

[a] VDD applied to terminals VDDc, WL, NWELL (BL and BLB for read, BL or BLB for write).
[b] Reduced voltage or capacitance on BL.
[c] Well bias also modulates pull-down NMOS device in most bulk technologies.

While the category types are useful for quickly analyzing the various assist techniques, they are not fundamentally exclusive, and in some cases both mechanisms influence the net assist effectiveness as we will discuss in more detail in Section 6.

The read and write assist methods listed in Table 1 can and in many cases are used in combination, and most can be implemented in either a static or dynamic mode. The categories can be further distinguished by the voltage terminal or terminals which are manipulated. For example a change in the WL voltage would involve modifying one voltage level while a change in the global VDD would involve changing the voltage on five of the seven available terminals associated with the 6T SRAM cell (VDDc, NWELL, WL, BL and BLB). Increased global VDD is unique for several reasons and will be discussed in more detail in Section 5. Modification of the cell design parameters such as device WL, or device threshold voltage by process change or by means beyond the control of the circuit designer, are outside the scope of this paper.

3. Review of assist methods

A brief overview of circuit assist methods published over the last 5 years will support the objectives of this paper, but the large number of publications prevents an exhaustive review here. It is sufficient for this purpose to provide a sample of the options that have been proposed and to allow us to discuss some of the major advantages and disadvantages in context of the categories and terminal access options given in Table 1.

3.1. Read assist

Those read assist methods we categorize as type 1 include methods that reduce the noise source amplitude or duration, which impact the storage latch. These include those methods we shall refer to as write-back [2–4], reduced word line gate voltage [5–9], increased word line (pass gate) threshold voltage through body bias [10,11], and reduced bit line charge by lowering the voltage or capacitance [3,12–14]. The methods we categorize as type 2, which are intended to improve the resilience of the latch, are increased array VDD [6,15–18], decreased array VSS [7]and reduction in the absolute value of the SRAM pull-up PMOS threshold voltage [10].While some techniques such as write-back (or read–modify–write) are purely dynamic in nature, those techniques which involve altering the well (NWELL or PWELL) bias are proposed as primarily static implementations due to the large RC delay or layout complexity that would be involved in making this technique dynamic. The embodiments proposed as assists in [10,19] are essentially fixed biases set at one point in time to provide some compensation for global variation.

3.2. Write assist

A roughly equal number of publications are invested in the challenge associated with writing the 6T SRAM. The read/write assist symmetry observed from Table 1 is worth noting, and all but one method (increased global VDD) have the not so surprising opposite effect on read stability versus ability to write. Publications that address the challenge of writing the cell following category 1 (increased amplitude or duration of the write signal through the pass gate device) have proposed some form of boost to the word line gate voltage [6,15,20,16] or negative bit line voltage [7,21,9] to increase the VGS 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 [2,5,6,8,

11,17], raising the array VSSc [12,22,23], or reducing the strength the pull-up PMOS by NWELL bias [10,19].

4. Assist metrics

The primary objective of this work will be focused on the functional metrics of margin sensitivity and performance. The metrics of power and cost will be addressed in Section 6 in context of this primary objective. In this section we define and quantify of the margin and performance metrics used in this analysis.

4.1. Margin sensitivity

We define the margin sensitivity as the change in margin with respect to the change in applied assist bias voltage for a given technique. This is expressed as:

$$Sensitivity = \frac{\partial(Margin)}{\partial V} \tag{1}$$

Margin may refer in this case to either SNM or WM. To compare the margin sensitivity of the specific assist methods, we perform noise margin analysis using custom predictive technology models (PTMs) [24,25] using pre-defined scaled SRAM dimensions consistent with the dense SRAM published values. The defined margin sensitivity is a useful metric for quantitatively comparing assist method effectiveness. It is applicable to all bias based assist methods, provides an objective means of comparing the assist methods to one another and also across the technology nodes.

Because bias limitations of some form exist for all assist methods, the margin sensitivity provides a means of quantitatively determining the attainable margin improvement. Depending on the assist method used, different limiting factors will constrain the terminal bias values that can be applied. In the case of boosting schemes such as +WL(write), neg BL(write), +VDDc(read) or −VSSc(read), the common limiting factor is the technology Vmax. Voltage suppression schemes such as +VSSc(write), −VDDc(write) or −WL(read) are limited by different mechanisms. For example, the bias used collapsing the supply voltage (+VSSc or −VDDc), becomes limited by data retention fails for unaccessed cells that share the collapsed supply. For −WL(read), performance limitations can quickly limit the allowable bias available for read stability margin gains obtained with reduced word line voltage.

The nominal VDD is based on published industry values for the nodes of interest. We used 1.2 V, 1.1 V, 1.1 V and 1.0 V for 65 nm, 45 nm, 32 nm, and 22 nm respectively. As part of the methodology defined in this investigation, we will place particular emphasis on the specific conditions that represent the worst case operation voltage (Vwc) for the technology. We define Vwc as the minimum voltage at which the SRAM must be able to perform both a read and write operation across the entire array without failure. Thus, we need to ensure that the VDDmin[1] for a given array is at or below our predefined Vwc for each technology node. Because Vwc is recognized as technology and application dependent, we will use 0.8X the nominal VDD as this value. This condition accounts for factors such as voltage droop, NBTI shifts over the product lifetime, and testing equipment variability.

In addition to the shift in the mean margin value, we also examine variation and the impact of the assist methods on the margin distribution in Section 5. This is a critical point since the ultimate goal of the assist technique is to improve the yield at the Vwc or lower the VDDmin of a particular array.

[1] While non-foundry or in-house designs may have the flexibility to push the operation voltage to the empirically defined VDDmin, foundry based design kits frequently specify a valid model operation voltage range. Designing outside this specified range (below Vwc) may produce invalid results.

4.2. Performance

The performance for a given assist method is evaluated using write delay for the write assist method and the time required for bit line signal development for read assist. For this analysis, we are concerned about the deltas between techniques. This simplifies the analysis and allows us to focus specifically on the two performance components of interest. The delay can be reduced to the time required to charge the word line (τ_{WL}), plus the time required to develop a sufficient differential voltage on the BL ($\tau_{\Delta BL}$)to set the sense amplifier.

$$\tau_{read} = \tau_{WL} + \tau_{\Delta BL} \tag{2}$$

To briefly illustrate how the assist method may impact the τ_{Read}, the read assist method of reduced WL voltage is considered. For this example, the τ_{WL} will be reduced while the $\tau_{\Delta BL}$ will be increased.

Following a similar approach as with the read performance evaluation, considering the deltas associated with the assist methods for comparison purposes, the write performance (τ_{write}) estimate will require three components as given in (3).

$$\tau_{write} = \tau_{BL} + \tau_{wcell} + \tau_{WL} \tag{3}$$

The value τ_{WL} is consistent with the previous definition, and τ_{BL} is the delay (or part of the delay that does not overlap with τ_{WL}) required to establish the BL–BLB voltage differential for the write operation. τ_{wcell} is the delay associated with the cell state change given the applied BL differential and WL voltage. Simulations will be used to quantify τ_{wcell} in this study.

4.3. Margin/delay analysis

A margin sensitivity factor and performance factor will be employed to derive a final effectiveness factor and a graphical (margin/delay) space analysis will also be used [26]. To illustrate the margin/delay approach, Fig. 3 shows a schematic diagram depicting the desired functional window, delineated by the margin and delay requirements of the memory. As the VDD is reduced to Vwc, the read/write margins and corresponding performance degrade. Use of assist methods generally improves margin and in most cases delay to some extent. We propose that plotting the margin versus delay of a memory with varying amounts of assist bias will illuminate the most effective assist methods for a given technology and set of functional requirements. This graphical approach provides additional insight into the net functional impact of a given assist method and allows us to readily understand potential advantages and trade offs of a given assist approach.

5. Results

Four read assist and four write assist methods were examined to provide a set of test cases for the assist evaluation methodology. A schematic representation of the specific assist methods explored is given in Figs. 1 and 2 for read assist and write assist respectively. Three of the read assist methods chosen for this evaluation were of type 2 category and one (WL droop) was type 1. Two of the write assist methods chosen for this evaluation were from type 1, and the remaining two were type 2. The four read assist methods shown are listed in Table 1 rows 1–4. The four write assist methods discussed in this work are given in Table 1 rows 2–5. Those assist methods that are inherently dynamic (influencing the duration of the noise source) must be evaluated using dynamic noise margin methods. These include reduced BL capacitance and read modify write or write-back.

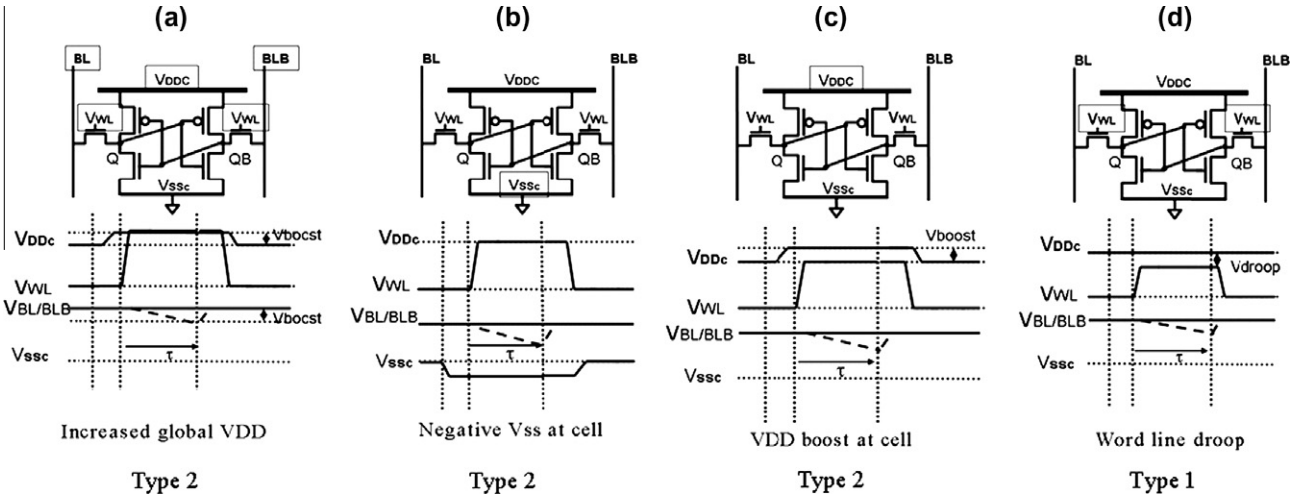


Fig. 1. Schematic timing diagram representations for read assist (a) raised array global VDD, (b) negative VSS at the cell, (c) VDD boost at the cell and (d) WL droop. τ represents the time for the sense amplifier to set. Text box denotes modulated terminal(s).

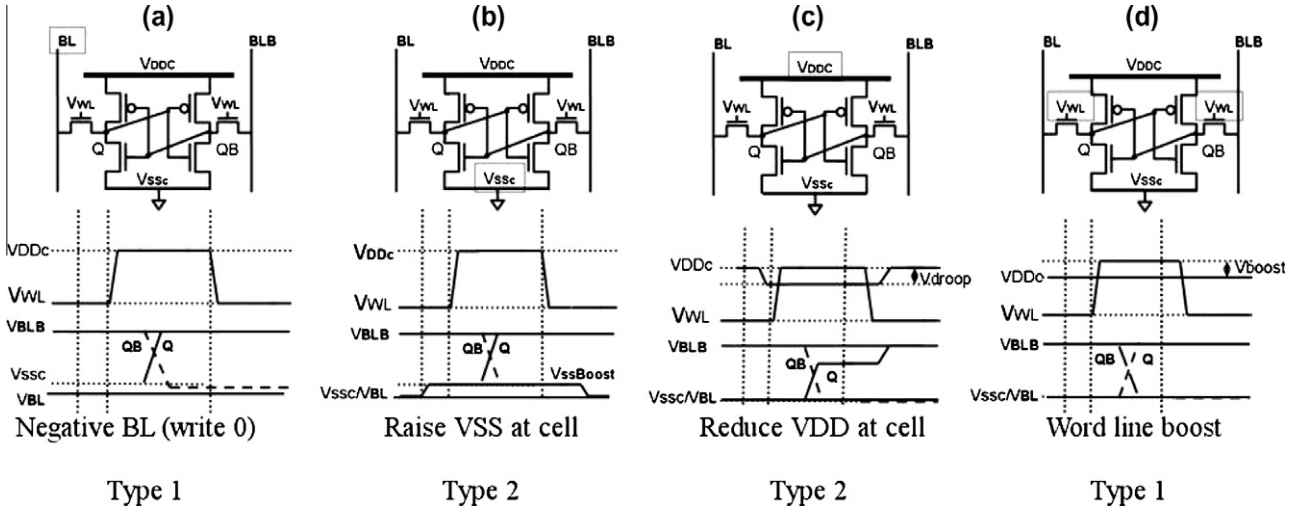


Fig. 2. Schematic representations for write assist (a) negative BL, (b) raised VSS at the cell, (c) VDD droop at the cell and (d) WL boost. Text box denotes modulated terminal(s). Node voltage Q represented by dashed line in schematic timing diagram.

5.1. Simulation results – margin

To quantify the margin sensitivities in this study, static metrics will be used to emulate the functional environment using the custom low power (LP) PTM bulk technologies [25]. For read assist, SNM based on the butterfly curve analysis is used [27]. For write assist, the ramped WL based metric will be employed [28], defined as the (VWLmax – VWLflip) to assess the margin due to its improved correlation to dynamic write margin [29]. A yield analysis will be used to establish quantitative relationships of the required margins.

Fig. 4a–d plots the SNM as a function of the assist bias for the four read assist techniques defined in Fig. 1a–d. The four technology nodes are represented in each of the four plots. Fig. 4c for example shows the change in SNM with increased array VDD (VDDc) as described schematically in Fig. 1c. There is a negative slope for methods (b) and (d) corresponding with the fact that these methods utilize a reduction in the terminal voltage. While all four methods produced some degree of improvement in the SNM, and the response or sensitivity is similar across the technology nodes, the sensitivity was most significant for VDDc, Fig. 1c and Fig. 4c. The initial voltage is either 0 V or varies consistently with the Vwc for each technology.

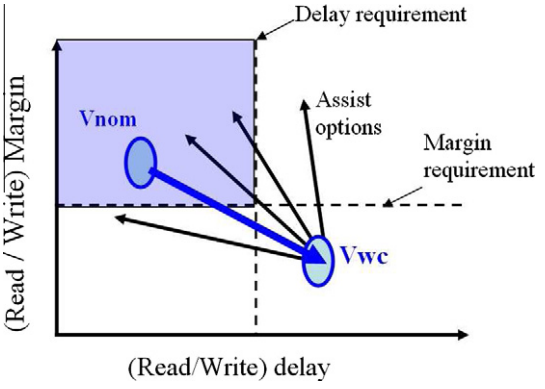


Fig. 3. Schematic diagram of read/write margin vs read/write delay and desired functional window based on margin limited yield and performance requirements for application [26].

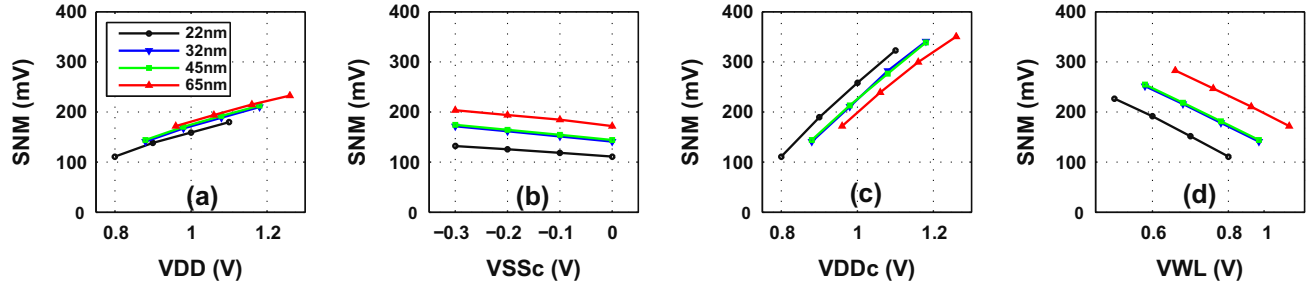


Fig. 4. Read static noise margin as function of (a) raised array global VDD, (b) Negative VSS at the cell, (c) VDD boost at the cell (VDDc) and (d) WL droop [26].

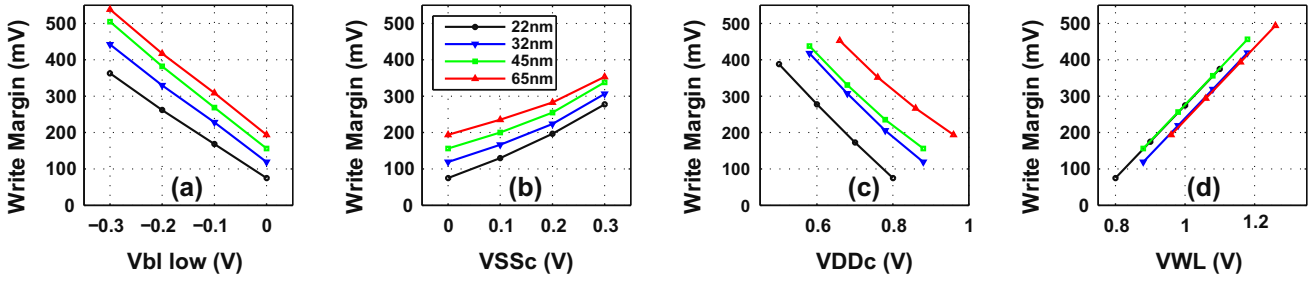


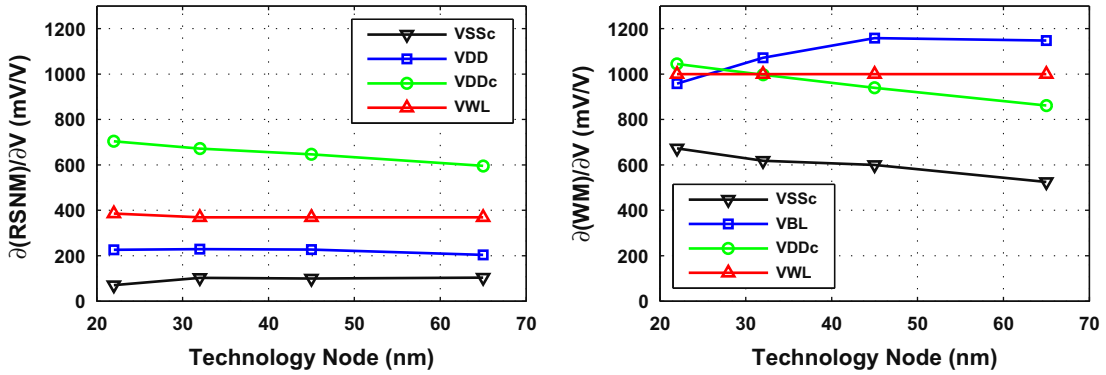
Fig. 5. Write margin as function of (a) negative BL, (b) raised VSS at the cell (VSSc), (c) VDD droop at the cell (VDDc) and (d) WL boost [26].

The simulation results for the write assist methods are shown in Fig. 5a–d corresponding with the conditions defined in Fig. 2a–d. For the write assist methods in this analysis, the VSSc response, Fig. 5b, was the least linear and showed the least sensitivity. Although there is some degree of non-linearity in the response characteristics of write margin and static noise margin, most exhibit a sufficient degree of linearity across the 300 mV range to allow us to characterize the responses using a first order linear model to allow a high level comparison. SNM sensitivities shown in Fig. 4a–d are summarized for each of the technology nodes in Fig. 6a. As a means of improving the SNM, raised cell voltage (VDDc) is the method that emerges as exhibiting the greatest sensitivity across the LP technology nodes. The trends also suggest that there is some increase in sensitivity as scaling continues.

Following a similar approach, we also characterize the functional sensitivities across the technology nodes for write margin sensitivity, Fig. 6b. In this case, three of the methods exhibit similar sensitivities to the applied bias. Raised array VSS (VSSc) showed less degree of linearity and had a weaker response. The unique and completely linear response of the WL boost was due to the fact that the write margin metric used in this investigation was defined as the difference between the final word line voltage and the voltage of the word line required to write the cell.

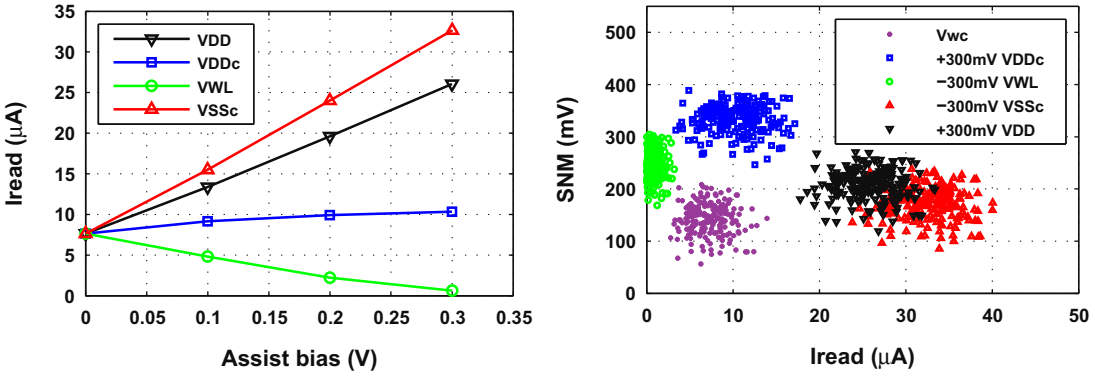
5.2. Simulation results – performance

The relationship between read current and read SNM is of particular concern with scaled technologies as the read currents are generally decreasing with successive generation. The read assist methods have an important and significant impact on the cell read current. The influence of the read assist methods on the read current for the 45 nm node is shown in Fig. 7a with the initial value representing no assist technique at the low voltage corner (Vwc). Fig. 7b further plots the spread of read current vs SNM at 300 mV assist bias. Although only the 45 nm technology data is shown, the other three technology nodes responded in a similar



(a) SNM sensitivity by node (b) WM sensitivity by node

Fig. 6. The margin sensitivities across LP technologies for the four read assist methods (a) and four write assist methods (b) investigated.



(a) 45nm LP read current vs assist bias (b) 45nm LP SNM vs read current (Monte Carlo data)

Fig. 7. The impact of read assist bias conditions on the bit cell read current (a) and SNM versus Iread for Vwc and 300 mV of assist bias (b). Data shown is for the 45 nm technology node.

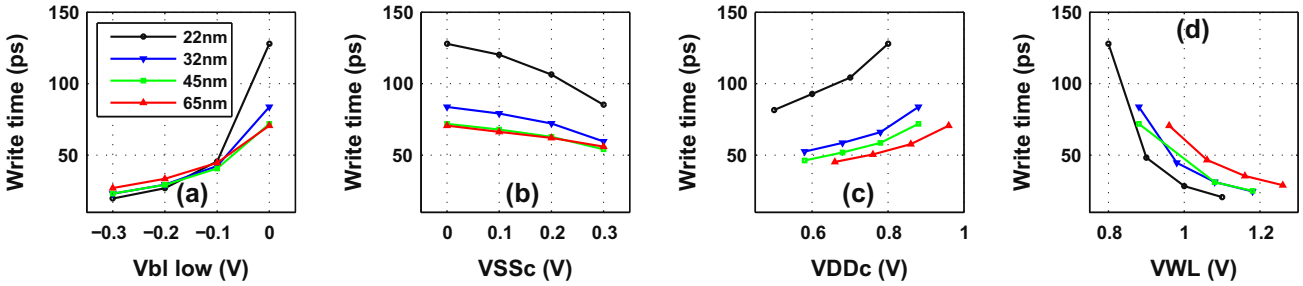


Fig. 8. Effect of write assist techniques on cell component of write time (a) negative BL voltage, (b) raised cell Vss, (c) reduced VDD as the cell and (d) boosted WL voltage.

way. Increased array voltage (VDDc) has only a small positive impact on the read current, while reduced word line voltage significantly degraded the read current. Decreasing the VSSc terminal below GND resulted in the strongest improvement in read current, exceeding that of conventional VDD increase. This results from both increased VGS and reduced threshold voltage in the SRAM cell pull-down (PD) NMOS device due to the body effect. The read performance impact of the read assist techniques can be estimated for each technique with (2). Based on the simple relationship provided in (3), the performance limitations associated with the WL droop can quickly become prohibitive.

The delay impact associated with the cell write time (τ_{wcell}) is shown in Fig. 8a–d for the four write assist methods evaluated. While all four methods improved the write time, WL boost and negative BL voltage bias schemes showed a more significant improvement in delay. Increasing the cell VSS and reducing the cell or array VDD had less impact. The delay response for cell write time was similar with scaling although the 22 nm node showed a stronger initial response to the applied bias conditions. For the negative BL and boosted WL cases, the 22 nm delay response is most dramatically influenced by the 0.3 V applied assist bias.

5.3. Impact of assist methods on variation

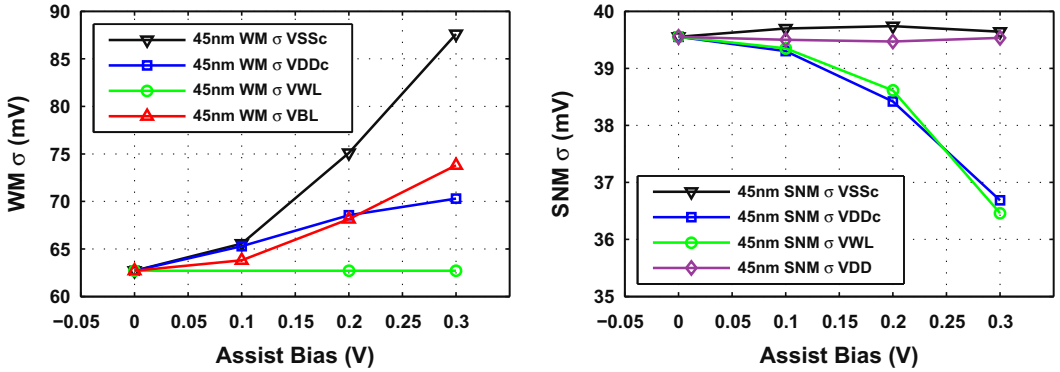
Until now, we have discussed only the impact of the voltage deviations of the assist methods on the mean values of SNM and WM at a given bias condition. However, to determine the functional yield expectation for a given array size at the worst case voltage, the local and global variation must be taken into account. Without the variation component, the required margin improvement will be unknown. For the small scaled SRAM devices, the local variation associated with random dopant fluctuations (RDF) dominates the variation sources. Although technology improvements offered by high-k and metal gate, may provide significant improvement due to the higher gate capacitance, continued scaling will quickly consume these gains.

To address the impact of the assist methods on the variation in both SNM and WM Monte Carlo simulations were run for each method that we explored in this paper. Fig. 9 plots the sigma for the WM distribution (a) and SNM (b) as a function of the assist voltage bias for the 45 nm node. A minimum of 200 Monte Carlo runs were performed for each bias condition. Several observations emerge from this analysis. We first observe that the assist method and bias both impact the standard deviation of the distribution. We account for this in assessing the overall contribution of the assist method which we will discuss in the next section.

An additional source of variation in assist response can be caused by voltage variations on the assist modulated terminal(s). This variation will strongly depend on the specific design and assist implementation scheme used. The sensitivity metric, discussed in Section 4.1, provides a means of assessing the overall impact of this variation source by relating changes in terminal voltage to margin.

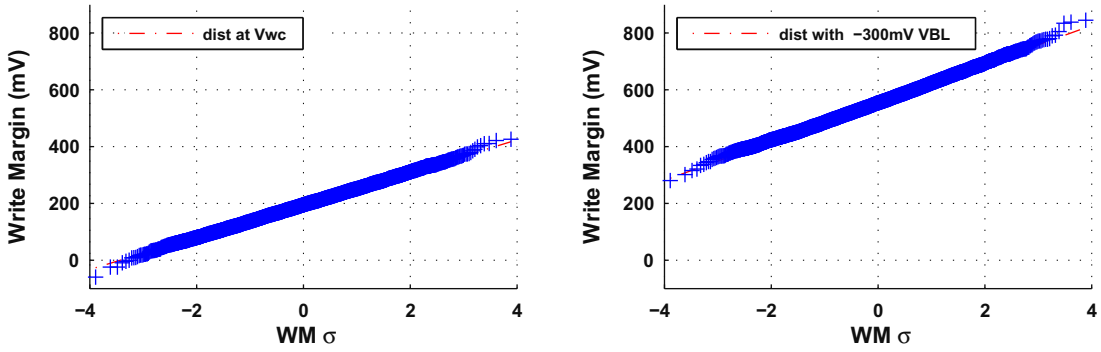
5.4. Yield quantification

To identify the functional window requirement as depicted in Fig. 3, it is necessary to be able to convert the simulated margin information into yield. Soft fails are 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 to set the sense amp), (3) stability upset, and (4) data retention. These four failure modes are not attributable to defects but are instead associated with a distribution tail stemming from variation sources. Although we do not address read fails and data



(a) 45nm LP WM σ vs assist bias (b) 45nm LP SNM σ vs assist bias

Fig. 9. Impact of assist method applied bias on the sigma of the resulting 45 nm LP technology distribution for write assist (a) and read assist (b).



(a) WM(0) distribution at Vwc for 45nm LP (b) WM(0) distribution at Vwc with-300mV VBL bias for 45nm LP

Fig. 10. 10,000 Monte Carlo cases showing WM(0) standard normal distribution for 45 nm LP technology at Vwc with no assist bias (a) and with 300 mV negative BL bias (b).

retention fails directly, assist method choices can clearly impact these mechanisms. The assist methods are directed at mechanisms 1 and 3. To address the write and stability related yields quantitatively we use the following approach.

SNM0/WM0 denote the read/write margin for data '0' and SNM1/WM1 denote the margin for data '1'. The definition of SNM/WM would be the minimum value for '0' and '1'. An important observation is that the distribution of SNM0 or SNM1 can be represented by a standard normal distribution under normally distributed parameter variation. This same observation is true for WM0 or WM1. For the cases we examined, the distributions remain normally distributed with assist bias, though the mean and the standard deviation may change. An additional set of Monte Carlo simulations (1000–10,000 cases) were run on selected assist bias conditions for distribution verification purposes. Fig. 10 shows the results of 10,000 cases for WM0 at Vwc (a) and with 300 mV negative BL bias (b) for the 45 nm LP technology. The linearity of the quantile plots confirms that the WM distribution remains normal even with the assist feature engaged. The failure probability (Pf) for the right or left node (probability of SNM0 < 0 or SNM1 < 0 for example) is given as:

$$Pf = \frac{1}{2} erfc\left(\frac{\eta_\sigma}{\sqrt{2}}\right) \tag{4}$$

where η_σ is defined as the number of random variable standard deviations from the mean based on the standard normal distribution. For large arrays with relatively few fails, the Poisson distribution will be used to estimate the soft fail limited yield. We can then compute (λ) defined as the number of bits (N) times the fail probability (Pf), including both states of the latch:

$$\lambda = N \cdot (Pf_{(0)} + Pf_{(1)}) \tag{5}$$

With the assumption that the RDF induced variations are random and non-clustered, the soft fail yield (without redundancy) for a given mechanism can be expressed as:

$$Yield = exp(-\lambda) \tag{6}$$

To obtain a 10 Mb SRAM with a SNM-limited yield of 99% would require a η_σ value of 6.12σ. In other words, to achieve this yield target, SNM0wc must be larger than the minimum noise margin threshold (in this case 0) for 99 of 100 10 Mb arrays. The limited yield for WM is computed with this same approach, to obtain a 99% WM-limited yield, which would result in an over all soft fail limited yield of 98% considering both WM and SNM. For our 45 nm LP technology, Fig. 11 shows that this is achieved with 180 mV for either word line boost or negative BL bias (a) and 100 mV assist bias for the most effective read assist technique (VDDc boost) (b).

6. Discussion

We have outlined the elements of both margin and delay referenced to Vwc, and provided a means of transforming the write and read margins into a soft fail limited yield value. This approach has been applied and demonstrated using the LP PTM platform of bulk technologies from 65 nm to 22 nm.

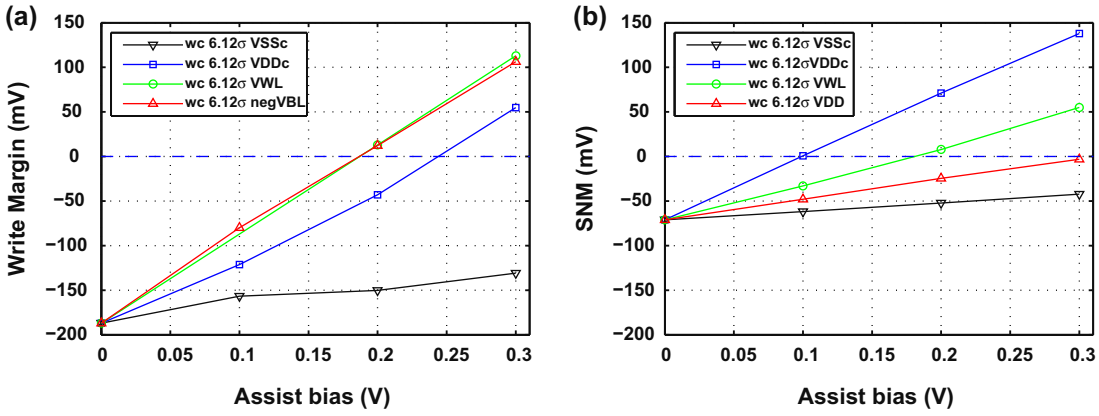


Fig. 11. The 6.12σ worst case (wc) write margin (a) and SNM (b) as a function of assist bias for the 45 nm LP technology.

6.1. Assessing functional effectiveness

The functional read/write margin sensitivity was evaluated over a 300 mV window to minimize non-linearity in the response and to ensure the bias conditions would not exceed the technology reliability limits. Because our reference (Vwc) condition was more than 200 mV below nominal VDD in all cases, the reliability requirement was preserved. Even for the 22 nm node where the Vwc was taken to be 0.8 V, the max voltage would be only 10% greater than nominal VDD, which is consistent with common technology specifications.

The sensitivity response for the assist methods studied is often influenced by more than one mechanism and can be understood when the device physics are taken into account. For example, the superior result associated with raised array voltage (VDDc) for read assist can be attributed to the fact that several mechanisms influence the result. The body effect causes the cell PFET to become stronger because of the modulated VSB for the PFET and the VGS is increased for the devices in the latch which are on.

6.2. Margin/delay space method

An example of the margin/delay plot introduced earlier is shown in Fig. 12a showing the write margin versus write delay for each of the four assist methods evaluated. The different assist methods portray varying trajectories in the margin/delay space, and the type 1 methods are shown to increase margin while decreasing delay most effectively. Fig. 12b shows the assist trajectories in margin/delay space for the read assist methods evaluated. A combined VDDc and VSSc assist method is shown in Fig. 12b which demonstrates how the assist techniques can be combined as required to optimize both delay and margin. This figure also points out that some assist methods, such as WL droop, may improve the margin while simultaneously degrading the performance. Using this analysis approach, the methods categorized as type 2 were more effective for read assist.

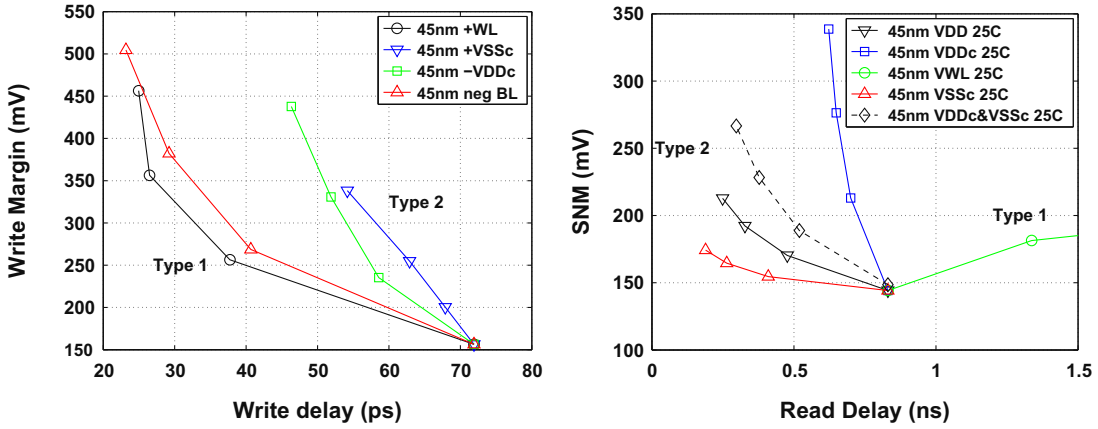
The effect of variation was examined in some detail and it was found that both assist method and bias had a non-negligible impact on the resulting WM and SNM distributions. For those cases where the assist method influenced the distribution, it was necessary to account for this in determining the effectiveness of a given method on the yield. While the SNM and WM distributions are intrinsically non-Gaussian for reasons previously discussed, relying on the distributions which are normally distributed, we compute the distribution tail. By this method, a required assist bias for a given array size and soft fail yield requirement for both WM and SNM can be established.

6.3. Practical considerations

To assess the complexity of implementation for specific assist methods, yield implications associated with the specific assist method should be considered. For example, of the four write assist methods we investigated, three (WL or VSSc boost, and VDDc droop) require a higher, yield related complexity. This is because WL boost increases the potential for stability upset in the cells along the asserted word line on the non-selected columns, and reduced voltage at the cell by VSSc boost or VDDc droop introduces data retention concerns. The trade off in the stability (SNM) impact of the half-selected bits during a write assist is shown in Fig. 13 for both negative BL and WL boost assists for 45 nm. Although the negative BL method partially avoids these yield implications, the added level shift circuit complexity of generating the negative voltage must be considered.

To address cell layout compatibility with a given assist method, it is noted that the 6T cell is typically provided by the foundry and therefore constrains the memory array designer to seek assist methods that best comply with the given layout. For example, the predominant industry 6T cell design style makes use of a VDD bus on metal level 2 (M2) level running parallel with the M2 bit lines. Although this layout style has advantages for density and performance reasons, the implementation of locally raising VDDc along the word line requires that all columns on the selected WL be boosted. Although pulsing the VSSc may be more consistent with this style cell layout (the metal 3 (M3) VSSc bus which runs parallel with the M3 word local line), we found this technique exhibited less margin sensitivity. It should also be pointed out that assist compatibility with dual port SRAM is of emerging importance, and some methods such as drooped VDDc for write assist are fundamentally incompatible. For those applications requiring both 1 and 2 port SRAM, the cost effectiveness for an approach such as the negative BL may become more compelling.

For those methods deemed most effective based on functional sensitivity and performance, the cell compatibility and yield complexity are considered together. Along with raised global VDD, four additional combinations of assist methods would need to be considered. Considering the predominant industry cell layout style, the comparison may then be summarized in Table 2. For the LP bulk technologies considered in this study, both read and write assist would be required to achieve high yield for large SRAM arrays beyond 65 nm. Combining both the functional effectiveness requirement with the requirement that the cell layout must be compatible with the predominantly used industry bit cell, results in five pairs of options. By introducing the additional constraint that the yield complexity be low, the viable assist combinations re-



(a) Write margin vs write delay (b) Read margin vs read delay

Fig. 12. Margin vs delay plots showing write (a) and read (b) for the 45 nm LP technology when assist bias is swept from 0 to 300 mV [26].

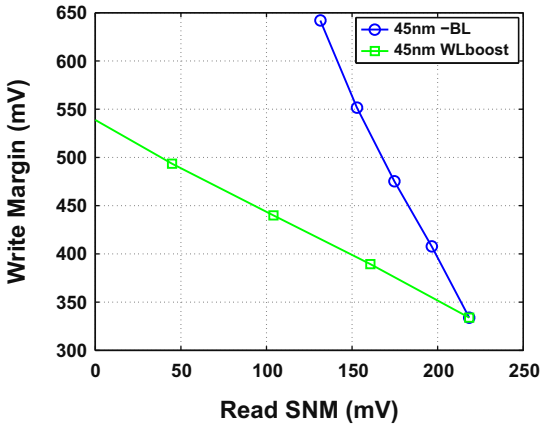


Fig. 13. Impact of write assist on stability of the half-selected bits on the asserted word line shown for 45 nm LP. As word-line-boost or negative-bit-line assist increases the write margin, the SNM is reduced for those bits on the word line subjected to a dummy read condition.

Table 2
Practical considerations for viable assist combinations.

Read assist method	Write assist method	Cell compatible	Low yield complexity
Raise VDD	Raise VDD	Yes	Yes
−VSSc	+WL	Yes	No
−VSSc	−BL	Yes	Yes
+VDDc	+WL	Yes[a]	No
+VDDc	−BL	Yes[a]	Yes

[a] VDDc boost required for all columns on asserted WL.

duce to three. For a final selection between the remaining combinations of assist methods, absolute margin and performance deltas should be considered along with factors such as power and area overhead. An assessment of area overhead is dependent on the specific implementation scheme and therefore beyond the scope of this paper, however, an area overhead of less than 4% would be expected for a competitive implementation [2,5,6,9,15].

6.4. Power

Power is a critical criteria for the ultimate selection of an assist method, however, power is dependent on both the assist method and implementation scheme. This is demonstrated by examining the essential components of SRAM array power. Both read and write operations are first described without assist and then for a specific read assist operation to illustrate this point.

The dominant components of power for a single read operation, without a circuit bias assist is given by:

$$P_{read} = P_{WL} + P_{\Delta BL} \tag{7}$$

where the components of read power are consistent with those in Eq. (2). The P_{WL} describes the power associated with the WL pulse and $P_{\Delta BL}$ refers to the power associated with the change in voltage on the BL's along the asserted WL. This read power may be expressed more fully as:

$$P_{read} = f(N_{BL}C_{WLc}V_{dd}^2 + N_{WL}C_{BLc}\Delta V_{BL}V_{dd}) \tag{8}$$

where f is the frequency, C_{BLc} and C_{WLc} are the bit line and word line capacitance per cell, N_{WL} and N_{BL} are the total number of word lines and bit lines in the array block of interest. The voltage differential required to set the sense amplifier is (ΔV_{BL}). The primary considerations for write power may be expressed as:

$$P_{write} = P_{BL\to 0} + P_{cell} + P_{WL} + P_{\Delta BL} \tag{9}$$

where the first three components of write power are consistent with those given in Eq. (3). Although not a contributor to write delay, $P_{\Delta BL}$ is a non-negligible component of the write power. The $P_{\Delta BL}$ term accounts for the power associated with the voltage change on the BL's along the asserted WL for the half-selected cells, i.e., those subjected to a dummy read operation. The $P_{BL\to 0}$ is the power associated with the BL discharge to ground for the write operation, P_{cell} is the power associated with writing the column selected cells on the word line, and P_{WL} describes the power due to the write WL pulse. The write power may be expressed more fully as:

$$P_{write} = f\Big(N_{SBL}N_{WL}C_{BLc}V_{dd}^2 + N_{SBL}C_{cell}V_{dd}^2 + N_{WL}N_{BL}C_{WLc}V_{dd}^2 + (N_{BL} - N_{SBL})\Delta V_{BL}V_{dd}\Big) \tag{10}$$

with N_{SBL} used to refer to the number of bit lines that are selected for the write operation.

A first order assessment in the change in power associated with a given assist method can be derived from these equations. For example, the change in power associated with the WL droop read assist can be expressed as:

$$\Delta P_{read} = f(N_{BL}C_{WLc}\left(V_{\Delta WL}^2 - 2V_{\Delta WL}V_{dd}\right)) + P_{assist} \tag{11}$$

where $V_{\Delta WL}$ is the voltage reduction on the WL, P_{assist} is the power expended by the specific assist scheme chosen. The power associated with achieving the dynamic voltage reduction in the WL (P_{assist}), would also need to be included in the analysis. For example, the use of a replica or set of replica pass gate devices [5], which lower the WL voltage but also provide a DC path to ground during the WL pulse, would constitute a non-negligible P_{assist} when assessing the overall power impact. A similar analysis can be used for each assist method and implementation scheme. It is also clear from this analysis that the power will be dependent on specific array configuration factors, e.g., N_{BL}, N_{WL}, and N_{SBL}. In addition to the specific assist implementation scheme and array configuration, the cell and array layout configuration is also an important factor. For example, it would follow from this analysis method that the power impact of dynamically modulating the array supply bus for VDDc assist, with the conventional 6T layout and the array layout configuration discussed in Section 6.3, may easily be large compared to other dynamic schemes.

A first principles analysis of relevant power components for both read and write without assist bias schemes was shown. Using this analysis it is also shown that determining the power for a given assist method requires specific details of the assist scheme and layout configuration. Because of the significant differences in margin sensitivity and performance across the assist methods, it is recommended that assessing the implementation costs and power be evaluated after determining the methods which are shown to satisfy the product functional requirements.

7. Conclusions

As competitive forces and industry scaling continue to erode the 6T SRAM functional margins, the use of assist methods will increase. A review and categorization approach for examining potential bias based assist methods is provided. For the assist methods evaluated in this study using the LP bulk CMOS technologies, those methods categorized as predominantly type 1 are more effective for write assist and the predominantly type 2 category of assist methods are more effective for read assist. The assist methods exhibited some degree of consistency across the platform of LP technologies studied. This suggests that the design infrastructure and assist method implementation learning can be reduced with reuse across multiple generations. The margin/delay analysis was demonstrated as an objective means of evaluating the influence on the functional metrics by the assist methods. Based on a margin/delay analysis and practical considerations, the more viable assist methods for future investment were identified, however, for a final selection additional factors such as implementation cost and power will need to be included in the analysis.

References

[1] Chang L, Fried DM, Hergenrother J, Sleight JW, Dennard RH, Montoye RK, et al. Stable SRAM cell design for the 32 nm node and beyond. In: Proc digest of technical papers VLSI technology 2005 symposium; 2005. p. 128–9.

[2] Pilo H, Barwin J, Braceras G, Browning C, Burns S, Gabric J, et al. An SRAM design in 65 nm and 45 nm technology nodes featuring read and write-assist circuits to expand operating voltage. In: Proc digest of technical papers VLSI circuits 2006 symposium; 2006. p. 15–6.

[3] Khellah M, Ye Y, Kim NS, Somasekhar D, Pandya G, Farhang A, et al. Wordline & bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65 nm CMOS designs. In: Proc digest of technical papers VLSI circuits 2006 symposium; 2006. p. 9–10.

[4] Kushida K, Suzuki A, Fukano G, Kawasumi A, Hirabayashi O, Takeyama Y, et al. A 0.7 V single-supply SRAM with 0.495 um^2 cell in 65 nm technology utilizing self-write-back sense amplifier and cascaded bit line scheme. In: Proc IEEE symposium on VLSI circuits; 2008. pp. 46–7.

[5] Ohbayashi S, Yabuuchi M, Nii K, Tsukamoto Y, Imaoka S, Oda Y, et al. A 65-nm SoC embedded 6T-SRAM designed for manufacturability with read and write operation stabilizing circuits. IEEE J Solid-State Circ 2007;42(4):820–9.

[6] Hirabayashi O, Kawasumi A, Suzuki A, Takeyama Y, Kushida K, Sasaki T, et al. A process-variation-tolerant dual-power-supply SRAM with 0.179 μm^2 cell in 40 nm CMOS using level-programmable wordline driver. In: Proc IEEE international solid-state circuits conference – digest of technical papers ISSCC 2009; 2009/ p. 458–9 and 459a.

[7] Wang DP, Liao HJ, Yamauchi H, Chen YH, Lin YL, Lin SH, et al. A 45 nm dual-port SRAM with write and read capability enhancement at low voltage. In: Proc IEEE international SOC conference; 2007. p. 211–4.

[8] Mohammad B, Saint-Laurent M, Bassett P, Abraham J. Cache design for low power and high yield. In: Proc 9th international symposium on quality electronic design ISQED 2008; 2008. p. 103–7.

[9] Nii K, Yabuuchi M, Tsukamoto Y, Ohbayashi S, Oda Y, Usui K, et al. A 45-nm single-port and dual-port SRAM family with robust read/write stabilizing circuitry under DVFS environment. In: Proc. IEEE symposium on VLSI circuits; 2008. p. 212–3.

[10] Mukhopadhyay S, Mahmoodi H, Roy K. Reduction of parametric failures in sub-100-nm SRAM array using body bias. IEEE Trans Comput – Aided Des Integr Circ Syst 2008;27(1):174–83.

[11] Yamaoka M, Maeda N, Shinozaki Y, Shimazaki Y, Nii K, Shimada S, et al. Low-power embedded SRAM modules with expanded margins for writing. In: Proc digest of technical papers solid-state circuits conference ISSCC. 2005 IEEE international; 2005. p. 480–611.

[12] Bhavnagarwala A, Kosonocky S, Radens C, Stawiasz K, Mann R, Ye Q, et al. Fluctuation limits & scaling opportunities for CMOS SRAM cells. In: Proc IEDM technical digest electron devices meeting IEEE international; 2005. p. 659–62.

[13] Bhavnagarwala AJ, Kosonocky S, Radens C, Chan Y, Stawiasz K, Srinivasan U, et al. A sub-600-mv, fluctuation tolerant 65-nm CMOS SRAM array with dynamic cell biasing. IEEE J Solid-State Circ 2008;43(4):946–55.

[14] Abu-Rahma MH, Anis M, Yoon SS. A robust single supply voltage SRAM read assist technique using selective precharge. In: Proc 34th European solid-state circuits conference ESSCIRC 2008; 2008. p. 234–7.

[15] Chen YH, Chan WM, Chou SY, Liao HJ, Pan HY, Wu JJ, et al. A 0.6 V 45 nm adaptive dual-rail SRAM compiler circuit design for lower VDDmin VLSIs. In: Proc IEEE symposium on VLSI circuits; 2008. p. 210–1.

[16] Chung Y, Song S-H. Implementation of low-voltage static RAM with enhance data stability and circuit speed. Microelectron J 2009;40:944–51.

[17] Zhang K, Bhattacharya U, Chen Z, Hamzaoglu F, Murray D, Vallepalli N, et al. A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply. IEEE J Solid-State Circ 2006;41(1):146–51.

[18] Yamaoka M, Osada K, Ishibashi K. 0.4-V logic-library-friendly SRAM array using rectangular-diffusion cell and delta-boosted-array voltage scheme. IEEE J Solid-State Circ 2004;39(6):934–40.

[19] Yamaoka M, Maeda N, Shimazaki Y, Osada K. 65 nm low-power high-density SRAM operable at 1.0V under 3σ systematic variation using separate Vth monitoring and body bias for NMOS and PMOS. In: Proc. digest of technical papers. IEEE international solid-state circuits conference ISSCC 2008; 2008. p. 384–622.

[20] Iijima M, Seto K, Numa M, Tada A, Ipposhi T. Low power SRAM with boost driver generating pulsed word line voltage for sub-1V operation. JCP 2008;3(5):34–40.

[21] Shibata N, Kiya H, Kurita S, Okamoto H, Tan'no M, Douseki T. A 0.5-V 25-MHz 1-mW 256-kb MTCMOS/SOI SRAM for solar-power-operated portable personal digital equipment – sure write operation by using step-down negatively overdriven bitline scheme. IEEE J Solid-State Circ 2006;41(3):728–42.

[22] Yang HS, Wong R, Hasumi R, Gao Y, Kim NS, Lee DH, et al. Scaling of 32 nm low power SRAM with high-K metal gate. In: Proc IEEE International Electron Devices Meeting IEDM 2008; 2008. p. 1–4.

[23] Suzuki T, Yamauchi H, Yamagami Y, Satomi K, Akamatsu H. A stable 2-port SRAM cell design against simultaneously read/write-disturbed accesses. IEEE J Solid-State Circ 2008;43(9):2109–19.

[24] Zhao W, Cao Y. New generation of predictive technology model for sub-45 nm design exploration. In: Proc 7th international symposium on quality electronic design ISQED '06; 2006. p. 6–590.

[25] Calhoun BH, Khanna S, Mann R, Wang J. Sub-threshold circuit design with shrinking CMOS devices. In: Proc IEEE international symposium on circuits and systems ISCAS 2009; 2009. p. 2541–4.

[26] Mann RW, Nalam S, Wang J, Calhoun BH. Limits of bias based assist methods in nano-scale 6T SRAM. In: Proc 11th international symposium on quality electronic design, ISQED'10; 2010. pp.1–6.

[27] Seevinck E, List FJ, Lohstroh J. Static-noise margin analysis of MOS. SRAM Cells 1987;22(5):748–54.

[28] Gierczynski N, Borot B, Planes N, Brut H. A new combined methodology for write-margin extraction of advanced SRAM. In: IEEE international conference on microelectronic test structures; 2007. ICMTS'07; 2007. p. 97–100.

[29] Wang J, Nalam S, Calhoun BH. Analyzing static and dynamic write margin for nanometer SRAMs. In: Proc int symp on Low power electronics and design, ACM, New York (NY, USA); 2008. p. 129–34.