

Stability Optimization of Embedded 8T SRAMs using Word-Line Voltage Modulation

B. Alorda, G. Torrens, S. Bota, J. Segura

Electronic Systems Group, Physics Dept.

Illes Balears University

Palma de Mallorca, Spain

tomeu.alorda@uib.es

Abstract

SRAM cell stability analysis is typically based on Static Noise Margin (SNM) evaluation when in hold mode, although memory errors may also occur during read operations. Given that SNM varies with each cell operation, a thorough analysis of SNM in read mode is required. In this paper we investigate the SNM of OAM cells during write operations. The Word-Line Voltage modulation is proposed as an alternative to improve cell stability when in this mode. We show that it is possible to improve 8T OAM cells stability during write operations while reducing current leakage, as opposed to present methods that improve cell stability at the cost of leakage increase.

1. Introduction

Current System on Chip (SoC) trends result in a significant percentage of the total die area being dedicated to memory blocks, thus making embedded SRAM yield dominate the overall SoC yield. The CMOS IC technologies have been constantly scaled down aggressively, and specially embedded SRAM cell layout has been significantly impacted. Some of these impacts are related to the critical dimension (CD) reduction in poly and diffusion features that entails an increase in physical parameters variation, which among other effects has a direct impact on SRAM cell stability. Therefore, a deep knowledge and analysis about the stability of the embedded SRAM cells and new approaches to improve the cell stability in post-production or during the circuit life-time is becoming a must in modern embedded SRAM CMOS designs.

The stability and robustness of a given SRAM 6T-based cell is usually evaluated by analyzing the Static Noise Margin during hold, read and write operations. The SNM during hold is defined as the minimum DC noise voltage needed to flip the cell state [1], and is used to quantify the stability of a SRAM cell using a static approach. During read operations a similar definition is used to quantify the static stability of the cell when the access transistors connect the internal cell nodes to the bit-lines. In that case, this static parameter is called Read-SNM or RSNM. The static stability during write operations cannot be defined in the same way because the objective of a write operation is to force a new value in the cell therefore breaking the cell stability. Several metrics have been defined to measure the write noise margin and in this paper the Write-SNM (WSNM) defined in [2] has been considered. Therefore, the WSNM is defined as the smaller of the two squares that can fit between the cell static characteristics during a write operation. RSNM has been identified as a critical memory

design parameter and great efforts have been done to increase stability during read operations. Some approaches improve this critical parameter by modifying the read circuitry [3], or by reducing the time required to sense the cell stored values [4]. In [5] the limits of different assist methods is discussed for 6T SRAM cells. Current approaches explore new cell architectures including those that increase the SRAM cell transistor count. Among several proposals, eight-transistor (8T) cell configurations are being adopted as an alternative to the traditional six-transistor (6T) cells in industry designs [6]. Most of the new 8T SRAM approaches are based on the traditional cross-coupled inverters arrangement as the core storage structure like in 6T designs. The 8T SRAM cell selected in this work incorporates, compared to the 6T cell, a dedicated read port circuit to improve read robustness at a given supply voltage for similar area constraint [7].

In previous works a detailed analysis about 6T SRAM cells static stability during read, and the comparison between different alternatives to improve the SNM during hold and read operations were discussed [8]. In that work, the major improvement in RSNM, when no changes are introduced in memory structure design, came from Word-Line maximum Voltage modulation. In [9], a dual word-line voltage strategy was proposed and analyzed to improve the RSNM without affect the write operation margins. Furthermore, the proposed strategy was used in combination with transistor width modulation to improve both static and dynamic stability.

In this paper, we present an analysis focused on the 8T-based memory cell and explore the benefits of decreasing the word-line voltage showing the drawbacks in terms of delay and power consumption. The rest of the paper is organized as follows: the following section presents the column architecture and the timing behavior of control signals. The weakest state of an 8T-based cell is identified and described. Section 3 proposes the word-line voltage modulation scheme to improve the cell stability. Impacts on the full memory characteristics are explored in terms of delay and consumption. Section 4 shows the impact of parameter variation on the proposed solution. Finally, Section 5 points out the main conclusions.

2. Architecture assumptions

The 8T SRAM memory cell is composed of two cross-coupled CMOS inverters with two pass transistors connected to complementary bit-lines and an additional read port for readout operations. Fig. 1 shows the selected 8T cell structure inserted in a 256-cell column architecture.

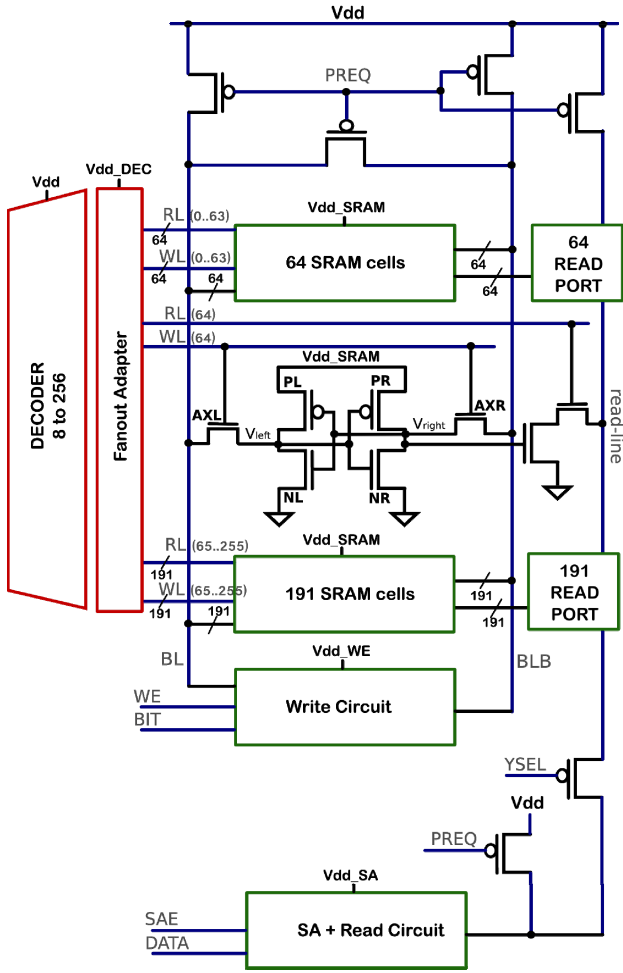


Figure 1. 6T SRAM column schematic.

The access transistors, AXR and AXL, have their gate node connected to the word-line (WL) for access during write operations through the column bit-lines (BL and BLB). The read port is connected to the read-line (RL) to read the internal value through the read line. Therefore, bit-lines act as input nodes carrying the data from write circuitry to the memory cells during write operations, and read-line act as output node carrying the data from SRAM cells to the sense amplifier during read operations.

The considered column consists of 256 8T SRAM cells implemented in commercial CMOS 65nm, as well as the write circuitry, the sense-amplifier to perform read operations and other circuitry like pre-charge modules and column multiplexing blocks used to share various columns with one sense-amplifier. The row decoder has been represented in Fig. 1 with two blocks: the logic block implementing the decoder function, and the buffer stage that performs the adaptation to word-line capacitance. This separation has been shown since the word-line voltage modulation approach proposed in this paper only involves the buffer stage circuit without changing the logic decoder block. Different power supply voltage nodes have been defined in Fig.1 to allow energy consumption computation associated to each subcircuit of the memory column.

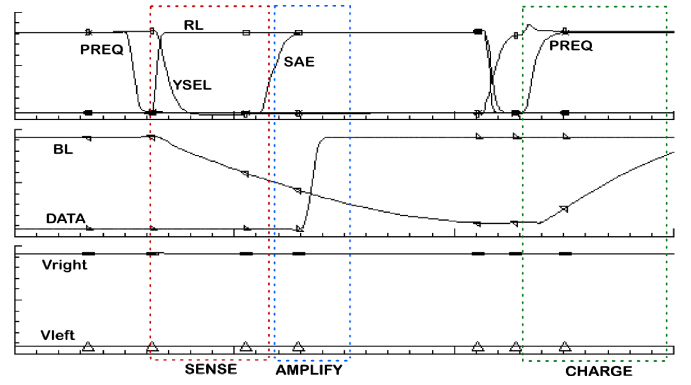


Figure 2. Read operation Timing.

Signal timing during a read operation is shown in Fig. 2. In this case, the 8T cell is accessed using the read port and the RL signal controls the evaluation process of internal value stored. A read operation cycle is composed of three sub-states: (i) SENSE: It starts when the RL signal activates the read port. In this sub-state, the read-line is discharged according to the internal cell value. The corresponding column pre-charge transistors are turned OFF while mux-column transistor is turned ON. (ii) AMPLIFY: The read-line value is amplified to a logic value and stored into the output latch. This sub-state begins when the SAE signal enables the read circuitry. (iii) CHARGE: It starts once after RL is turned OFF. This eliminates current contention that may cause unnecessary power consumption and leaves the bit-lines ready for the next operation cycle.

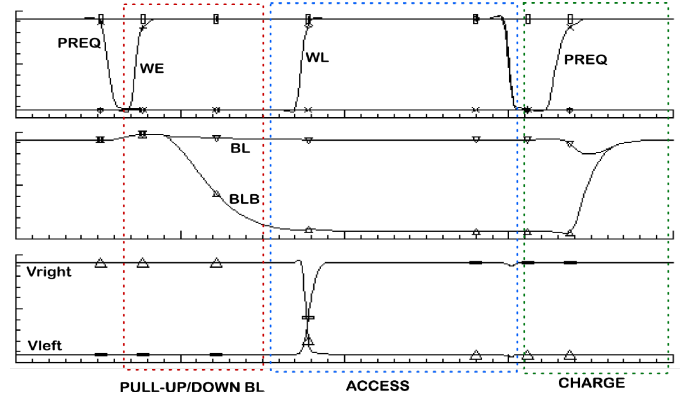


Figure 3. Write operation Timing.

Write operations timing is shown in Fig. 3. In this case, the write circuit controlled by WE signal pulls-up and down the bit-lines to induce the write operation according to the desired value indicated by the BIT signal. The write operation in a 8T-based SRAM is similar to the write operation in the traditional 6T-based SRAM.

A write operation cycle is composed of three sub-states: (i) PULL-UP/DOWN BL: It starts when the write circuit is activated (WE signal is ON) and the bit-lines are pulled according to the BIT signal. In this sub-state, the pass transistors of the cell to be written are OFF, as well as the corresponding column pre-charge transistors. (ii) ACCESS: The cell access transistors are turned ON and the new data

value is stored into the memory cell. (iii) CHARGE: It starts after the WL is turned OFF and leaves the bit-lines ready for the next operation cycle.

In a typical SRAM memory organization, only one word-line signal driven by the row decoder is set high during memory access (read or write) according to the input address, while the column decoder selects which specific cell in each block is actually read-out or write. With this access mechanism not only the cells with the columns selected are connected to their corresponding bit lines, but also the cells sharing the same word-line. These cells are commonly called half-selected cells or cells being in Open Access Mode (OAM). Any cell being in OAM has its internal nodes connected to the corresponding bit-lines, and it is in its worst-case cell stability mode. Read operations in 8T-cell memories use the read port thus keeping the cell internal nodes isolated from the external bit-lines. However, 8T write operations keep all half-selected cells with the same word-line in OAM.

The timing of an OAM cell during a write operation is shown in Fig. 4. The WL signal controls the memory cell pass transistors while the column decoder does not connect the bit-lines to the write circuitry, therefore the pre-charged bit-lines forces the internal cell nodes to reduce the noise immunity and a discharge process of bit-lines is produced.

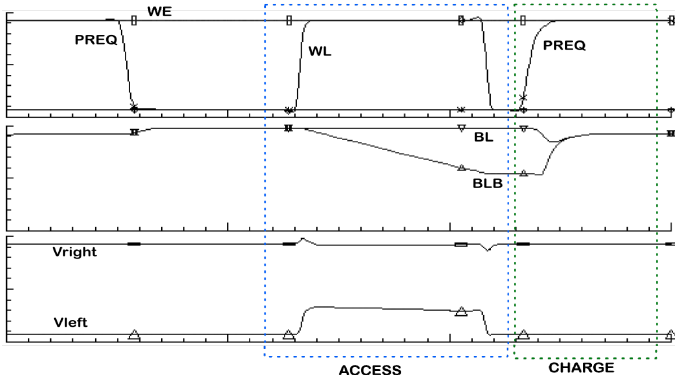


Figure 4. Write operation Timing for OAM cells.

Only two write sub-states out of the three can be identified during the write operation timing for OAM cells. In this case, during the ACCESS sub-state the bit-lines initially pre-charged senses the internal value stored in the accessed SRAM cell. During this sub-state the selected cell is written to the final value, but for half-selected cells the situation is similar to a 6T readout process.

Some research has been done to avoid OAM cells in 8T memories to maintain high robustness in all memory modes. These solutions propose eliminating half-selected cells during write operations by designing a memory array where all 8T cells sharing the same word-line are selected for the write process. Therefore, when a memory cell is selected through word-line activation, a new written data will be stored [10]. The main drawback of these designs is that the benefits of block-based memory architectures (i.e. high interleaving distance) are eliminated and the cost in terms of multi-event

radiation stability may be high. This work proposes a new approach to minimize the impact of OAM in 8T-based SRAMs and allows using the interleaving distance technique to increase the robustness of memories to multi-bit soft-errors.

3. Word-Line Voltage Reduction

Given that the SNM is significantly degraded for OAM cells during write operations, in this work we evaluate the benefits of word-line voltage (V_{WL}) reduction to increase the SNM parameter for these cells compatible with memory structures architectures that implement interleaving distances. In previous works [8,9] a dual word-line voltage (V_{WL}) methodology was proposed to reduce the voltage of word-line during 6T read operations while remaining at their nominal value during write operations. A more careful analysis, when the memory is in write state, reveals that apart from the cells that are being written, there are other cells that operate in OAM. These cells are those that share the same word-line with the cells being written, so their access transistors are ON and the respective bit-lines are connected to the stored value. In the following subsections we will explore the benefits and drawbacks of word-line voltage reduction in 8T-based memories with half-selected cells.

3.1. Benefits on OAM cells.

The main benefit of V_{WL} modulation is based on the SNM behavior observed when the V_{WL} is reduced from the nominal memory bias voltage (VDD). Fig. 5 shows the evolution of SNM and how reducing the V_{WL} by 200mV from VDD (17% reduction), SNM improves by about 50%.

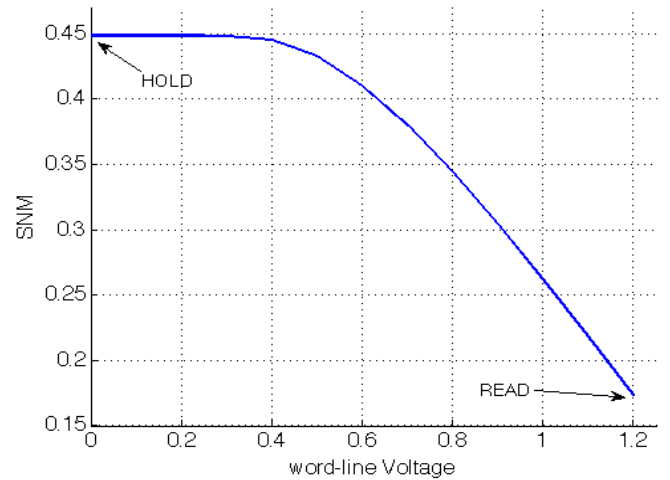


Figure 5. Impact of V_{WL} on SNM.

The reduction of V_{WL} during a write operation on a 8T-based cell does not only affect the cell stability, but also increases the write latency. When V_{WL} is lowered the internal cell node being high gets a smaller voltage. In the following discussion we will assume that in Fig. 1 $V_{left}=0$, $V_{right}=1$.

When the 8T cell is in OAM during a write operation, the internal node being initially low (V_{left}), increases its voltage as a consequence of charge redistribution coming from the pre-

charged bit-lines (see V_{left} node behavior during ACCESS sub-state in Fig. 4). Fig. 6 shows how the maximum voltage reached by V_{left} when access transistors are opened is reduced when the V_{WL} is lowered. In that case, a 17% of reduction in V_{WL} produces a 33% of reduction in maximum V_{left} voltage during write operations. This behavior is directly related to cell stability improvement and affect all half-selected cells during write operations.

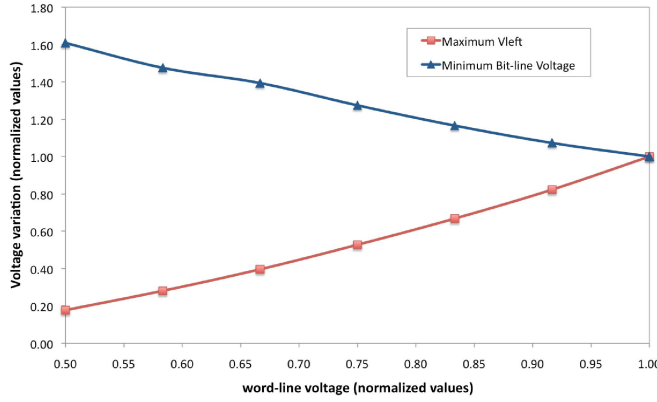


Figure 6. Variation of maximum degradation voltage of internal cell nodes and final BL discharge voltage due to V_{WL} reduction.

A smaller V_{left} degradation implies a higher value of the final minimum related bit-line voltage. This situation is produced during SENSE period (Fig. 4) and is shown in Fig. 6, where a V_{WL} reduction of 17% implies a 17% less bit-line discharged voltage. The smaller discharge experienced by the bit-lines, result in less power consumption during the next pre-charge process. If this impact is applied to all SRAM columns in OAM during write operations, the impact on power saving may be significant for low-power applications.

To calculate the energy saved per write cycle, some assumptions have been made: (i) energy is computed considering only one column, following the structure shown in Fig. 1; (ii) a write cycle is formed by the three steps described in Fig. 3; and (iii) in write operations only a few SRAM columns are selected for the operation, while the rest of memory columns are in half-selected mode. Therefore, for a 256KB embedded memory organized in 16 blocks of 64 columns of 256 cells per column, considering that the write operation affect to 16 bits simultaneously, only 16 cells will be selected for write process, while $(16 \times 64) - 16$ cells will be in OAM.

Fig. 7 shows the impact of V_{WL} reduction on energy consumption for the case of selected and OAM cells. In the case of write-selected cells, the energy saving during a write cycle is negligible, because the behavior of this column is not modified so much. The V_{WL} reduction enlarges the write process at the cell level, but the rest of circuits (bit-lines discharge, decoders) remain unchanged. Fig. 7 represents the column total energy consumption per cycle, but not all circuits contribute in the same way to the overall energy as shown in Table 1.

And in the other case, the OAM cells during a write process try to discharge the bit-lines following the internal values stored, so, due to V_{WL} reduction, the bit-lines are less discharged and the stability of memory cell is increased, see Fig. 6. This feature is interesting from the point of view of energy saving, because less bit-lines discharge means less energy to pre-charge again those bit-lines. It may be observed in Fig. 7 where if a V_{WL} reduction of 17% is applied, the energy saved per half-selected column is about 5% per OAM column in each write operation.

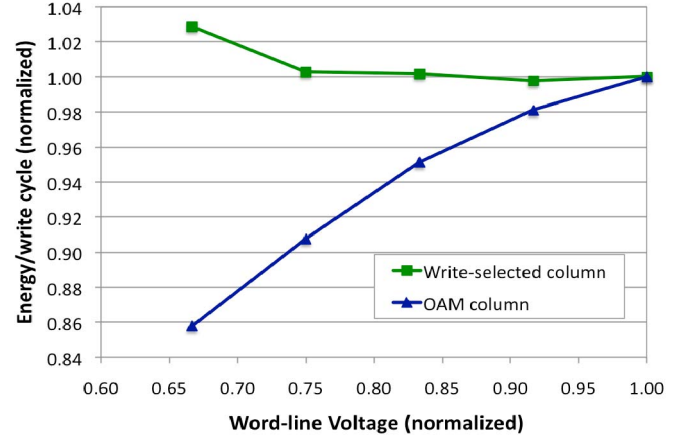


Figure 7. Impact of V_{WL} reduction on total energy/write cycle consumption of write-selected and OAM columns.

This benefit in energy savings per OAM column must be considered taking into account that the number of OAM columns in one memory architecture during a write operation will be higher than the number of write-selected columns.

To evaluate the different sinks of energy consumption, four different power supply nodes were used to quantify each block contribution on the total energy/cycle during write operation. The blocks considered are shown in Fig. 1 and the energy of each block is defined as:

- $E(V_{\text{dd}})$ energy consumption of pre-charge circuitry, control logic, and decoder logic.
- $E(V_{\text{dd_SRAM}})$ energy consumption of 8T memory cells.
- $E(V_{\text{dd_WE}})$ energy consumption of write circuitry.
- $E(V_{\text{dd_DEC}})$ energy consumption of the buffer stage used to drive the word-line voltages from decoder logic.

Table 1 shows the energy consumption of each block during a write operation for the columns having a cell selected for a write access. Table 2 shows the energy consumption for the columns having OAM cells during that period. Each table column shows how the consumption of each block is modified by V_{WL} reduction. For example, the column $E(V_{\text{dd_SRAM}})$ of Table 1 shows how the energy consumption of the write-selected column raises due to the latency increase in the cell write process. Therefore, the same column in Table 2 represents the reduction in consumption of OAM column due

to the smaller effort required to maintain the stored value in the memory cell.

Table 1. Energy/write cycle evolution for write selected columns.

V_{WL}	E(Vdd)	E(Vdd_WE)	E(Vdd_SRAM)	E(Vdd_DEC)
100%	1	1	1	1
92%	1	1	1.12	0.81
83%	1	1.03	1.27	0.65
75%	1	1.04	1.58	0.52
67%	1	1.08	3.26	0.42

Table 2. Energy/write cycle evolution for OAM columns.

V_{WL}	E(Vdd)	E(Vdd_WE)	E(Vdd_SRAM)	E(Vdd_DEC)
100%	1	1	1	1
92%	0.99	1	0.88	0.81
83%	0.96	1.03	0.85	0.65
75%	0.92	1.04	0.81	0.52
67%	0.87	1.08	0.8	0.42

Although from previous tables it seems that all blocks contribute in the same way to the total energy consumption, the contribution of E(Vdd) represents the 93% of the total energy consumption, the contribution of E(Vdd_DEC) is 2% of the total, 2% is related to E(Vdd_SRAM) and the remaining 3% comes from E(Vdd_WE). In the case of a half-selected column, the distribution is 97% from E(Vdd), 1% from E(Vdd_SRAM) and 2% from E(Vdd_DEC). From these distributions, it is clear that the main energy saving comes from the reduction in the bit-line minimum voltage of OAM cells during write operations.

3.2. Impact on Write Operation

Decreasing the V_{WL} during write operations may result in SRAM cells being less capable to retain written data. Fig. 8 shows the variation of WSNM depending of V_{WL} . According to the slope, two regions are observed. To ensure a lower impact on write margin, it is advisable to maintain the V_{WL} in the lower slope region. This means that the V_{WL} should not drop below 20% of V_{DD} . For example, with a V_{WL} reduction of 17%, the WSNM is reduced only by 7.8%.

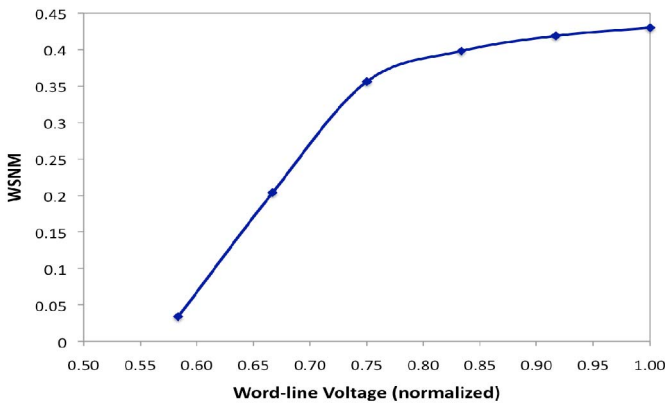


Figure 8. Impact of WL on WSNM.

The write latency was considered as the time between the 50% of rising edge of WE signal and the 50% of maximum value of last edge transition of V_{left} and V_{right} (Fig. 3).

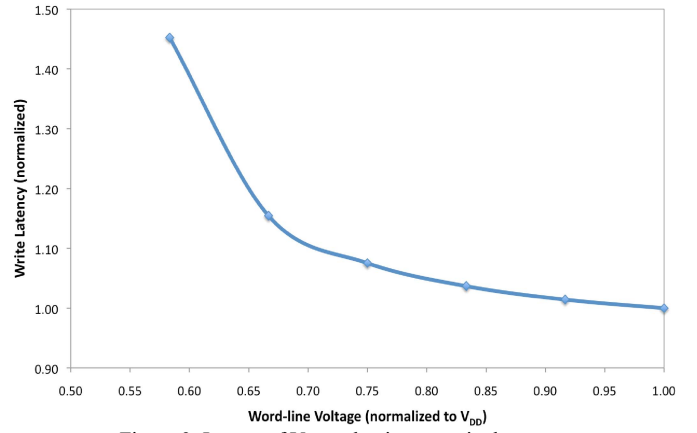


Figure 9. Impact of V_{WL} reduction on write latency.

Similarly to WSNM, two trends are observed: the write latency increases fastly when the V_{WL} reduction is beyond 20% of V_{DD} . Therefore, if the V_{WL} is reduced by 17%, the variation in write latency will remain low, just 4% higher than at maximum V_{WL} .

4. Impact of parameter variation

The impact of process and transistor mismatch has been evaluated using the write latency as key parameter. As shown in Fig. 9, write latency has a strong dependence on V_{WL} . Figure 10 shows Monte Carlo simulation results where the write latency mean is represented in solid line and is obtained as a mean value over 500 iterations. The dashed lines represent the standard deviation obtained from the corresponding histogram. The upper point in the write latency mean plus the standard deviation, while the lower point is the write latency mean minus the standard deviation. Both values are represented normalized to the minimum write latency histogram parameters.

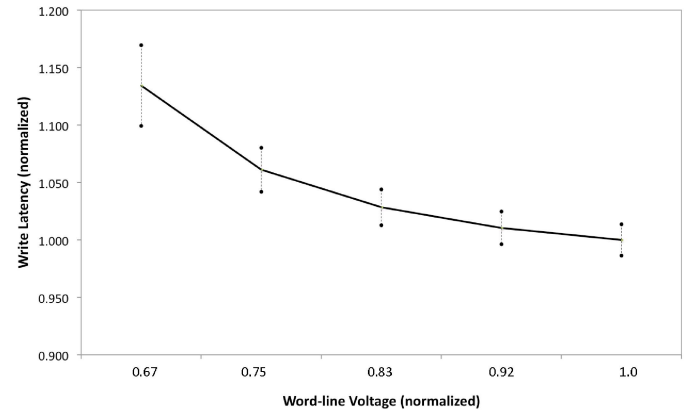


Figure 10. Impact of parameter variation on Write delay.

Not all iterations have produced a successful write operation. For V_{WL} being 75% of V_{DD} , at least 0.2% of the cases lead to a wrong result. This trend increases quickly as V_{WL} was further reduced: when V_{WL} was 67% of V_{DD} the write process was wrong in the 29% of cases. Therefore, considering a V_{WL} reduction of 17%, the write latency mean is 3% higher than the nominal value with an increment of 1.4% in the standard deviation. Taking into account the impact of variability in

nanometer technologies, an adaptive solution may be implemented to adjust the level of V_{WL} . Applying a method similar to the one proposed in [11], the specific V_{WL} value for each circuit may be obtained from an iterative methodology by using a set of cells as being representative of the whole memory. This subgroup of cells is written in a normal way, using an initial value of V_{WL} , and then it is read to estimate the writability conditions. If any cells of the specific subgroup fail the operation, then V_{WL} is slightly increased and the procedure is repeated. In addition, if a local charge-pump is used, this procedure may be used for each memory in the chip. The result of this adaptive methodology is that each memory circuit will be as much robust as possible.

5. Conclusions

The V_{WL} reduction has been proposed as a way to increase the robustness of OAM 8T-based cells during a write operation. This requires an increase of write latency, and a decrease in WSNM. To avoid an excessive WSNM degradation, the approach proposed in this work consists in a V_{WL} reduction based on an iterative algorithm to determine the nominal voltage during all kind of SRAM access operations in order to increase SNM. The final V_{WL} value used could be selected according to post-process calibration and adjusted for each memory block in the SoC.

An additional benefit of this proposal is the increase of the final bit-line discharge voltage value in OAM columns, therefore reducing the energy consumption in the pre-charge period. These improvements can be implemented without changing the SRAM architecture; only the buffer stage in the row decoder must be redesigned to work at a lower bias supply.

The use of this technique in embedded SRAMs designs could contribute to achieve most stable operations, best soft-errors robustness and reduce power consumption.

ACKNOWLEDGMENTS

This work has been supported by the Cluster for Application and Technology Research in Europe on NanoElectronics under grand CA303 OPTIMISE, the Spanish Ministry of Science and Innovation under grant No. AP2006-03170 and the project TEC2008-04501/MIC.

References

- [1] Seevinck, E., et al. 1987, "Static-Noise Margin Analysis of MOS SRAM Cells". IEEE Journal of Solid-State Circuits, SC-22, 5. 748-754.
- [2] J. Wang, et al. 2008, "Analysing Static and Dynamic Write Margin for nanometers SRAMs", Proc. Of 13th Inter. Sympo. on Low Power Electronics and Design, 129-134.
- [3] H. Pilo, et al., "An SRAM Design in 65-nm Technology Node Featuring Read and Write-Assist Circuits to Expand Operating Voltage", IEEE Journal of Solid-State Circuits, vol-42, 2007, pp. 813-819.
- [4] K. Zhang, et al., "SRAM Design on 65-nm CMOS Technology With Dynamic Sleep Transistor for Leakage Reduction", IEEE J. Sol. St. Circuits, 2005, pp 895-901.
- [5] R.W. Mann, et al. "Limits of Bias Based Assist Methods in Nano-Scale 6T SRAM", Proceedings of Quality Electronic Design Symposium, pp. 1 – 8, 2010.
- [6] H. Akamatsu, et al. "A 45nm 2-port 8T-SRAM Using Hierarchical Replica Bitline Technique With Immunity From Simultaneous R/W Access Issues", IEEE Journal of Solid-State Circuits, Vol. 43, 4, 2008
- [7] T. Kim, et al., "A Voltage Scalable 0.26V, 64kb 8T SRAM with Vmin lowering techniques and Deep Sleep Mode", IEEE Journal of Solid-State Circuits, vol. 44, n° 6, 2009, pp. 1785-1795.
- [8] G. Torrens, et al., "Design Hardening of Nanometer SRAMs Through Transistor Width Modulation and Multi-Vt Combination", IEEE Transactions of Circuits and Systems II: Express Briefs, pp. 280-284, 2010.
- [9] B. Alorda, et al., "Static and Dynamic Stability Improvement Strategies for 6T CMOS Low-power SRAMs", Proceedings of Design Automation and Test in Europe, 2010.
- [10] L. Chang, et al., "An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches", IEEE Journal of Solid-State Circuits, vol. 43, no. 4, April 2008.
- [11] A. Raychowdhury, et al. "PVT-and-aging adaptive wordline boosting for 8T SRAM power reduction" IEEE International Solid-State Circuits Conference, 352-354. 2010.