RAPIDO Testing of Assisted Write and Read operations for SRAMs

Joseph Nguyen[†‡], D. Turgis[†], D. Bonciani[†], B. Lhomme[†], Y. Carminati[†], O. Callen[†], G. Guirleo[†], Lorenzo Ciampolini[†], G. Ghibaudo[‡]
[†]STMicroelectronics, 850 rue Jean Monnet, 38920 Crolles
[‡]IMEP-LAHC,Minatec/INPG, Grenoble, France

***Abstract*– Lowering the supply voltage of Static Random-Access Memories (SRAM) is key to reduce power consumption, however since this badly affects the circuit performances, it might lead to various forms of loss of functionality. In this work, we present silicon results showing significant yield improvement, achieved with write and read assist techniques on a 6T high-density bitcell manufactured in 40 nm technology. Data is successfully modeled with an original spice-based method that allows reproducing at high computing efficiency the effects of static negative bitline write assist, the effects of static wordline underdrive read assist, while the effects of read ability losses due to low-voltage operations on the yield are not taken into account in the model.**

Keywords- SRAM, write assist, write margin, read assist, half-selected cell, yield, negative bit line, wordline underdrive.

Introduction

Low-Power (LP) applications are gaining interest, since they allow wearable or portable devices to execute, at a given battery capacity, more complex tasks than applications that were operated at standard power levels. For devices that prioritize autonomy over performance, scaling the suppply voltage is one of the best way to reduce power consumption. Since using low supply voltage may cause failures in some circuits, several ways exist to tackle this issue, for example using a separate, higher-voltage supply and connect it to the faulty circuits [1], at the cost of an additional voltage source. For Static-Random-Access-Memory (SRAM), it can be easier to use assist techniques when required. 6T SRAM bitcells performance and stability depend on the strength ratio of the transistors composing the bitcell. Process variations may cause read and write failures, and the lower the supply voltage, the greater effects these variations have. If pass-gates are weak due to the manufacturing variability, the bitline voltage levels during a write operation may not be transmitted correctly to the internal nodes (blti, blfi) and the trip point at which blti=blfi might never be reached, leading to a write failure, see Figure 1. For such write-limited bitcell, applying write assist allows the SRAM to function at lower voltage until read operation starts to fail because of half-selected read disturb [2].

To improve write margin, negative bitline (NBL)[3–6] can be used by applying a negative voltage NBL on the bitline carrying a logic zero, as shown in Figure 2. To improve read margin, Wordline Underdrive (WLUD) can be used, where the wordline voltage is lower than the voltage of the array [7], to weaken the pass-gates during read operation.

In this work, NBL and WLUD are used to assist the SRAM 6T bitcell to lower the minimum operating voltage V_{MIN} on a test chip that is described in the next section. The extra voltage levels are generated using external voltage supplies. Their effects are shown on silicon, and a model to estimate the write margin taking into account the NBL is proposed.

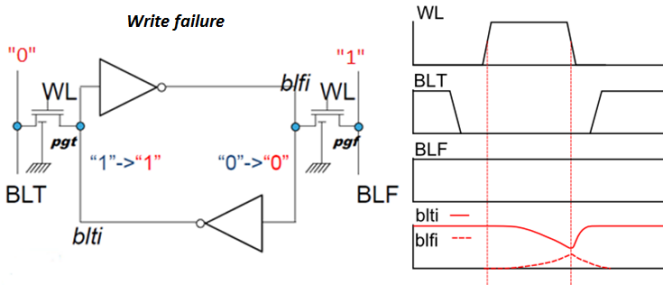

Fig. 1. Schematic showing a write failure and the correspondent waveforms showing no trip point.

RAPIDO test vehicle and experimental setup

Figure 3 shows the floorplan of the test chip manufactured by STMicroelectronics in 40nm technology node, showing in green different test vehicles for various capacities of different bitcells. The test vehicles are called RApid Memory PrototypIng for Process Defectivity Optimization (RAPIDO). The rest of the chip is used to access the memory cuts, or to test other IPs.

The main advantage of RAPIDO is to get very rapidly silicon-based results on the functionality and manufacturability of new memory cells, as early as the first processed test-chip of a technology is available. This is possible beacause RAPIDO allows to generate very robust memory cuts, so that process yield can be directly correlated to the bit-cell (BC) yield, suppressing problems linked to design styles. The RAPIDO periphery operates at a fixed nominal voltage and at low-frequency (MHz range) to minimize marginalities at nominal voltage. The current design allows to vary the array supply voltage V_{ddma} and allows emulating statically the NBL write assist and WLUD read assist through two

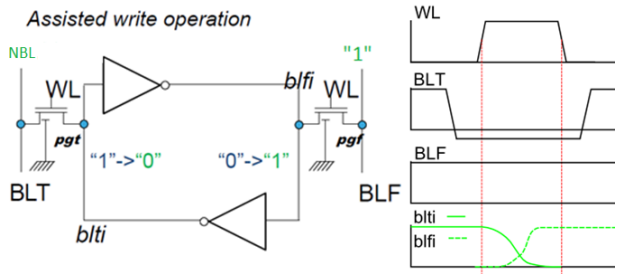

Fig. 2. Schematic showing the NBL (Negative bitline) write assist and its waveforms showing a successful write.

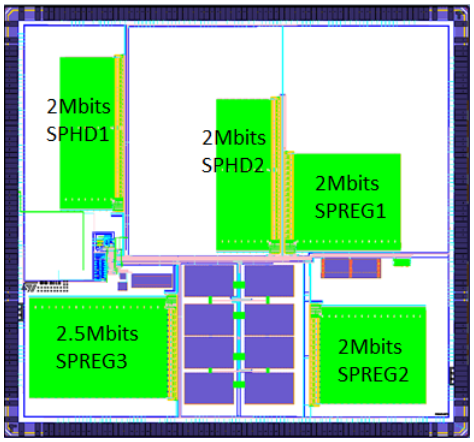

Fig. 3. Floorplan of the test chip with 5 RAPIDOs.

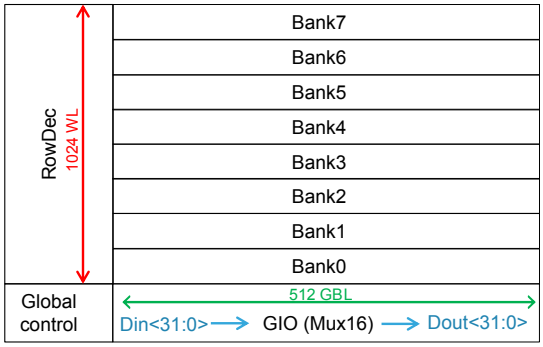

Fig. 4. Architecture of the RAPIDO SPHD2 SRAM.

independent power supplies for bit lines and word line drivers, respectively.

In this paper, the SPHD2 that is investigated, is a 2Mb cut of a 0.242 μm^2 high-density bitcell [8] containing 65536 words of 32 bits, at a multiplex level of 64. Figure 4 shows how the memory is divided in eight memory banks, with 1024 WLs connected to the banks, and 512 global bitlines. The timing is handled in the global control circuit and then distributed locally in each bank, see Figure 5. The same figure shows that each bank is separated in two arrays, made of 128 rows and 1024 columns. Local control block circuits manage timings and local I/O circuits to exchange bit lines at mux4 with the global I/O, see Figure 5. Data read-out is sped up by the use of a differential Sense Amplifier (SA), triggered by a dedicated self-timing circuit.

NBL write assist is tested as follows: each testable memory out of a population of 67 dies is written at a given ground bitline voltage, then the memory is read at nominal voltage. WLUD is applied during all write operations to prevent possible half-selected cell disturbs. If the read code corresponds to the written code, a PASS indicator is returned in a datalog file, otherwise its a FAIL. Yield is then calculated by the ratio of the number of PASS conditions with the number of testable memory cuts. The whole is iterated for different ground bit line voltages.

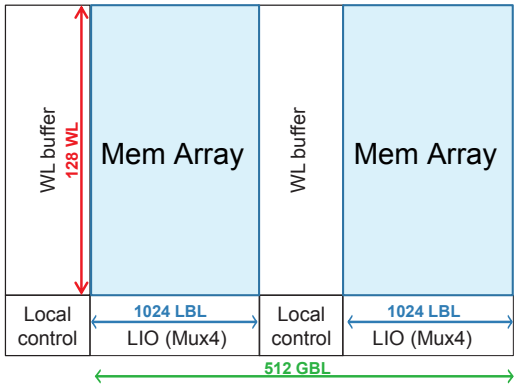

Fig. 5. Structure of one bank of the RAPIDO SPHD2 memory.

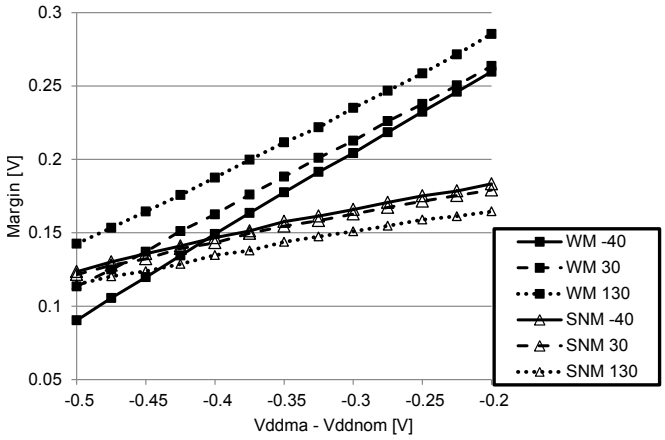

Fig. 6. Simulations of WM and SNM vs ($V_{ddma} - V_{ddnom}$), at TT -40 °C, 30 °C and 130 °C.

Modeling of bitcells marginalities

3.1 *Yield at no assist*

The variability of the SPHD bitcell used in the RAPIDO test vehicles has been modeled with spice-based simulations using the model card available at STMicroelectronics for this technology node. Monte Carlo (MC) technique is used to simulate the fluctuations due to random effects around typical conditions (TT) of Write Margin (WM, [9]) and Static Noise Margin (SNM, [10]). With respect to the die population, the models are slightly pessimistic for PMOS performances, and slightly optimistic for NMOS performances. Figure 6 shows the average WM and SNM results for typical process conditions vs ($V_{ddma} - V_{ddnom}$) at three temperatures: -40 °C, 30 °C, 130 °C. Margins are expressed in volts, against the difference between the operating supply voltage V_{ddma} and the nominal supply voltage V_{ddnom} 1.1V [8]. At these conditions, the bitcell is clearly read-limited, and write assist would be of no help, but read assist can be used to modify this situation.

Bitcell write failure probability is calculated from the mean and sigma μ_{WM} and σ_{WM} assuming a simple normal distribution [11]. Failure probability due to read disturb is calculated in the same manner from μ_{SNM} and σ_{SNM}. Yield

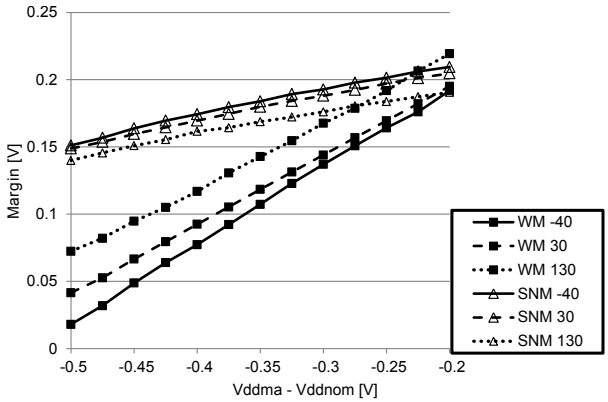

Fig. 7. Simulations of WM and SNM vs ($V_{ddma} - V_{ddnom}$) with 60mV WLUD, at TT -40 °C, 30 °C and 130 °C.

is basically obtained from the SNM and WM mean-over-sigma ratios μ/σ [12].

3.2 Read assist modeling

Static WLUD can be modeled by lowering the voltage on the passgate gates in static margin simulations and calculating yield as described in the previous section. Figure 7 shows results obtained in the same condition than Figure 6, but with read-assist applied through a constant 60mV drop on the WL supply. SNM is now larger than WM on the whole voltage range, except at 130 °C and V_{ddma} close to V_{ddnom}. In these conditions, the bitcell is more write-limited and it is expected to observe effects of NBL.

3.3 Write assist modeling

In NBL write assist, a negative voltage is applied column-wise only on the bitlines carrying the '0' logical state, with only one row selected through the WL, therefore it has no effect on SNM. It is more subtle to model the effect on WM. Figure 8 shows how the bitcell write failure probability at TT $V_{ddnom} - 0.45V$, 30 °C evolves against WL pulse width. This low supply operating voltage was chosen to increase the number of failures and to reduce the number of MC runs to obtain reliable results. The precision of dynamic results depends critically on the number of MC runs: if the dynamic probability decreases by one order of magnitude, the minimum number of MC runs to feature at least some fails increases by one order of magnitude and might soon become practically unfeasible.

For no NBL assist (squares), the shorter the WL pulse width is, the higher the probablity of failure. As the WL pulse width extends, failure probablity lowers and saturates to the probability obtained by using WM as described above (solid line), obtained with a single, small set of 4k MC runs. When NBL assist is applied, failure probablity decreases but follows the same trend, as shown by results obtained from different dynamic write MC simulations (symbols).

We can perform a static estimation of static NBL by adding the negative bitline voltage directly to the WM:

$$\mu'_{WM} = \mu_{WM} - NBL \quad (1)$$

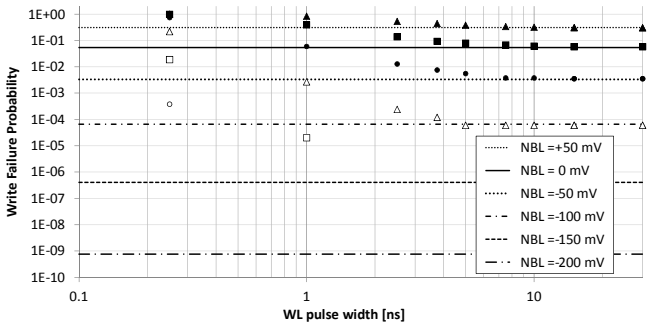

Fig. 8. Effects of write assist on write failure probability vs WL pulse width.

Bitcell failure probability with static write assist can then be modeled as in the previous section, and its results are shown in Figure 8 (dotted lines). For long-duration pulses, the proposed model fits well the results obtained through dynamic simulations. The figure also shows that the effect of NBL is very strong, decreasing by nearly 8 orders of magnitude the write failure probability for NBL=-200 mV. As the RAPIDO is used in the MHz range, the WL pulse width is in the range of microseconds and therefore, the proposed method can be used to estimate WM with NBL assist.

Experimental yield measurements vs models

All measurements have been carried out at three temperature values, -40 °C, 30 °C and 130 °C, as shown in Figure 9, Figure 10 and Figure 11. The array supply voltage V_{ddma} varies between V_{ddnom}-0.2V and V_{ddnom}-0.35V. A WLUD of 60mV is used across all measurements, while NBL ranges from NBL=0V to NBL=-0.3V. For all temperature, yield is zero or very poor (20 % at 130 °C and $V_{ddma} = V_{ddnom} - 0.2V$) when NBL is not applied. When NBL=-0.1V is applied, one fully recovers yield at $V_{ddma} = V_{ddnom} - 0.2V$ for all temperatures, meaning that it is possible to extend the operating voltage range safely down to this low supply level.

For lower supply levels, one observes that NBL=-0.2V can be sufficient to recover yield. Nevertheless, at the lowest temperature, yield is never recovered fully, but rather saturates well below 100%. Basically one sees that, with the current RAPIDO design, it is useless to apply large NBL values, since at low-voltage and low temperature yield is badly affected by a different phenomenon. To explain this phenomenon, we have calculated the worst-case current level amongst 512 bitcells (this amount corresponds to the number of bitcells which are connected to each SA, when one takes into account the multiplex level) for all temperatures and supply levels. Results are plotted in Figure 12 against the observed yield level for large NBL: one observes that the latter is well below 100 only for very weak currents. The yield might therefore limited by the very weak performances of the worst-case bitcells, which might fail due to read-ability, a failure in which the bitcells current is not sufficient to discharge the bit line capacitance within the due time.

At larger temperatures, the small misalignment between the particular silicon on which dies were manufactured and typical models might introduce some uncertainty in the comparison. Results at NBL=-0.1V, for example, show some

minor discrepancies with the model, which might be due to the fast (with respect to the model) PMOS, which might explain why silicon writing is slightly more difficult than expected by the model.

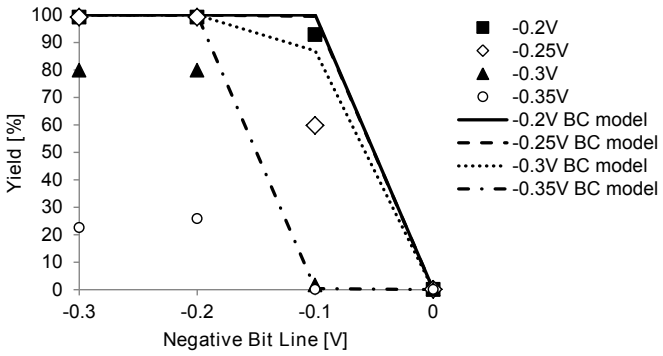

Fig. 9. Experimental (symbols) and modeled (lines) yield vs NBL, at -40 ° C.

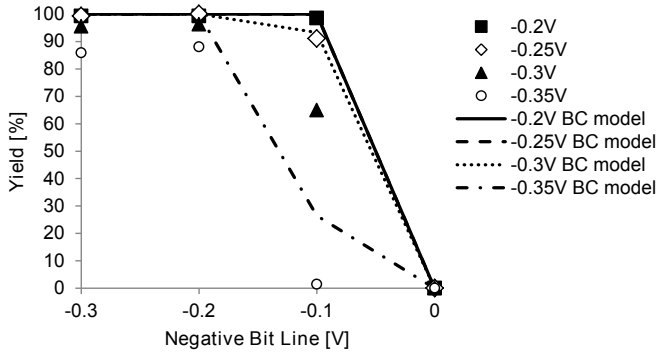

Fig. 10. Experimental (symbols) and modeled (lines) yield vs NBL, at 30 ° C.

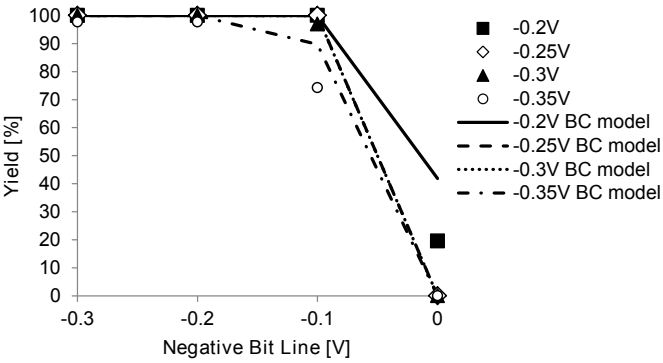

Fig. 11. Experimental (symbols) and modeled (lines) yield vs NBL, at 130 ° C.

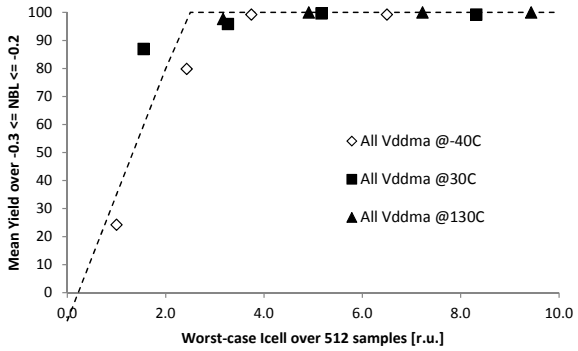

Fig. 12. Mean experimental yield with NBL=-0.2V and NBL=-0.3V for all temperatures and supply levels vs worst-case read current amongst 512 bitcells.

Conclusion

In this work, assist techniques were used to lower the minimum operating voltage V_{MIN}, when NBL=-0.1V is applied, the memory supply can be lowered to $V_{ddnom}-200$mV safely. Yield is successfully modeled with an original spice-based method that allows reproducing at high computing efficiency the effects of static negative bit line write assist and the effects of static word line underdrive read assist, while the effects of read-ability losses due to low-voltage operations on the yield are not taken into account.

References

[1] K. Zhang et al. A 3 ghz 70 mb sram in 65 nm cmos technology with integrated column-based dynamic power supply. *JSSC*, 41(1), 2006.

[2] T. Suzuki et al. A stable 2-port sram cell design against simultaneously read/write-disturbed accesses. *JSSC*, 43(9) :2109–2119, 2008.

[3] N. Shibata et al. Bitline-overdriven writing circuitry for ultralow-voltage mtcmos srams. *Proc.IEICE Electron*, page 100, 2003.

[4] E. Karl et al. The impact of assist-circuit design for 22nm sram and beyond. *Electron Devices Meeting (IEDM)*, page 25, 2012.

[5] Chien-Yu Lu et al. A 0.33-v, 500-khz, 3.94-uw 40-nm 72-kb 9t subthreshold sram with ripple bit-line structure and negative bit-line write-assist. *Circuits and Systems II*, 59(12) :863–867, 2012.

[6] T. Song et al. A 14 nm finfet 128 mb sram with vmin enhancement techniques for low-power applications. *JSSC*, 50(1) :158–169, 2015.

[7] S. Keshavarapu et al. A new assist technique to enhance the read and write margins of low voltage sram cell. *International Symposium on Electronic System Design*, pages 97–101, 2012.

[8] C Y. Chen et al. A high-performance low-power highly manufacturable embedded dram technology using backend hi-k mim capacitor at 40nm node and beyond. *VLSI-TSA*, pages 1–2, 2011.

[9] K. Kim et al. Asymmetrical sram cells with enhanced read and write margins. *VLSI-TSA*, pages 162–3, 2007.

[10] E. Seevinck et al. Static-noise margin analysis of mos sram cells. *IEEE, J. of Sol.-State Circuits*, 22 :748, 1987.

[11] H. Makino et al. Reexamination of sram cell write margin definitions in view of predicting the distribution. *Circuits and Systems II*, 58(4) :230–234, 2011.

[12] L. Ciampolini et al. Circuit-level modeling of sram minimum operating voltage vddmin in the c40 node. *Journal of Low Power Electronics*, pages 106–112, 2011.