Circuit Techniques for Ultra-Low Power Subthreshold SRAMs

Tae-Hyoung Kim, Jason Liu, John Keane and Chris H. Kim

Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN
{thkim, jliu, jkeane, chriskim}@umn.edu

***Abstract*—Subthreshold operation has become an important area in applications where minimal power consumption and energy efficiency are the critical constraints. In particular, ultra-low power SRAM designs are critical for implementing such applications due to the large portion of the systems that they account for. However, sub-threshold SRAMs have many design issues such as cell stability, readability, and writability. In this paper, we give an overview of sub-threshold SRAM design issues and discuss several circuit techniques. We will focus on SRAM cell stability during read and write operation, improved writability, and read port circuits for the design of an ultra-low power sub-threshold SRAMs.**

I. INTRODUCTION

Digital subthreshold logic is becoming increasingly popular for ultra-low power applications where minimal power consumption is the primary design constraint [1][2][3]. Subthreshold static CMOS logic can operate while consuming roughly an order of magnitude less power than in the normal strong-inversion region. Designing robust SRAM for subthreshold systems is extremely challenging due to the reduced voltage margin, degraded *Ion-to-Ioff* ratio, and increased device variability. Conventional 6-T SRAMs in the subthreshold region fail to deliver sufficient density and yield requirements due to the reduced Static Noise Margin (SNM), poor writability, limited number of cells per bitline, and reduced bitline sensing margin.

This paper introduces various circuit techniques for designing robust and high-density SRAMs in the subthreshold regime, covering both SRAM cell and peripheral circuit design. The following techniques are proposed to enable a fully functional 480kb SRAM operating at 0.2V with 1k cells in a bitline. First, a decoupled 10-T SRAM cell is proposed for SNM improvement. This design also eliminates data-dependent bitline leakage, thereby enabling 1k cells per bitline. Second, the Reverse Short Channel Effect (RSCE) is utilized for write margin improvement. Third, a Virtual Ground (VGND) replica scheme improves the bitline sensing margin. Finally, a writeback scheme is adopted to preserve the row data in unselected columns during write operation. A 130nm SRAM test chip incorporating each of these techniques was fabricated and successfully measured.

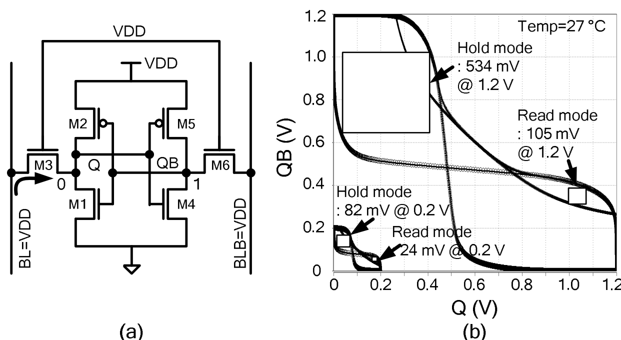

Figure 1. (a) Worst case SNM scenario. (b) SNM simulation results.

This work has been sponsored by Intel, IBM, and United Microelectronics Corporation (UMC).

II. DESIGN ISSUES IN 6-T SUB-THRESHOLD SRAMS

A. 6-T SRAM Cell Stability

SNM is an important criterion in estimating the stability of an SRAM cell. SNM is decided by the combination of cross-coupled inverters and access devices. The worst case in stability happens in the accessed cell during a read operation, and in the unselected columns during write operation. Fig. 1 shows the worst case scenario in SNM and simulation results comparing the read mode SNM to the hold mode SNM. The worst case SNM becomes so small as the supply voltage is reduced that the 6-T SRAM cell cannot be used in the subthreshold regime. Several techniques have been proposed to improve the SNM [4][5][6][7][8][9][10]. A widely used method is decoupling the cell node from bitline by using additional read port transistors [5][8]. By doing this, the SNM in read mode becomes equal to that in hold mode.

B. Degraded Writability due to PVT Variations

The write margin is decided by the ratio between pull-up PMOSs (M2, M5) and access transistors (M3, M6). Those access transistors are typically sized to be stronger than the pull-up PMOSs. However, the write margin problem is exacerbated in the subthreshold regime due to the increased current sensitivity to PVT variations. Using a higher supply level for driving the wordline, and a collapsed cell supply can improve the write margin at the cost of additional power consumption, generation circuitry, routing, and degraded stability [5][8][9].

C. Impact of Bitline Leakage Current on Readibility

As the supply voltage is scaled, the *Ion-to-Ioff* ratio decreases exponentially [11]. The small *Ion-to-Ioff* ratio in the subthreshold region limits the number of cells per bitline. As the number of cells in a bitline increases, bitline leakage from the unaccessed cells is comparable to the read current of the accessed cell, making it difficult to distinguish the bitline high and low levels. When reading a '1', the worst case read bitline

(RBL) voltage is determined based on the contention between the pull up current from the accessed cell and the pull down bitline leakage currents from the unaccessed cells. Likewise, when reading a '0', the contention between the pull down current of the accessed cell and the pull up bitline leakage currents of the unaccessed cells decides the worst case RBL voltage. As the number of cells per bitline increases, the worst case RBL for data '1' decreases and that for data '0' increases due to the bitline leakage current. As a result, the bitline voltage for data '1' may be lower than that for data '0' under the worst case data patterns.

III. PROPOSED CIRCUIT TECHNIQUES FOR SUB-THRESHOLD SRAMS

A. *Decoupled SRAM Cell and Writeback Scheme for Improved Stability*

Several methods have been proposed to improve SNM without decoupling the cell node from the bitline [6][7][9]. However, these optimization techniques offer limited improvement due to the disturbed cell node. Fig. 2 shows two decoupled SRAM cells for improving SNM without cell node disturbance [5]. Each SRAM cell consists of a cross-coupled inverter pair, write access devices, and read port circuits which decouple the cell node from the bitline. When the read wordline is enabled (RWL=1), RBL is conditionally discharged through the pull-down transistors depending on the stored data. The cell node is not disturbed by the read bitline, retaining a hold mode SNM during the read operation.

Another source of stability degradation happens during write operations. A column muxing scheme is widely used in SRAM designs for efficient area usage and large density. However, this scheme causes a write disturbance in unselected columns due to the shared wordline. One method for solving the write disturbance problem is using a distributed wordline driver within a cell [12]. However, this comes at the cost of two additional transistors in each cell. A writeback scheme shown in Fig. 3 is proposed to solve the disturbed stability problem during write operations. A write operation is executed by enabling both a read wordline (RWL<i>) and a write wordline (WWL<i>) simultaneously. In the unselected columns (Y<i>=0), the write bitlines are held at VDD and a read operation is executed. The writeback enable signal (WB) is activated by the rising edge of RWL, after a sufficient delay created by a replica bitline enabling the writeback path. The read data from the sense amplifier is transferred to D_INT and written back to the write bitlines (WBL, WBLB). By writing the read data back to WBL and WBLB, there is no voltage difference between write bitlines and the cell nodes, eliminating the contention current. That contention current is only seen between the write wordline rising edge and the rising edge of WB, corresponding to the read finish. The area overhead of this scheme is small because the additional gates in the write driver are shared by all of the SRAM cells in one column.

B. *Write Margin Improvement Techniques*

Maintaining a sufficient write margin is challenging in subthreshold SRAMs due to the small gate overdrive and large impact of process variations in the write access devices (M3

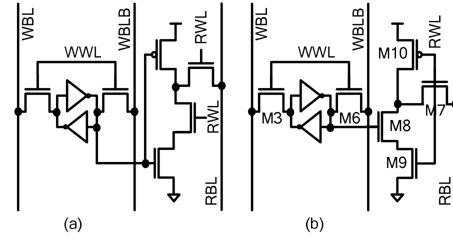

Figure 2. Decoupled SRAM cells: (a) 10-T in [5]. (c) Proposed 10-T.

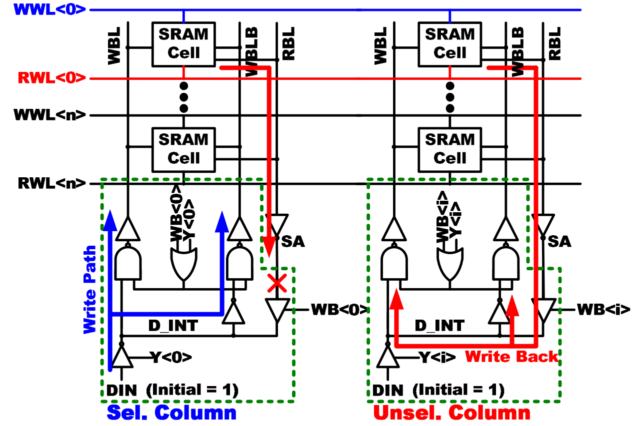

Figure 3. Writeback scheme for minimizing disturbance.

and M6 in Fig. 2 (b)). Virtual supply rails, as shown in Fig. 4 (a), have been used in previous work to improve cell writability [5][8][9]. The cell supply voltage of the selected column is left floating during a write operation. The virtual supply rails collapse making it easier for the write access devices to flip the cell value due to the weakened PMOS transistors. However, this technique is not suitable in subthreshold SRAMs as the virtual supply droop cannot be controlled accurately, and the SNM is already close to its functional limitation. Fig. 4 (b) shows another technique using a wordline voltage which is higher than the cell voltage to increase the drive current of the write access transistors [5]. However, this technique increases power consumption, and requires an additional high VDD to be generated and routed.

In this work, we utilize the RSCE in the subthreshold region to improve cell writability without introducing a separate high VDD or collapsing supply rails. RSCE is observed in modern CMOS devices due to the HALO pocket implants used to compensate for the V_{th} roll-off [13]. RSCE is not a major concern in conventional strong-inversion designs since SCE is dominant in minimum channel length devices in that region. However, in the subthreshold region, the RSCE is dominant due to the significantly reduced Drain Induced Barrier Lowering (DIBL) [11]. This causes the V_{th} to decrease monotonically with increasing channel length, as shown in Fig. 5 (a). The optimal device length for our application is found at the channel length where a minimum delay is achieved with a fixed amount of current drivability, which is illustrated in Fig. 5 (b). In the 0.13µm technology used in this paper, we find that minimum energy point at a channel length of 0.36µm, which is 3X larger than the minimum value.

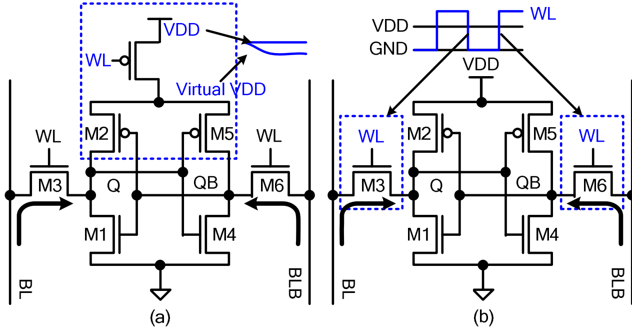

Figure 4. Previous write margin improvement techniques. (a) Collapsed supply rails. (b) Boosted wordline voltage.

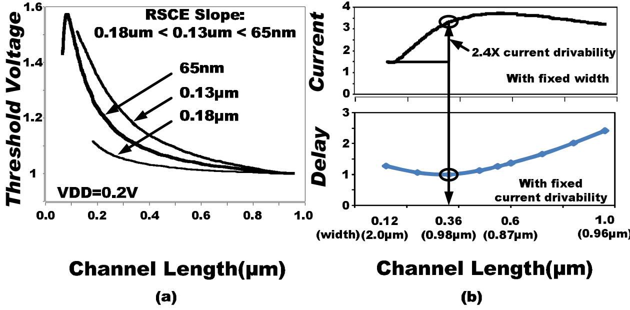

Figure 5. (a) Vth normalized with Vth at long channel to show the derivative of Vth in different technology nodes. (b) Normalized current drivability and energy consumption.

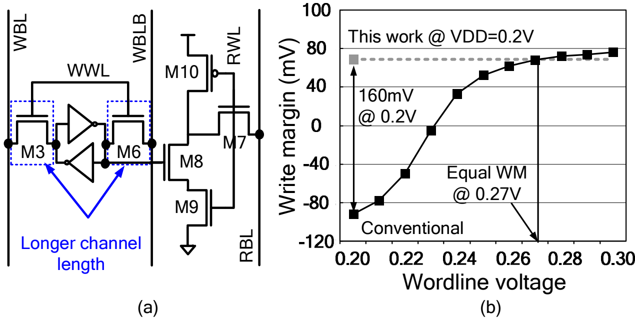

Figure 6. (a) Proposed 10-T SRAM cell utilizing RSCE. (b) Write margin versus wordline voltage.

Therefore, we took advantage of RSCE by using write access transistors that are 3X the minimum length to improve cell writability in our design (Fig. 6 (a)). The stronger drive current enables a robust write operation and hence lowers the minimum operating voltage. Simulation results in Fig. 6 (b) show that the writability of the proposed SRAM at 0.2V is equivalent to that of a cell using the conventional sizing scheme with a WWL voltage of 0.27V.

C. *Read Port Circuits*

The small *Ion-to-Ioff* ratio in the subthreshold region limits the number of cells per bitline, which negatively impacts the SRAM density. As the number of cells in a bitline increases, bitline leakage from the unaccessed cells can rival the read current of the accessed cell, making it difficult to distinguish between the bitline high and low levels. A hierarchical bitline

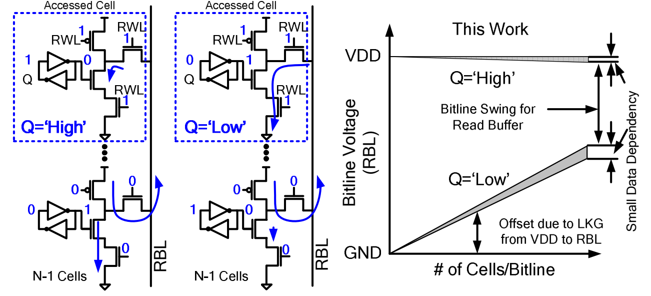

Figure 7. (a) Bitline without data dependency. (b) Read bitline voltage.

can alleviate this bitline leakage problem at the cost of large area overhead due to additional logic gates in each hierarchy [14]. Another technique is reducing the amount of bitline leakage current by utilizing charge-pump circuit in each row [8]. However, the data dependent leakage current still makes it difficult to reliably read data. A 0.3V subthreshold SRAM with 256 cells on a single bitline was reported in [5]. However, our simulations indicate that the maximum number of cells per bitline of the prior design quickly reduces to 16 at a supply voltage of 0.2V.

The proposed 10-T SRAM cell with a 4-T read port circuit (Fig. 6 (a)) eliminates the data-dependent bitline leakage problem by making bitline leakage current flow in one direction. In unaccessed SRAM cells, the drain voltage of M10 becomes VDD, forcing the leakage current to always flow from the cell into the bitline regardless of the stored data. The logic low level is generated by the balance between the pull-up leakage current of unaccessed cells and the pull-down read current of the accessed cell, as shown in Fig. 7. The logic high level is close to VDD because the M10 drain voltage is high in all unaccessed cells and the leakage into the accessed cell is negligible. As a result, we get a fixed bitline low and high level irrespective of the column data pattern.

D. *Sense Amplifier Design for Optimal Sensing Margin*

In subthreshold SRAMs, sense amplifiers are replaced with static inverter type read buffers because noise margin is the key design concern rather than speed [5][8]. That is because these read buffers have the maximum sensing margin for a given supply voltage, if there is a full swing on the bitline. However, the reduced bitline swing due to the leakage explained in the previous subsection necessitates a careful sense amplifier design. A sense amplifier with redundancy was introduced in [8] to reduce the probability of sense amplifier failure. This scheme requires an extra step to select one sense amplifier based upon the measurement results. In this paper, a VGND replica scheme is devised to maximize the sensing margin of the sense amplifiers based on the fact that the bitline high and low levels are insensitive to the column data pattern (Section III.C). The VGND represents the logic low level on RBL. It is generated from a replica bitline and shared by several sense amplifiers as the ground level (Fig. 8 (b)). The proposed VGND replica scheme automatically tracks the optimal read buffer trip point to obtain the largest possible sensing margin as shown in Fig. 8 (a). The trip point of the read buffer is set to the middle of the logic high and low levels

by using the VGND level as the ground level of the read buffer as shown in Fig. 8 (a).

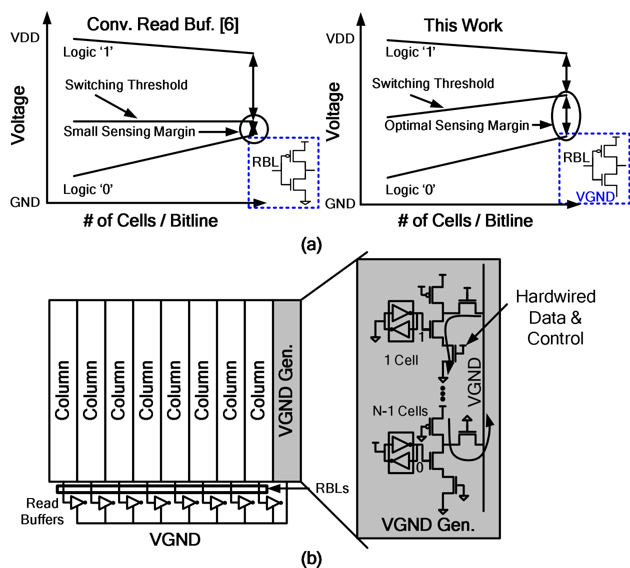

Figure 8. (a) Bitline with data-dependent leakage. (b) Read bitline voltage.

IV. Experimental Results

A 4.1x1.5mm^2 SRAM with 480kb cells was fabricated in a 130nm, 8-metal CMOS technology. The test chip microphotograph is shown in Fig. 10 (c). Fig. 9 (a) and (b) show the VGND measurement results. VGND becomes as high as 50% of the supply voltage at 0.2V for a bitline with 1k cells attached. This voltage level is relatively independent of temperature. Fig. 9 (c) and (d) illustrate the leakage current, power, and the maximum operating frequency. The leakage current of the 480k SRAM is 10µA for a supply voltage of 0.2V at 27°C. The maximum operating frequency is 100kHz at 0.2V and 27°C with 1k cells per bitline. The access time and maximum operating frequency of four quadrants are shown in Fig. 10 (a) and (b). As supply voltage increases, maximum operating frequency increases exponentially. The quadrant with 1k cells per bitline was readable at a supply voltage of 0.17V, which is shown in Fig. 9 (d). A delay improvement of 28% was obtained in row decoder by utilizing RSCE.

V. Conclusions

Several circuit techniques for ultra-low power sub-threshold SRAMs have been demonstrated. A 10-T SRAM cell is proposed to eliminate the read failure caused by data-dependent bitline leakage. A VGND replica scheme is proposed to track the logic 'low' level of the bitlines under PVT variations to achieve the maximum read sensing margin. The strong RSCE in the subthreshold region was utilized to improve cell writability, reduce power consumption, improve logic performance, and enhance circuit immunity to process variations. Combining the proposed circuit techniques, we successfully realized a fully functional subthreshold SRAM with 1k cells per bitline operating at 0.2V and 27°C.

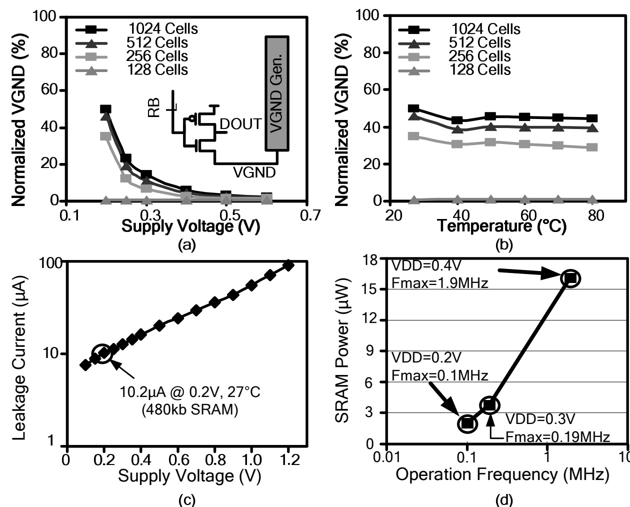

Figure 9. Measured results: (a) Supply voltage dependency of VGND. (b) Temperature dependency of VGND. (c) Leakage current vs. supply voltage. (d) SRAM power and max. operating frequency versus supply voltage.

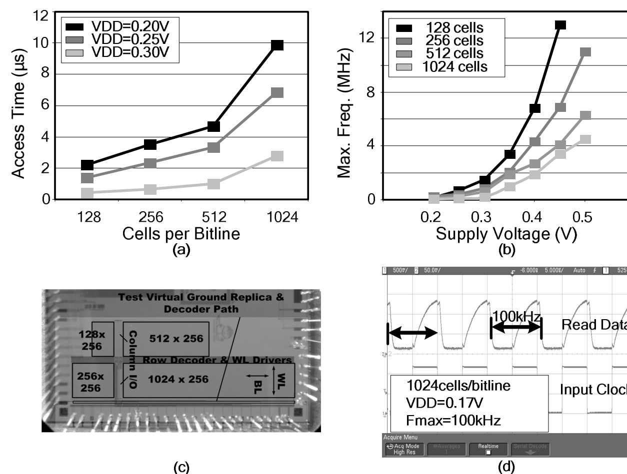

Figure 10. (a) Access time of four quadrants versus supply voltage. (b) Maximum operating frequency of four quadrants versus supply voltage. (c) Test chip microphotograph showing different sized quadrants. (d) Read data waveform at minimum supply voltage.

References

[1] H. Soeleman et al., IEEE Trans. VLSI Systems, pp. 90-99, Feb. 2001.
[2] B. Zhai et al., ISLPED Dig., pp. 20-25, Aug. 2005.
[3] A. Wang et al., IEEE J. SSC, pp. 310-319, Jan. 2005.
[4] J. Kim et al., IEEE Int'l. SOI Conf. Dig., pp. 211-212, Oct. 2005.
[5] B.H. Calhoun et al., ISSCC Dig., pp. 628-629, Feb. 2006.
[6] L. Chang et al., Symp. VLSI Technology Dig., pp. 128-129, June 2005.
[7] M. Khellah et al., Symp. VLSI Circuits Dig., pp. 9-10, June 2006.
[8] Naveen Verma et al., ISSCC Dig., pp. 328-329, Feb. 2007.
[9] Bo Zhai et al., ISSCC Dig., pp. 332-333, Feb. 2007.
[10] L. Chang et al., Symp. VLSI Circuits Dig., pp 252-253, June 2007.
[11] T. Kim et al., ISLPED Dig., pp 127-130, Oct. 2006.
[12] V. Ramadurai et al., CICC, pp 25-28, Sep. 2007.
[13] B. Yu et al., Symp. VLSI Technology, pp. 162-163, June 1996.
[14] J. Chen et al., IEEE J. SSC, pp. 2344-2353, Oct. 2006.