Cross-coupled Bit-Line Biasing for 22-nm SRAM

David Halupka and Ali Sheikholeslami
Department of Electrical and Computer Engineering
University of Toronto, Toronto, ON, CANADA
{halupka,ali}@eecg.utoronto.ca

***Abstract*—Sub-22 nm processes necessitate the use of larger SRAM cells in order to counter the effects of increasing silicon variation. However, storage density is reduced when the SRAM cell grows in size. This paper proposes the use of a cross-coupled bit line (*BL*) biasing scheme that retains SRAM's fast access speed while reducing the read-access failures in the presence of V_t variation, without excessively increasing the SRAM cell size. We have shown, by extensive Monte-Carlo simulations using 22-nm predictive CMOS models, that the proposed scheme reduces the cell area by 6.5% compared to the conventional BL biasing schemes also analyzed in this paper.**

I. Introduction

Silicon variation is predicted to increase as transistors scale down to 22 nm and beyond. The ITRS predicts a 3-sigma V_t variation of 58% of nominal V_t for 22-nm CMOS [1]. Random V_t variation causes SRAM memories, such as ones based on the six-transistor (6T) cell shown in Fig. 1, to experience functional and timing failures.

Several techniques can be used to mitigate the effects of increasing V_t variation. *1.* Use row or column redundancy to bypass faulty cells, but as faulty cells tend to be randomly distributed in the array the amount of redundancy required increases rapidly with increasing variation [2]. *2.* Use error correcting codes (ECC) to detect and correct faulty bits. While ECC is able to correct multi-bit errors, the area and access time overhead can be undesirable for some applications. *3.* Use non-standard SRAM cell to address the read functionality [3], [4] or the write functionality [5] of a cell. However, these cell structures typically require extra transistors, and hence area. *4.* Use the 6T SRAM cell, but increase the size of the cell's transistors. This has been shown to be a preferred approach over row/column redundancy when total memory area, including any row/column redundancy, is considered [2]. *5.* Use row or column assist techniques to address the write [6], [7] or read (this work) functionality. While these techniques do mitigate the effects of V_t variation, most of them do so at the cost of cell area or performance overhead.

In this paper we show that bit-line (BL) biasing and BL conditioning circuits can also be used to mitigating the effects of V_t variation, without any cell area or performance overhead. Two common methods for biasing the BLs are shown in Fig. 2a and Fig. 2b. The first method involves pre-charging the BL to V_{DD} during the pre-charge phase, while the second technique resistively ties the BL to V_{DD}. In this paper, we propose the use of a cross-coupled BL biasing scheme (Fig. 3) to mitigate the effects of V_t variation on the read functionality of a 6T SRAM cell. The proposed scheme allows for the use of smaller SRAM cell transistors, in comparison to the two conventional BL biasing methods. We will show later that the cross-coupled BL biasing scheme can use SRAM cells that occupy 6.5% less silicon area, while maintaining functionality and performance in the face of V_t variation. Area savings for the proposed scheme are due to simultaneously improving two opposing factors: the readability of a cell and the cell's read access time. The interaction of these factors and how they can be simultaneously improved will be discussed and analyzed in this paper using a predictive 22-nm CMOS technology [8].

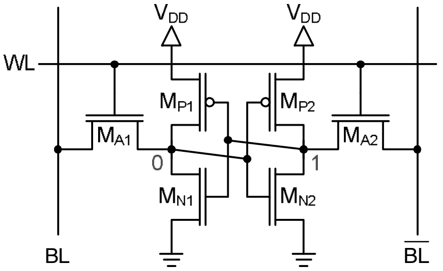
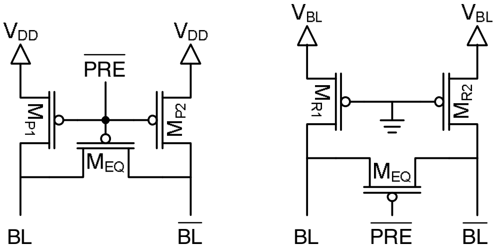


Fig. 1. 6T SRAM cell, storing a "0" on the left side and a "1" on the right.



(a) Pre-charged BLs (b) Resistive BLs

Fig. 2. Typically, BLs are (a) pre-charged or (b) resistively tied to V_{DD}.

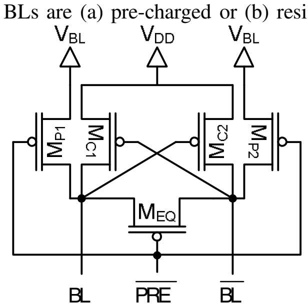


Fig. 3. In the proposed scheme, BLs are pre-charged to a voltage below V_{DD}, and the two cross-coupled PMOS transistors help amplify the data being read from a cell.

This paper is organized as follows: Section II discusses the

978-1-4244-3732-0/09/$25.00 ©2009 IEEE

impact of varying the 6T SRAM cell's BL pre-charge voltage. Section III introduces the proposed BL-biasing scheme and its advantages. Section IV compares the proposed BL-biasing scheme with the conventional biasing schemes and with the write column-assist scheme [7] in terms of cell area required for an increasing amount of V_t variation. Section V discusses leakage current savings due to the proposed scheme.

II. Background

Two important metrics that are used to characterize an SRAM cell are the read noise margin (RNM) and the read access time. RNM is defined as the maximum amount of noise a cell can tolerate during a read operation without data loss [9]. Read access time is defined as the time required to develop a sufficient voltage difference between *BL* and $\overline{BL}$ following *WL* activation. Both of these metrics are influenced by the BL pre-charge voltage, which we discuss next.

According to our analysis, a lower $V_{BL}(t=0)$ reduces the voltage rise on the cell's logic-0 storage node during a read operation, thereby increasing the RNM. A plot of RNM versus $V_{BL}(t=0)$ is shown in Fig. 4a for a typical SRAM cell in 22 nm CMOS. Note that the RNM is largest when $V_{BL}(t=0)$ is approximately 460 mV. RNM begins to drop below a $V_{BL}(t=0)$ of 460 mV, as the voltage of the logic-1 storing node begins to be pulled low (via the access transistor) to $V_{BL}(t=0)$, resembling a write operation. For $V_{BL}(t=0)$ below 300 mV all read accesses are destructive. On the other hand, read access time increases with decreasing $V_{BL}(t=0)$, up until the point where the read operations become destructive. A plot of the read access time versus $V_{BL}(t=0)$ is shown in Fig. 4b for a purely capacitive *BL* and $\overline{BL}$. The minimum read access time occurs when $V_{BL}(t=0)=V_{DD}$.

The optimum points for RNM and read access time do not correspond to the same $V_{BL}(t=0)$, as shown in Fig. 4a and Fig. 4b, thereby creating a trade-off between the RNM and the read access time. A plot of this trade-off curve for a typical 22 nm 6T SRAM cell and varying $V_{BL}(t=0)$ is shown in Fig. 4c. Transistor sizing does not change the general shape of this trade-off curve; it merely stretches and displaces the curve on the RNM vs. read access time plane. The proposed cross-coupled BL-biasing scheme allows us to operate at a point that is off this curve, as indicated in Fig. 4c. By operating off this trade-off curve we can achieve a higher RNM without sacrificing read access time or cell area.

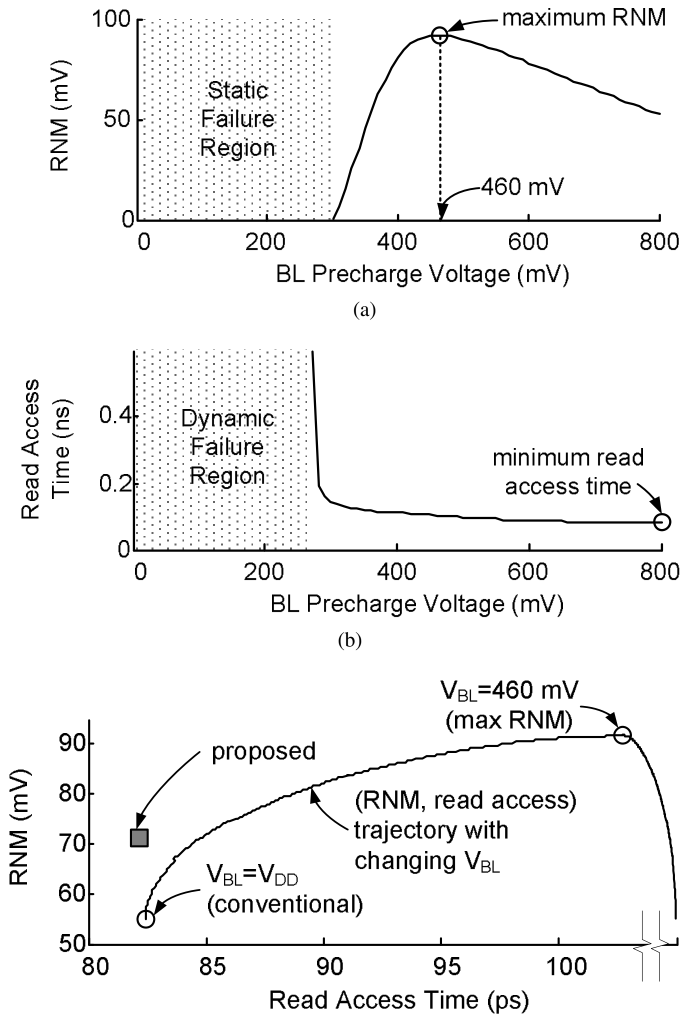
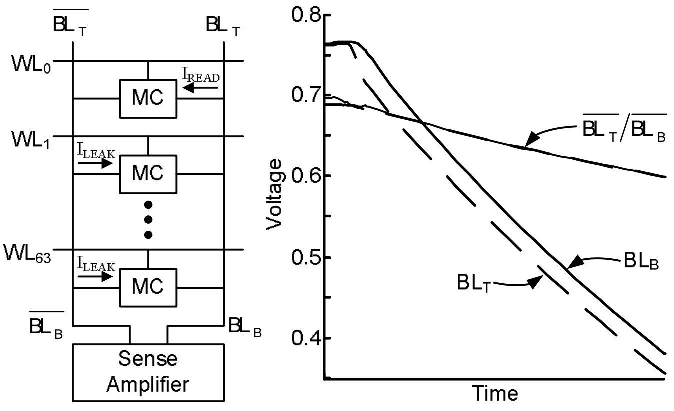


Fig. 4. (a) Maximum RNM occurs when the BLs are pre-charged to 460 mV, but (b) minimum read access time occurs when the BLs are pre-charged to V_{DD}. (c) RNM can be traded for read access time with decreasing BL pre-charge level. The proposed scheme allows for a 30% increase in RNM without loss of read access time.

III. Cross-Coupled BL Biasing Scheme

The cross-coupled BL biasing scheme, shown in Fig. 3, uses two cross-coupled PMOS transistors (M_{C1} and M_{C2}) along with an equalizing transistor (M_{EQ}) and pre-charge transistors (M_{P1} and M_{P2}) that are controlled by the pre-charge signal ($\overline{PRE}$). During the pre-charge phase ($\overline{PRE}$ is low) *BL* and $\overline{BL}$ are charged to $V_{DD}-|V_{tp}|$ ($\sim$0.5 V). Note, during pre-charge M_{C1} and M_{C2} are effectively diode connected due to M_{EQ}. At the end of the pre-charge phase transistors M_{C1} and M_{C2} should be barely on, ready to help amplify the voltage difference between *BL* and $\overline{BL}$. Pre-charge transistors M_{P1} and M_{P2}, which are in parallel with transistors M_{C1} and M_{C2}, are used to shorten the required pre-charge time and to provide better control of the pre-charge voltage level.

At the start of a read operation ($\overline{PRE}$ is high) transistors M_{C1} and M_{C2} are no longer diode connected, but are in a cross-coupled configuration instead. This is similar to a dynamic sense-amplifier configuration [10]. This PMOS configuration helps to amplify the voltage being developed on the BLs by the accessed cell, resulting in a shorter read access time as indicated in Fig. 4c. For example, when a word line (*WL*) is driven high, the corresponding cell starts to discharge *BL* to ground, whereas the cross-coupled PMOS transistors charge $\overline{BL}$ to V_{DD}. Hence, if the voltage on *BL* begins to decrease, the current supplied to $\overline{BL}$ by M_{C2} will begin to increase, slowly charging $\overline{BL}$ further up to V_{DD} and thereby amplifying the *BL*/$\overline{BL}$ voltage difference. Therefore, the dynamic amplification provided by M_{C1} and M_{C2} decreases the read access



(a) Read/Leakage currents

(b) *BL* at column extremes

Fig. 5. (a) Distributed capacitance, resistance and cell leakage along the column cause worst-case read access time to be determined by the cell farthest away from the sense amplifiers. (b) Timing simulation shows how the *BL* signal at the bottom of the column (BL_B) develops relative to the top of the column (BL_T). Also shown is the effect of distributed leakage on $\overline{BL}$.

time, compensating for the increased read access time due to a lowered $V_{BL}(t=0)$.

A. Effect on read access time

The location of transistors M_{C1} and M_{C2} also has an impact on the read-access time, as the data that is being read from a cell needs to propagate through the BLs to reach the sense amplifiers, typically at the bottom of the column. Hence, a cell at the top of a memory column, such as in Fig. 5a, experiences the largest delay to the sense amplifiers that are located at the bottom of the column. In contrast, a cell at the bottom of the column experiences the smallest delay. The transient simulation plot in Fig. 5b shows the impact of signal propagation from a cell at the top of the column along the *BL*, and the impact of non-accessed distributed cell leakage on $\overline{BL}$.

As transistors M_{C1} and M_{C2} are not controlled by any external signals, it is possible to place them at the top of the memory column. This location allows these transistors to help propagate the data from the top of the array, towards the bottom, by locally amplifying the voltage difference between *BL* and $\overline{BL}$. This local amplification creates the required difference in voltage between *BL* and $\overline{BL}$ at the input of the sense-amplifiers in a shorter amount of time.

B. Effect on write noise margin

For a 6T SRAM cell, there is also a trade-off between RNM and write noise margin (WNM), where WNM is defined as the maximum amount of noise a cell can tolerate while remaining writeable [9]. For the conventional biasing scheme, this trade-off curve is shown by the dashed line in Fig. 6, where the aspect ratio of M_A in the 6T cell (in Fig. 1) is varied for each simulation. To achieve a high RNM M_A needs to be weaker than M_N, whereas for a high WNM M_A needs to be stronger than M_P. These two constraints make M_A a balancing point for the cell, forcing the designer to trade-off WNM in favor of RNM, under a fixed cell-area constraint.

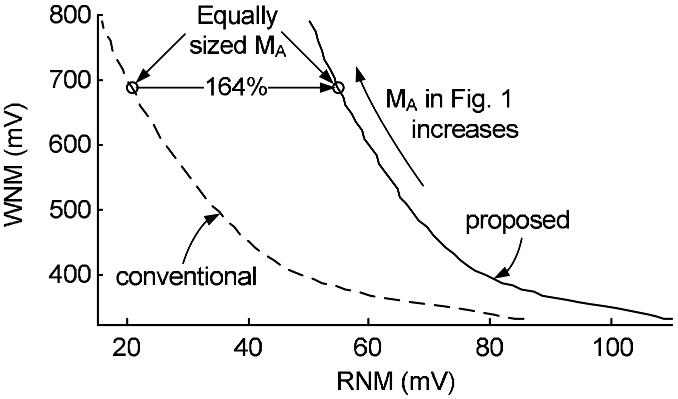


Fig. 6. WNM & RNM trajectory as the size of M_A is varied. For the same sized M_A, the proposed scheme achieves a 164% larger RNM without any sacrifice to WNM.

The cross-coupled BL scheme allows for a larger RNM, without any change to WNM, as shown by the solid line in Fig. 6. This enables the designer to automatically obtain more RNM (164% as indicated in Fig. 6) without any change to the WNM or read access time. Alternatively, by using the cross-coupled BL scheme the designer can keep the cell's RNM constant by increasing the size of M_A, which improves both the WNM and read access time; effectively, moving vertically from the dashed curve to the solid curve.

C. SRAM cell area under increasing V_t variation

The effects of V_t variation can be mitigated by increasing the channel area of the SRAM cell's transistors. However, this has a direct impact on the silicon area required for each SRAM cell, and hence the memory array's bit density. Having outlined some of the advantages of the proposed scheme, we will now examine how these advantages change the required silicon area as V_t variation increases. A plot of the minimum silicon area required for an SRAM cell using proposed cross-coupled BL-biasing scheme (Fig. 3), the conventional pre-charged BL biasing scheme (Fig. 2a), the pre-charge biasing scheme using a pre-charge voltage of $\sim$0.5 V, a resistively tied BL-biasing scheme (Fig. 2b), and the scheme proposed by [7] is shown in Fig. 7. For any amount of V_t variation, the proposed cross-coupled BL biasing scheme is able to use SRAM cells that require less silicon area then the conventional schemes, while attaining equivalent functionality and access time. Cell area is reported for a commercially available 90-nm CMOS technology, using digital logic layout rules, and a thin-cell layout style [11]. Transistor sizes are mapped from the 22-nm predictive technology onto the 90-nm technology by a constant scaling factor. The area required for transistors M_{C1} and M_{C2} is not included in the plot cell area, as the size of these transistors is independent of V_t variation.

The increased RNM of the cross-coupled BL biasing scheme can be obtained by using the conventional pre-charge scheme with a lowered pre-charge voltage (i.e. V_{BL}=0.5 V). However, this modified pre-charge BL biasing scheme has a much slower read access time, which is compensated for by

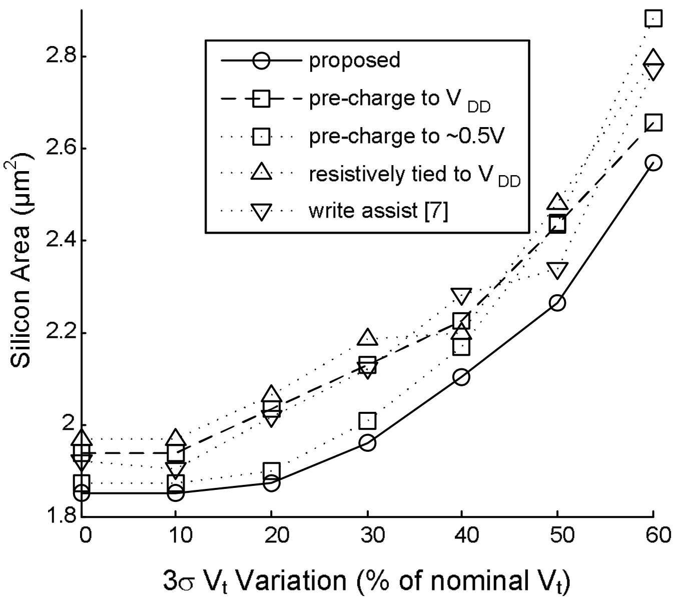


Fig. 7. 22 nm cell area required to compensate for an increasing V$_t$ variation.

making the cell's access transistors larger, thereby making the cell use more silicon area as indicated in Fig. 7. Note that the write enhancing scheme proposed by [7] does not result in the ability to use SRAM cells with a smaller cell area.

The BL biasing schemes and the write-enhancing scheme of [7] are compared using the cell optimizer described in [2]. The optimization function used by this optimizer consists of dynamically weighted design goals: the cell's failure probability, read and write access times. The goal of the optimizer is to find a cell with a minimum cell area that achieves a failure rate below 5 ppm, and a read and write access time below 500 ps and 350 ps, respectively.

The cell optimizer makes use of the equivalent half-cell circuits described in [2] to quickly estimate the cell's probability of a functional read and write failure. These half-cell circuits provide a method of measuring the cell's threshold voltage (V_{TRIP}), the logic-0 voltage under a read condition (V_{READ}) and the logic-1 voltage under a write condition (V_{WRITE}). A cell fails to be functionally readable when $V_{READ}>V_{TRIP}$, an inadvertent write to the cell. A cell fails to be functionally writeable when $V_{WRITE}>V_{TRIP}$, the cell contents are not changed. Read and write access times are measured for a 64-bit column, with the remaining 63 non-accessed cells configured to create worst-case BL leakage for the read or write access. Parasitic BL wire capacitance and resistance is incorporated into the model, based on 90-nm extracted values, as values for the 22-nm CMOS are not readily available.

The cell optimizer systematically analyzes all possible cell transistor sizes in order to find a cell with minimum area that meets the design constraints. One million Monte-Carlo simulations are run for each cell under consideration by the optimizer in order to obtain an estimate of the failure probability of a cell. Furthermore, over one thousand Monte-Carlo transient simulations are used to obtain the average read and write access times, and to count any additional dynamic read and write failures.

D. *Effect on cell leakage current*

The lower $V_{BL}(t=0)$ greatly decreases a cell's BL leakage current. For example, in a 22-nm cell a 12.5% reduction in $V_{BL}(t=0)$ yields a 51.1% reduction in BL leakage current. Therefore, cells optimized for the cross-coupled BL biasing scheme waste 56.6% less leakage current overall, than cells optimized for the conventional pre-charge to V_{DD} scheme. The combined worst-case BL leakage current is insufficient to trigger the cross-coupled PMOS transistors between $\overline{PRE}$ deactivation and *WL* activation.

IV. Conclusion

The proposed cross-coupled BL-biasing scheme increased the 6T cell's RNM by 30%, without significantly impacting cell's read access time. The cross-coupled BL-biasing scheme allows for the use of SRAM cells that use 6.5% less silicon area and waste 56.6% less leakage current than the conventional pre-charge to V_{DD} BL-biasing scheme. We have compared the cross-coupled BL-biasing scheme to various other BL biasing schemes and a write-enhancing scheme proposed by [7]; the cross-coupled BL biasing scheme is able to use 5% less silicon area then these other schemes as well.

Acknowledgements

This work was sponsored by the National Sciences and Engineering Research Council of Canada. Our sincere thanks to O. Tyshchenko for his helpful discussions and feedback.

References

[1] International Technology Roadmap for Semiconductors, "ITRS 2008 update," Website, February 2009, http://www.itrs.net/Links/2008ITRS/Home2008.htm.

[2] S. Mukhopadhyay *et al.*, "Modeling of failure probability and statistical design of SRAM array yield enhancement in nanoscaled CMOS," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 12, pp. 1859–1880, 2005.

[3] L. Chang *et al.*, "Stable SRAM cell design for the 32 nm node and below," in *VLSI Technical Digest*, 2006, pp. 128–129.

[4] K. Takeda *et al.*, "A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications," *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 113–121, 2006.

[5] R. E. Aly and M. A. Bayoumi, "Low-power cache design using 7T SRAM cell," *IEEE Trans. Circuits Syst. II*, vol. 54, no. 4, pp. 318–322, 2007.

[6] M. Yamaoka *et al.*, "90-nm process-variation adaptive embedded sram modules with power-line-floating write technique," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 705–711, 2006.

[7] H. Yamauchi *et al.*, "A differential cell terminal biasing scheme enabling a stable write operation against a large random threshold voltage (Vth) variation," *IEICE Trans. on Electronics*, vol. E89-C, no. 11, pp. 1526–1534, 2006.

[8] Nanoscale Integration and Modeling (NIMO) Group, "Predictive technology model," Website, June 2007, http://www.eas.asu.edu/~ptm/.

[9] A. Bhavnagarwala *et al.*, "Fluctuation limits & scaling opportunities for CMOS SRAM cells," in *IEDM Techical Digtest*, 2005, pp. 659–662.

[10] N. Wang, "On the design of MOS dynamic sense amplifiers," *IEEE Trans. Circuits Syst.*, vol. 29, no. 7, pp. 467–477, 1982.

[11] M. Ishida *et al.*, "A novel 6T-SRAM cell technology designed with rectangular patterns scalable beyond 0.18 μm generation and desirable for ultra high speed operation," in *IEDM Technical Digest*, 1998, pp. 201–204.