A TECHNIQUE FOR DESIGNING VARIATION RESILIENT SUBTHRESHOLD SRAM CELL

AMINUL ISLAM

Department of Electronics and Communication Engineering,
Birla Institute of Technology (Deemed University), Mesra, Ranchi, Jharkhand, India.
aminulisam@bitmesra.ac.in

ABSTRACT: This paper presents a technique for designing a variability aware subthreshold SRAM cell. The architecture of the proposed cell is identical to the standard read-decoupled 8-transistor (RD8T) SRAM cell with an exception that the access FETS are replaced with transmission gates (TGs). In this work, different design metrics are assessed and compared with RD8T SRAM cell. The proposed design offers 2.14× and 1.75× improvement in $T_{RA}$ (read access time) and $T_{WA}$ (write access time) respectively compared with RD8T. It proves its robustness against process variations by featuring narrower spread in $T_{RA}$ distribution (2.35×) and $T_{WA}$ distribution (3.79×) compared with RD8T. The proposed bitcell offers 1.16× higher read current ($I_{READ}$) and 1.64× lower bitline leakage current ($I_{LEAK}$) respectively compared with RD8T. It also shows its robustness by offering 1.34× (1.58×) tighter spread in $I_{READ}$ ($I_{LEAK}$) compared with RD8T. It exhibits 1.42× larger $I_{READ}$ to $I_{LEAK}$ ratio. It shows 2.2× higher frequency at 250 mV with read bitline capacitance of 10 fF. Besides, the proposed bitcell achieves same read stability and write-ability as that of RD8T at the cost of 3 extra transistors. The leakage power of the proposed design is close to that of RD8T.

KEYWORDS: variability; robust, subthreshold; random dopant fluctuation (RDF); read static noise margin (RSNM); write static noise margin (WSNM).

1. INTRODUCTION

Due to severe increase in threshold voltage ($V_t$) fluctuation caused by global and local process variations in ultra-short-channel devices, 6T SRAM cell and its variants cannot be operated at further scaled supply voltages without parametric and functional failure causing yield loss. Single-ended 6T SRAM cell [1] suffers from
write delay. Low power 6T SRAM cell [2] could reduce access delay and write power but could not improve stability. Read-decoupled 8T (RD8T) proposed in [3-6], could improve RSNM (read static noise margin), but failed to improve variability significantly.

Battery operated mobile platforms such as PNA, PDA, cell phone, RFID tag, hearing aid, defibrillator, iPod, Smartcard, SmartPhone, SmartPen demand ultralow power SRAM to prolong battery life as SRAM is an integral part of all these platforms. Subthreshold operation holds promise for ultralow power operation for these emerging applications. However, the design of robust SRAM is challenging in subthreshold regime due to more severe mismatch in device characteristics induced by process variability compared with superthreshold region [7, 8].

To circumvent process variation, robust and variation tolerant SRAM cell design technique capable of absorbing $V_t$ shift due to random dopant fluctuation (RDF), and variation in other device and process parameters (such as length, width, oxide thickness, sub-wavelength-lithography, etching, etc.) and still be able to perform expected functions need to be investigated. Trade-off must generally be made in terms of area to achieve this goal. To solve process variation issues and minimum $V_{DD}$ operability in the face of extreme $V_t$ variation in ultrashort-channel devices at aggressively scaled technology node such as 16 nm, this paper proposes a transmission gate (TG)-based subthreshold 11T SRAM cell (hereafter called P11T) and compares its performance with standard read-decoupled 8-transistor SRAM cell (hereafter called RD8T).

The remainder of this paper is organized as follows. Section-2 presents the proposed design and its variability analysis. Simulation results are discussed and compared in Section-3. Section-4 concludes the paper.

2. PROPOSED TG-BASED 11T SRAM CELL

An attempt is made in this work to mitigate the impact of process, voltage and temperature (PVT) variations on design metrics of traditional RD8T SRAM cell by incurring minimum area penalty and retaining its dual port architecture.

2.1 Cell Sizing and Array Architecture

As fluctuation in $V_t$ due to RDF exhibits an inverse dependence on square root of device area [9], it is apparent that the simplest solution to the variability problem is to use larger device area. Therefore, the design metrics of the proposed design are compared with those of RD8T to demonstrate that the proposed design offers narrower spread in its design metrics even at iso-device area (device width over length (W/L) ratios are mentioned in Figs. 1 and 2). The array architecture of RD8T [5] is applicable to the proposed SRAM array organization as well. An array can be optimized by the usage of peripheral circuits designed in LSDL (limited switch dynamic logic) [10] and gated diode sense amplifier [11] as was done in RD8T [5]. The proposed design utilizes one read port and one write port and does not require much architectural changes except adding a PFET (MP3, MP4, and MP5) in parallel with each access NFET, thereby making it an 11T SRAM cell.
Two additional controls WWLB (write word line bar) and RWLB (read wordline bar) are required to switch the access PFETs. They are non-overlapping complementary signal of WWL (write wordline) and RWL (read wordline) respectively. Therefore, while accessing the cell, all relevant access FETs are switched simultaneously ON for reading or writing. During hold mode all access FETs remain OFF.

2.2 Current Analysis of Proposed Design

Read current (I_{\text{READ}}) can be expressed analytically by solving Kirchoff’s current law at node QM (with internal node “QB” storing “1”), where current flows out through MN6 and flows in through MN5 and MP5.

\[
I_{\text{READ}} = I_{\text{MN5}}(V_GS = V_{RWL} - V_{QM}, V_{DS} = V_{RBL} - V_{QM}) \\
+ I_{\text{MP5}}(V_GS = V_{RWLB} - V_{RBL}, V_{DS} = V_{QM} - V_{RBL}) \\
= I_{\text{MN6}}(V_{GS} = V_{QB}, V_{DS} = V_{QM}) .
\] (1)

Thus, I_{\text{READ}} is the sum of the currents I_{\text{MN5}} and I_{\text{MP5}} that passes through MN6 while discharging precharged RBL.

The write-ability of a one half cell is determined by applying complementary signal to BL and BLB. With BL = 1, I_{\text{WRITE}} is defined as a net current flowing out of internal node “QB” storing “1’
\[ I_{\text{WRITE}} = I_{MP5} (V_{GS} = V_Q - V_{DD}, V_{DS} = V_{QB} - V_{DD}) - I_{MN5} (V_{GS} = V_{WL}, V_{DS} = V_{QB}) - I_{MP5} (V_{GS} = V_{WLB}, V_{DS} = V_{QB}) - I_{MN1} (V_{GS} = V_Q, V_{DS} = V_{QB}) \] 

(2)

Since \( V_Q \) is usually small, the last term of \( I_{\text{WRITE}} \) is neglected. Thus, \( I_{\text{WRITE}} \) is the difference in current between pull-up device resisting the write and access devices discharging the node “QB”.

In the proposed P11T cell (Fig. 1), the read current \( (I_{\text{READ}}) \), defined as RBL discharge current, is the summation of two currents flowing through NMOS(MN5) and PMOS(MP5) transistors of a transmission gate (TG) respectively. This summation of read current \( I_{\text{READ}} \) over two different transistors in parallel is primarily responsible for the lower value of variability in case of P11T. This averaging is impossible in case of standard RD8T due to the existence of only one access transistor. It is clear that the total current will remain almost constant in case parameters of MN5 and MP5 vary in opposite direction due to summation. It is highly unlikely that parameters of both MN5 and MP5 of TG will vary in the same direction because they are of different types and MP5 exists in a different N-well. Moreover, it can be deduced that the current \( I_{\text{dn5}} \) (through MN5) varies to stabilize the total read current, in case current \( I_{\text{dp5}} \) (through MP5) varies due to RDF induced \( V_t \) shift and vice versa. This can be understood by assuming that only the threshold voltage of MN5 of TG goes down due to RDF whereas the threshold voltage of MP5 remains unchanged. This increases the current through MN5 and the total current \( I_{\text{READ}} \) through TG and hence \( V_{QM} \) rises quickly and \( V_{RBL} \) falls rapidly. This rapid rise of \( V_{QM} \) and fall of \( V_{RBL} \) increases threshold voltages of MN5 and MP5 due to body effect respectively, thereby, reducing the current through both MN5 and MP5. This helps in reducing (stabilizing) the total current \( I_{\text{READ}} \). Similarly, if the threshold voltage of MN5 increases due to RDF or any other reasons, the current through MN5 and also the total current \( I_{\text{READ}} \) decrease and hence, \( V_{QM} \) rises slowly and \( V_{RBL} \) reduces gradually. This results in lower value of threshold voltages for both MN5 and MP5 due to less pronounced body effect compared with the first case. The lower value of threshold voltage increases the current through both MN5 and MP5, thereby, stabilizing the total current \( I_{\text{READ}} \). The same explanation can be given in case the threshold voltage of MP5 varies due to RDF or any other reason. Therefore, it can be concluded that \( I_{\text{READ}} \) in TG is much more stable compared to \( I_{\text{READ}} \) in NMOS pass transistor. Similar explanation can be given for stability of \( I_{\text{WRITE}} \).

3. SIMULATION RESULT

This section presents various design metrics which are estimated during Monte Carlo (MC) simulations in HSPICE using 16 nm CMOS PTM [12].

3.1 Simulation Set-up

Threshold voltage \( (V_t) \) is assumed to have independent Gaussian distribution with \( 3\sigma \) variation of 58% [13]. Though, this ITRS anticipation of 58% \( V_t \) variation is quite aggressive, this value is selected to demonstrate that the proposed technique is still able to perform better than RD8T (however, 10% \( V_t \) variation is also used for certain simulation setup, which is mentioned in relevant figures). Authors in [14] showed that variation in \( C_{ox} \) (gate oxide capacitance) due to variation in \( L \) (channel length), \( W \) (width), and \( t_{ox} \) (oxide thickness) is negligibly small. It is to be noted that the effect of all parameter
variations can be translated into the effective variation in threshold voltage [15]. Therefore, only 58% $V_t$ variation is applied taking ITRS projection into consideration.

ITRS 2009 also anticipates temperature variation within the range from − 55 °C to 155 °C. Therefore, standby power dissipation is measured over this temperature range. As per ITRS 2009, expected variation in supply voltage is 10% in the future technology generations such as 16 nm [13]. Hence, most of the design metrics are estimated by scaling the supply voltage from 400 mV down to 250 mV and below.

The authors in [14] showed that a sample size of 2000 ensures a lower than 4% inaccuracy in the estimation of standard deviation. This work uses 5000 sample size while estimating design metrics with MC simulation [16]–[18] to achieve even higher accuracy. Moreover, read and write access failure probabilities are assessed by MC simulations with 25000 run [16]. Unless specified otherwise, all parameters are estimated under the above variation/simulation setup.

3.2 Read Current, Read Bitline Leakage Current and Their Variabilities

In subthreshold, the $I_{ON}/I_{OFF}$ ratio for NFET (PFET) in 16 nm technology becomes as low as 150 (170) @ 200 mV, which may be even lower in the face of process variation (see Fig. 3) (this nominal value of $I_{ON}$ to $I_{OFF}$ ratio is obtained by simulating an NFET and PFET (W/L = 32 nm/16 nm) in 16 nm technology). This may result in read failure and impose a theoretical upper bound on the number of cells sharing a common bitline. The devices in read buffer of both designs are therefore up-sized to mitigate these issues in subthreshold operation. The higher $I_{ON}$ to $I_{OFF}$ ratio of PFET compared to NFET leverage the $I_{READ}/I_{LEAK}$ ratio in the proposed design (see Fig. 3).

![](image.png)

**Fig. 3:** ON-current to OFF-current ratio of NFET and PFET (W/L = 32 nm/16 nm) in 16-nm technology.

$I_{READ}$ of both designs are estimated during simulation (see Fig. 4). RD8T shows higher $I_{READ}$ because the drive current of NFET is higher than PFET (~1.5× with W/L = 32 nm/ 16 nm). Therefore, though the width of MP5 (2×W) and MN5 (1.5×W) in P11T taken together is (3.5×W) equal to that of MN5 in RD8T, their drive strength is not equal. However, improved $I_{READ}$ variability is achieved in the proposed design due to the averaging effect of TG (see Fig. 5). This improved variability in $I_{READ}$ offers narrower spread in other design metrics of the proposed design.
Fig. 4: Read current of proposed bitcell (P11T) and standard RD8T.

Fig. 5: Read current variability versus $V_{DD}$. Proposed bitcell (P11T) offers tighter spread in $I_{READ}$ @ all considered supply voltages compared with RD8T.

The proposed design offers lower RBL (read bitline) leakage current $I_{LEAK}$ (see Fig. 6). It happens even though the pull-down stack (MN6) in read buffer has equal width (4xW) and the read buffer switch MN5 and MP5 forming TG in P11T taken together has equal width to that of MN5 in RD8T. This can be understood by conceiving the hot-carrier injection mechanism in short-channel devices. In a short-channel device, due to high electric field near the Si/SiO$_2$ interface, electrons or holes can gain sufficient energy from the electric field to cross the interface potential barrier and enter the oxide layer. This effect is so called hot-carrier injection. The injection of electrons from Si to SiO$_2$ is more likely than injection of holes, as electrons have lower effective mass than that of holes, and the barrier height for holes (4.5 eV) is more than that for electrons (3.1 eV) [19].

It can be observed that the long tail of the $I_{LEAK}$ distribution in case of RD8T extends to the right beyond that of P11T, which results in significant increase in $I_{LEAK}$ (see Fig. 7). $I_{LEAK}$ distribution curves of P11T and RD8T cross at 1.3 nA. Based on the simulation data, the estimation shows that 18% of statistical samples in case of P11T have $I_{LEAK}$ higher than 1.3 nA signifying higher column height because larger number of cells can be connected to RBL compared with RD8T (63% of statistical samples in case of RD8T have $I_{LEAK}$ higher than 1.3 nA).
Fig. 6: Read bitline leakage current versus $V_{\text{DD}}$. Proposed bitcell (P11T) offers lower $I_{\text{LEAK}}$ @ all considered supply voltages compared with RD8T.

Fig. 7: Impact of process variations on the bitline leakage current ($I_{\text{LEAK}}$) in 16 nm technology with $3\sigma V_t$ variation of 10% @ $V_{\text{DD}} = 0.25$ V.

Fig. 8: $I_{\text{LEAK}}$ variability versus $V_{\text{DD}}$. Proposed bitcell (P11T) offers narrower spread in $I_{\text{LEAK}}$ @ all considered supply voltages proving its robustness.

The proposed cell exhibits improved RBL leakage current variability compared to RD8T (see Fig. 8). It shows its robustness by offering 1.34× and 1.58× tighter spread in $I_{\text{READ}}$ and $I_{\text{LEAK}}$ respectively compared with RD8T @ 250 mV. Higher $I_{\text{READ}}$ to $I_{\text{LEAK}}$ ratio is highly expected in subthreshold SRAM design to achieve higher sense margin and column height. The proposed design exhibits higher $I_{\text{READ}}$ to $I_{\text{LEAK}}$ ratio even in the face of severe $V_t$ variability mentioned in the simulation setup. It exhibits 1.42× higher $I_{\text{READ}}$ to $I_{\text{LEAK}}$ ratio @ 250 mV implying its suitability in deep subthreshold operation (see Fig. 9).
3.3 Read Stability and Write-Ability

RD8T/P11T isolates the storage nodes Q and QB, resulting in RSNM free structure. And the operating margin, which is significantly improved, is set by the hold SNM and the write margin. The read stability of SRAM cell (RSNM) relies on $V_{DD}$. This is illustrated in Fig. 10, which shows that RSNM gradually decreases with $V_{DD}$ scaling. The proposed design shows three roots even at $V_{DD} = 250$ mV. This implies that the proposed bitcell can be operated in the subthreshold region at this voltage. RSNM VTCs of both cells, starting from 400 mV down to 250 mV have three distinct roots assuring its functionality as a bistable circuit in read and hold mode [8-9], [17-18], [20].

The write-ability of a cell is measured using read and write VTCs as shown in Fig. 11. The write VTC while writing “1” at QB is measured by sweeping $V_{QB}$ (see Fig. 11, y-axis) with BLB & WWL high and BL low while monitoring $V_Q$ (see Fig. 11, x-axis). This write
VTC is used in combination with the read VTC measured by sweeping $V_Q$ (see Fig. 11, x-axis) while monitoring $V_{Qb}$ (see Fig. 11, y-axis). The side length of the smallest square, that can be inscribed between the read and write VTCs of the same SRAM cell at the lower half of the curves, passed the trip voltage of INV1, represents WSNM [9], [17], [18], [21]. When WSNM falls below zero, write VTC intersects read VTC, indicating positive write margin and signifying write failure. WSNM of P11T and RD8T are illustrated in Fig. 11 @ 250 mV.

![Fig. 11: Static voltage characteristics of SRAM cells during write operation in 16 nm process technology @ supply voltage $V_{DD} = 250$ mV.](image)

**3.4 Read/Write Access Times and Their Variabilities**

The estimated result of $T_{RA}$ (read access time or read delay) and its distribution is plotted in Fig. 12. $T_{RA}$ at $V_{DD} = 250$ mV is also reported in Table I. Normalized values are reported in bracket. As observed from Table I, the proposed design (P11T) exhibits 2.14x improvement in read delay compared with RD8T @ 250 mV. It offers 2.35x improvement in $T_{RA}$ distribution @ 250 mV compared with RD8T, thereby proving its robustness against process variations. This improvement comes due to the use of TG in place of pass transistor in the read buffer.

![Fig. 12: Read delay and its variability versus $V_{DD}$.](image)
The estimated results of $T_{WA}$ (write access time or write delay) and its distribution are plotted in Fig. 13. $T_{WA}$ at $V_{DD} = 250$ mV are also reported in Table II. The values are normalized with respect to that of P11T and are reported in bracket. As observed from these simulation results, the proposed design offers $1.75\times$ improvement in $T_{WA}$ compared with RD8T @ 250 mV. It proves its robustness against process variations by featuring tighter spread ($2.79\times$) in $T_{WA}$ distribution compared with RD8T @ 250 mV. This improvement comes due to the PFETs employed in TGs.

![Fig. 13: Write delay and its variability versus $V_{DD}$](image)

### Table 1: Read Access Time and Its Spread @ VDD = 250 mV.

<table>
<thead>
<tr>
<th>SRAM</th>
<th>Std. Dev. (s)</th>
<th>Mean (s)</th>
<th>Std. Dev./Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>P11T</td>
<td>8.179e-07</td>
<td>1.894e-07 (1)</td>
<td>4.32 (1)</td>
</tr>
<tr>
<td>RD8T</td>
<td>4.101e-06</td>
<td>4.046e-07 (2.14)</td>
<td>10.14 (2.35)</td>
</tr>
</tbody>
</table>

### Table 2: Write Access Time and Its Spread at VDD = 250 mV.

<table>
<thead>
<tr>
<th>SRAM</th>
<th>Std. Dev. (s)</th>
<th>Mean (s)</th>
<th>Std. Dev./Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>P11T</td>
<td>1.186e-08</td>
<td>3.980e-09 (1)</td>
<td>2.98 (1)</td>
</tr>
<tr>
<td>RD8T</td>
<td>5.798e-08</td>
<td>6.981e-09 (1.75)</td>
<td>8.31 (2.79)</td>
</tr>
</tbody>
</table>

### 3.5 Leakage Power Estimation

The standby power consumption in an SRAM cell is more critical than dynamic power consumption, since the whole part of the cache remains idle most of the time except the row being accessed. Leakage power is estimated when the cell is in hold mode. It is observed, that the standby power of the proposed design is close to that of RD8T at all temperatures ranging from – 55 °C to 155 °C (plot not shown). And standby power increases exponentially as the temperature is increased due to the subthreshold device behaviour. Figure 14 shows leakage current ($I_{\text{Standby}}$) distribution of both designs in hold mode with $3\sigma$ threshold voltage ($V_t$) variation of 10% @ $V_{DD} = 0.25$ V.
3.6 Failure Analysis

Read access and write access failure probability analyses are carried out to assess the practicability of both cells to function under subthreshold conditions (see Fig. 15). The number of trials $N = 25000$ in MC simulations is set to achieve sufficient accuracy in the estimation of the access failure probability [16]. The probability of access failure is estimated as the ratio of the number of occurrence of read/write failure to the total number of experiments.

It can be observed from the figure that the probabilities of both the access failures are lower in case of P11T compared to RD8T. This indicates that there will be a least number of faulty cells in case of P11T-based SRAM offering maximum yield in the subthreshold regime compared with RD8T. RD8T fails to improve write-ability, though it offers good RSNM. In this analysis, it clearly exhibits its poor write-ability by showing higher write access failure probability although its write access NFETs are up-sized to match with TGs in P11T. On the other hand, the proposed P11T, offers lower value of read access and write access failure probability. This is attributed to the fact that TG employs a PFET that is less velocity saturated and relatively stronger in subthreshold region.

In particular, the proposed bitcell exhibits no read and write access failures even @ 205 mV, whereas RD8T shows read access failure of $0.2e-3$ and write access failure of $0.2e-1$ at the same supply voltage. Thus, the proposed bitcell will provide voltage and dimension scaling advantages. Low voltage operation will result in reduction of standby power (linearly, $I_{Standby} \times V_{DD}$) and dynamic power (quadratically, $C_LV_{DD}^2$) with reduced $V_{DD}$. One should note that the failure mechanisms illustrated above are function of environmental parameters. Hence the failure map of a cache will change if these parameters are altered.
Aminul Islam

3.7 Performance Analysis

Figure 16 shows the operating frequency versus CRBL plot @ $V_{DD} = 250$ mV. The proposed bitcell achieves 2.2× higher frequency @ 250 mV with read bitline capacitance of 10 fF (equivalently 229 cells connected to RBL, neglecting contact and wire capacitance). Higher frequency also implies that the column height using the proposed design will be higher. The shorter column height has negative impact on chip area. Fragmentation of column incurs extra silicon area because of supply power grid distribution in the chip. Thus, longer bitline is preferred to save chip area.

To observe the effect of varying CRBL on read delay variation, read delay variability versus CRBL is plotted in Fig. 17, which again shows that the proposed design is robust against process variation.

![Graph showing operating frequency versus read bitline capacitance at $V_{DD} = 250$ mV.](image)

Fig. 16: Operating frequency versus read bitline capacitance.
3.8 Area Comparison

The layout of both cells was drawn for fair area comparison (layout not shown). The proposed bitcell adds 20% area overhead compared to RD8T SRAM cell of [5]. Although, the proposed cell-size expands by 20%, its overwhelming improvement in variability justifies the additional overhead. Moreover, overall area penalty is less since more cells can be included on the bitline, which is evident from the previous performance analysis.

4. CONCLUSION

Due to aggressive scaling of device dimensions, variability has become a metric of equal importance as SNM (static noise margin) and WSNM (write static noise margin) in SRAM cell. This paper proposes a TG-based variability aware 11T SRAM cell. It analyzes the impact of PVT variations on its read access time, write access time, read current, bitline leakage current, and standby current. The results show significant improvement in most of the design parameters over standard RD8T SRAM cell demonstrating its robustness and functionality up to 250 mV and below. It mitigates impact of PVT variations mainly due to access TGs employed in the design. The proposed design is an attractive choice for application in scaled technology in subthreshold region in the presence of severe variation.

REFERENCES


