Functionality and stability analysis of a 400 mV quasi-static RAM (QSRAM) bitcell

Adam Teman\(^a,^b\), Anatoli Mordakhay\(^b\), Alexander Fish\(^a,^b\)

\(^a\) Ben-Gurion University of the Negev, Electrical and Computer Engineering, PO Box 653, 84105 Be’er Sheva, Israel
\(^b\) Ben-Gurion University of the Negev, Faculty of Engineering, Ramat Gan, Israel 52900

**A R T I C L E  I N F O**

Article history:
Received 20 September 2012
Received in revised form 10 December 2012
Accepted 11 December 2012
Available online 19 January 2013

Keywords:
CMOS memory integrated circuits
SRAM
Leakage suppression
Ultra low power
Sub-threshold SRAM

**A B S T R A C T**

The development of low-voltage SRAM bitcells with ultra-low static power consumption has become a primary focus of memory design in recent years. The analysis of these bitcells requires the evaluation of dynamic noise margin metrics in addition to the traditional static noise margins. In this paper, we extend the presentation of our recently proposed quasi-static RAM (QSRAM) cell that employs an aggressive internal feedback technique for leakage suppression. In addition to the presentation of the QSRAM circuit topology and operation, a broad stability analysis of the cell is introduced, proving the functionality and bi-stability of the bitcell. Many of the recently proposed dynamic stability metrics used in this analysis have been demonstrated on standard SRAM bitcells; however, this is one of the first times these metrics have been used to analyze the functionality of an alternative implementation. Functionality of the proposed bitcell is shown for a sub-threshold 400 mV supply voltage, providing a typical leakage reduction of 21X–45X as compared to a standard two-port bitcell operating at its nominal voltage. An 8 kb QSRAM array was implemented and fabricated in a commercial low-power 40 nm process demonstrating full functionality and ultra-low power consumption under a sub-threshold 400 mV supply.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

The continuous rise in leakage power of on-chip SRAM arrays is the main factor behind recent efforts to reduce the operating voltage of these memories. Lowering the supply voltage results in an aggressive reduction of both sub-threshold and gate leakage currents. However, lower supply voltages also cause the noise margins to decrease, leading to an inevitable degradation of robustness. Designing robust, low voltage circuits is further complicated by the large variations in circuit behavior that are caused by high variability in the fabrication process. Ratioed circuits, such as the standard six-transistor (6T) SRAM cell (Fig. 1a), are even more susceptible to these fluctuations, as device drive strengths can vary by as much as three orders of magnitude at sub-threshold voltages in modern nano-scaled processes [1]. Both theoretical and empirical analyses have shown that 6T SRAM cells fabricated in sub-90 nm technologies are limited to super-threshold minimum operating voltages ($V_{DDmin}$) [2,3]. These limitations occur during read and write operations, when the drive strength ratios determine circuit functionality. Accordingly, many research groups have shifted their focus to the development of alternative bitcell topologies and peripheral circuit techniques to enable the operation of on-chip SRAM arrays at ultra-low operating voltages, deep into the sub-threshold region [3–15,36].

The traditional method for measuring the stability of an SRAM cell is with the well-known static noise margin (SNM) metric [16]. This metric ensures data retention in the presence of a pair of serial voltage noise sources at the bitcell’s internal data nodes. These sources are applied with opposite polarities to represent a worst case situation, as shown in Fig. 1b. Write and read stability are tested in a similar fashion, measuring the ability of the bitcell to change or retain its state, respectively, following an infinitely long access pulse. In older process technologies, obtaining sufficient margins was achieved solely based on device sizing, as $I_{read}/I_{write}$ ratios were large and process variations were limited. However, it is increasingly more difficult to meet stability requirements at scaled down processes, due to the increased device fluctuations that are inherent to these technologies. In addition, the static write margin (WM) criterion has been found to be overly optimistic, due to the finite duration of write access operations, as defined by the system’s operating frequency. Device variations are further emphasized at low operating voltages, since the drive strengths of the devices are more severely affected by threshold voltage ($V_t$) fluctuations [3]. Accordingly, recent years have seen increased research into the analysis of the dynamic operation and stability of SRAM cells [17–23,36].
Contra...
feedback SRAM (SF-SRAM) cell [7], but its functionality is very different, as will become apparent in the following sections.

In addition to the low-threshold (LVT) feedback device (M9), the cell core comprises a pair of high-threshold (HVT) nMOS pull-down devices (M1, M4); a pair of standard-threshold (SVT) pMOS pull up devices (M3, M6); and a pair of SVT nMOS access transistors (M2, M5). An additional two nMOS devices (M7, M8) create a single-ended read buffer, similar to that employed by the dual-port 8T bitcell. The gate of M7 is driven by the inverted data node (QB) and is completely independent of the non-inverted data node (Q); this is an essential observation to understand the functionality of the QSRAM bitcell. Since the read-out is done exclusively according to the level stored at QB, the level at Q has no influence on read functionality.

It should be noted here that the QSRAM cell, as presented in this manuscript, operates with standard peripheral circuits as required for the operation of a dual-port 8T bitcell. These include a pair of word lines for write and read operations (WWL and RWL); a pair of bitlines for write operations (BL and BLB); and a single-ended bitline for read operations (RBL). The simulations and measurements presented hereafter were achieved without the employment of non-standard techniques, such as overdrive or negative biasing. However, these methods, as well as other previously proposed techniques, could easily be incorporated with this topology for improved performance.

2.2. The hold ‘1’ state

Following a write ‘1’ operation (as will be described later), Q is charged to a high level, while QB is fully discharged to ground. The biases applied to the bitcell at this state and the resulting leakage currents are illustrated in Fig. 3a. The gate bias (V_GS) of M4 is initially much higher than that of M1; M4 is conducting, while M1 is firmly cut-off, ensuring that QB stays discharged. The discharged gate bias of M3 (V_GS = V_DQ = 0) causes a charge share between Q and V_DQ, resulting in a high level at V_DQ. At this point, the feedback of QB back to M9 creates a negative V_GS X (V_GS = V_DQ < 0), completely cutting off the power supply to the cell, and essentially floating the high data node (Q). The stable state of Q is set according to the leakage currents through the left side cut-off devices (M2, M9, and M1—the leakage through M6 is negligible); the aforementioned device threshold implants ensure that a discharged voltage at QB is maintained, providing the required gate voltage at M7 for a read ‘1’ operation. This is shown through Monte Carlo simulation of the steady state voltage at node QB, as plotted in Fig. 3b. This figure clearly shows that QB is fully discharged throughout the distribution, providing a robust readout under process variations and device mismatch.

2.3. The hold ‘0’ state

Contrary to many of the other SRAM bitcell implementations, the QSRAM is an asymmetric circuit, requiring a separate discussion for each of its stable states. An initial observation may lead to the assumption that the hold ‘0’ state is quite similar to the corresponding state of the 6T cell, as M9 is biased with a high voltage and therefore conducting. A secondary observation could make one believe that the cell is completely non-functional; assuming that passing the supply through M9 leads to a threshold drop, then this lower voltage is subsequently passed through M6 to QB. This would lead to an additional threshold drop, and so on, until QB would eventually fully discharge. However, an even closer look shows that neither of these observations is correct.

Starting with a write ‘0’ operation, Q is completely discharged, while QB is charged to a high level\(^1\). At this point, assuming V_DQD is at some median level, M3 is cut-off (V_GS = V_GB = V_DQD > 0) and the source of M6 is QB, such that QB charges V_DQ through M6. Ultimately, V_DQ = V_GB; and therefore V_GB = 0, cutting off M9. The voltage of QB following a write ‘0’ operation is, therefore, slightly lower than V_DQ. With the access transistors (M2 and M5) closed, the leakage ratio (I_D + I_D)/4 sets the final state of this node. Sizing and V_T implants are utilized to ensure that this level is as close as possible to V_DQ to ensure a correct readout according to the QB voltage sampled by M7 (V_GSS = V_GB). The Monte Carlo distribution of the steady state voltage at node QB is plotted for this state in Fig. 4b. The figure clearly shows that QB retains a high level throughout the distribution, providing a robust readout under process variations and device mismatch. Note that corners that cause a drop in the stable state voltage of QB will incur a slightly slower readout, as the overdrive voltage of M7 will be lower. However, this penalty is dampened at low voltages, as the current through M7 and M8 must

---

\(^1\) For super-threshold operation, we would claim that this would be a threshold drop under V_GS; however, the main operating region of this cell is in the sub-threshold region, where the drop off is not well defined, but is not negligible.
be equalized, trading off the values of overdrive and $V_{DS}$ as a function of the transient voltage at the node connecting the two.

2.4. Write operations

The QSRAM bitcell was designed to minimize leakage; however, as a positive side-effect, it also demonstrates a very robust write operation. Considering that write margin is the voltage limiting level, this presents quite an advantage during low-voltage operation. Considering that write margin is the voltage limiting as a positive side-effect, it also demonstrates a very robust write operation. Therefore, in order to lower $V_{DD}$, many of the low voltage SRAMs incorporate specialized peripheral write-assist techniques that either strengthen the pull-down devices or weaken the pull-up ones. Weakening the pull-up is an inherent property of the QSRAM design, as the feedback device (M9) adds additional serial resistance to the supply path. In fact, M9 is cut-off during both hold states, presenting virtually no contention to the bitline that is trying to pull down its adjacent node. Accordingly, the QSRAM retains write-ability at sub-threshold voltages across process corners and under local mismatch without the need for specialized write-assist circuits.

A closer look into the QSRAM write operations will show a slight difference between the write ‘0’ and write ‘1’ actions. Starting with the write ‘0’ operation, the non-inverted bitline (BL) is discharged to ground, while the inverted bitline (BLB) is charged to $V_{DD}$. Subsequently, the write word line (WWL) is asserted, initializing the write operation. Assuming that the accessed bitcell was holding a ‘1’, the voltages at node Q and $V_{DD}$ are at a median level, $V_{Q}$, and M9 is strongly cut-off with a negative overdrive voltage ($V_{GSO} = V_{Q} - V_{Q} = -V_{Q} < 0$). At this point, M2 is turned on with a strong overdrive ($V_{GSO} = V_{DD}$), easily discharging $Q$, as it needs only to overcome the very low leakage through the serial connection of M9 and M3. Even at the problematic slow $nMOS-fast pMOS$ (SF) corner, the operation is easily achieved, as both drive networks include nMOS devices.

Whereas for most SRAM designs, observing a successful discharge of one of the nodes is enough to ensure a write operation, the QSRAM doesn’t incur the same positive feedback; therefore, the behavior of the other node must be considered, as well. In this case, QB is charged through M5 with a weak overdrive voltage ($V_{GSO} = V_{Q} - V_{Q} < 0$). This is achieved without the contention of a pull-down node, as once Q is discharged, M4 is cut-off. At this point, M6 is conducting ($V_{GSO} = V_{Q} - V_{Q} < 0$), so $V_{DD}$ is charged up to $V_{Q}$, as defined by the uncontended charging operation through M5. At the end of the write operation, the final state of $V_{Q}$ is slightly lower than $V_{DD}$ and can fluctuate according to the leakage ratios of the cell’s pull-up and pull-down networks. This fluctuation can be seen in the distributions of Fig. 4b, above.

Continuing on to the Write ‘1’ operation, the biases of the bitlines are inverted ($BL = V_{DD}, BLB = 0$) and M5 is strongly over-driven to discharge QB. Assuming Q was discharged and QB was high, $V_{GSO} = V_{Q}$. As previously described, then M9 is cut-off with $V_{GSO} = 0$. Therefore, as in the write ‘0’ operation, the access transistor meets very little contention from the pull-up network (made up of M9 and M6). Moreover, as $V_{Q}$ drops, the overdrive of M9 becomes (at least temporarily) negative, further weakening the pull-up current. For this operation, observing the QB side is almost sufficient, as this is the side that controls the cell’s readout value; all that is required from the Q node is that it is charged higher than QB, which is easily accomplished through M2.
The final voltage at Q is lower than \( V_{DD} \) in accordance with the leakage ratios; this voltage is generally between 40 and 60% of \( V_{DD} \), depending on the process corner and mismatch parameters.

2.5. Read operation

As described above, read operation of the QSRAM cell is achieved in a similar fashion to the 8T read. Fig. 3b and Fig. 4b show the steady state voltage distributions at the QB node, through which the read operation is sensed. In our implementation, we chose a 256 row array with a divided bit-line, such that every 64 bits in a column are connected to the same RBL. In order to correctly read out of the cell, the frequency must be set between the worst case time it takes to read out a ‘0’ (i.e., QB is low with all other cells on the same RBL holding a ‘1’) and the shortest time it takes for leakage currents to discharge the RBL while reading a ‘1’ (i.e., QB is low with all other cells on the same RBL holding a ‘0’). Fig. 5 shows the distribution of the RBL discharge time for 5k Monte Carlo samples of cells holding a ‘1’ and a ‘0’. The plot shows that the time it takes to discharge the RBL due to leakage is substantially higher than the time it takes to discharge the bitline through M7 and M8. Note that the time axis is logarithmic, such that the gap between the two distributions is sufficient for read sensing.

2.6. Leakage power

The hold states of the QSRAM cell display a very non-conventional approach to SRAM design. Rather than having a fully static cross-coupled feedback structure, the stable state is set according to initial conditions and held by leakage currents. While this does make the bitcell more susceptible to injected noise, it retains functionality and provides enhanced leakage suppression over the standard topologies.

The majority of the leakage in a standard 6T SRAM cell is set by sub-threshold leakage through cut-off devices with \( V_{GS} = 0 \) and \( V_{DS} = V_{DD} \). Assuming the bitlines are precharged to \( V_{DD} \), the leakage current of the 6T cell in the hold ‘1’ state (as depicted in Fig. 1a) can be estimated as the sum of the currents through M1, M3, and M6. Neglecting gate leakages, which will be referred to separately, the leakage can be written according to the EKV based sub-threshold current model\(^{28}\) as\(^{2} \):

\[
I_{LT} = I_{0} + I_{S} = 2 \times I_{th} e^{\frac{V_{DD} - \Delta V}{V_{n} \mu_0 n_{thr} r_{P}}} + I_{th} e^{\frac{V_{DD} - \Delta V}{V_{n} \mu_0 n_{thr} r_{P}}},
\]

where \( n \) and \( p \) indices indicate nMOS and pMOS devices, respectively; \( I_{th} \) is the sub-threshold current coefficient; \( V_{T} \) is the device threshold voltage; \( \eta \) is the drain-induced barrier lowering (DIBL) coefficient; \( n \) is the sub-threshold swing coefficient; and \( r_{P} \) is the thermal voltage.

For the QSRAM in the hold ‘1’ state, the sub-threshold leakage can be estimated as the sum of the currents through M9 and the access transistors, M2 and M5\(^{5} \). Assuming that the steady state voltage of Q is a median voltage \( V_{QQ} \), and that \( V_{DD} \) equals to \( V_{Q} \) through M3, we can write:

\[
I_{LT, \text{hold1}} = I_{0} + I_{S} = \frac{I_{th} e^{\frac{V_{DD} - \Delta V}{V_{n} \mu_0 n_{thr} r_{P}}}}{2} + I_{th} e^{\frac{V_{DD} - \Delta V}{V_{n} \mu_0 n_{thr} r_{P}}},
\]

where the \( V_{lvt} \) index indicates a low-threshold nMOS device. In comparison to (1), the exponentials in the currents of M9 and M2 have additional negative values \( V_{Q} \) and \( \eta V_{Q} \) that make the QSRAM leakage substantially lower than that of the 6T bitcell. This is due to the negative \( V_{CS} \) of the feedback device and the lower DIBL current through both devices. If gate leakages are taken into account, a small additional reduction is achieved. The 6T presents two primary gate leakages (through M3 and M4 in the hold ‘1’ state), while the QSRAM in the hold ‘1’ state has three contributors (M3, M4 and M6); however, the gate leakage is exponentially dependent on the potential that falls across the gate oxide, such that the sum is lower for the QSRAM cell.

A similar observation can be applied to the QSRAM in the hold ‘0’ state as shown in Fig. 4a. In this case, the current through M5 is negligible, so the total sub-threshold leakage can be estimated as the sum of the currents through M9 and M2. Assuming that Q is fully discharged and the steady state voltage of QB is \( V_{QQ} \) (which is also the bias at \( V_{DD} \)) we can write:

\[
I_{LT, \text{hold0}} = I_{0} + I_{S} = I_{th} e^{\frac{V_{DD} - \Delta V_{QQ}}{V_{n} \mu_0 n_{thr} r_{P}}} \left( 1 - e^{\frac{V_{DD} - \Delta V_{QQ}}{V_{n} \mu_0 n_{thr} r_{P}}} \right) + I_{th} e^{\frac{V_{DD} - \Delta V_{QQ}}{V_{n} \mu_0 n_{thr} r_{P}}}
\]

(3)

The current described by (3) is usually slightly larger than that of (2), as the exponent of \( I_{th} \) is more negative; but this depends on the ratio between \( V_{Q} \) and \( V_{QQ} \). In fact, the higher the steady state of \( V_{QQ} \) is, the lower the leakage in the hold ‘0’ state. However, this increases the gate leakage through M9 and M6 in the hold ‘0’ state, and ultimately, these factors can trade off. In particular, this is the case for low-K processes, like the 40 nm LP technology that we used.

To summarize the leakage current dissipation, Fig. 6 shows the distribution of the leakages of the two states of the QSRAM as compared to a standard 8T bitcell at 400 mV. The figure clearly shows the significant leakage suppression of the QSRAM cell in the hold ‘1’ state, with an average reduction of 3.7X over the standard 8T cell operated at the same voltage. The hold ‘0’ state also provides a vast improvement with an average reduction of 1.8X. It should be noted that this reduction is on top of the exponential reduction achieved by operating the array at such a low voltage, under which the standard implementations are non-functional. When compared to an 8T bitcell at its nominal operating voltage (1.1 V), the average leakage reduction is 45X and 21X for the hold ‘1’ and hold ‘0’ states, respectively.

\[\text{Fig. 6. Monte Carlo statistical distribution of the ratio of leakage current between a standard 8T SRAM bitcell and the proposed 9T QSRAM bitcell. The figure includes ratioed plots for both the hold ‘1’ and hold ‘0’ stable states (with } V_{DD} = 400 \text{ mV).}\]
3. Stability analysis of the QSRAM bitcell

Section 2 presented the QSRAM bitcell and its advantages; it provides low leakage and retains write-ability at low voltages with standard peripheral circuitry. However, any experienced SRAM designer will immediately contemplate the stability of this non-conventional approach. In this section, we will provide a broad stability analysis of the QSRAM bitcell, through a large array of techniques and metrics.

3.1. Standard static noise margin metrics

The de-facto standard for measuring stability analysis is the static noise margin, originally proposed by Hill [29]. This method plots the voltage transfer characteristics (VTC) of the inverters that comprise the bitcell core, finding the largest static disturbance that the cell can withstand without losing its state. Seevinck et al. [16] proposed a method for efficiently measuring this metric using a single DC sweep simulation. For the hold state, the SNM metric assumes an infinite, serial voltage noise source (Fig. 1b), while for read and write operations, it assumes an infinitely long access pulse. These assumptions lead to a pessimistic evaluation of read and hold operations, as the duration of the actual noise is finite. In addition, a serial voltage noise source is non-physical and ignores the temporal pattern of injected noise and circuit dynamics [21,30]. For older technologies, this pessimism was reasonable, as process variations were not as severe, and the resulting bitcell sizes scaled with Moore’s Law [31]. However, for nano-scaled technologies with extreme process variations and local device mismatch, meeting large SNM requirements can lead to overdesign, and therefore, impede array size scaling. On the other hand, when discussing write operations, the SNM metric is an overly optimistic evaluation. Since the actual access pulse is finite, the cell won’t reach its intended state if the pulse is too short. Therefore, by exclusively measuring the SNM metric for write operations, potential write failures could be overlooked.

To conclude, it has been shown that for read and hold operations, the SNM metric is a sufficient, but not a necessary condition for stability, whereas for write operations, it is insufficient [21].

Measuring the SNM metrics of the QSRAM cell presents an additional aspect. This cell is quasi-static, rather than fully static, as its steady state is set by leakage ratios, rather than strong positive feedback. This leads to rather depleted SNMs for hold (and read, which is equivalent to hold for bitcells employing a non-penetrative readout buffer [3]). However, as we will show in the following subsections, the dynamic stability of the QSRAM cell is higher than would be expected based on its SNM. But first, as the initial stage of our stability analysis, we will present the SNM of the QSRAM cell.

Fig. 7a shows the “butterfly curves” [16] of the QSRAM cell for hold/read under typical conditions with both a 1.1 V and a 400 mV supply. The DC transfer functions from both Q→QB and QB→Q provide the expected inverting function; albeit, with a relatively low switching threshold, as the conducting nMOS has less contention when pulling down against the gated supply. A large difference is displayed between the SNM of the hold ‘0’ vs. hold ‘1’ states, as is expected due to the asymmetric nature of the circuit. One of the interesting aspects of these plots is that despite the extreme reduction of supply voltage (from 1.1 V to 400 mV), the SNM reduction is relatively small. In fact, for a 63% decrease in supply voltage, the SNM is only about 25% lower. This is another justification for designating this topology for low-voltage operation.

Fig. 7b shows the Monte Carlo statistical distribution of the SNM for 5k points at VDD=400 mV. The median value of the hold ‘0’ SNM is in the order of 100 mV, whereas for hold ‘1’ it is in the order of 70 mV. These values are sufficient, yet the standard variations are quite large (23 mV for hold ‘0’ and 27 mV for hold ‘1’), such that the 6σ distribution would seem to fail. However,
the implications of this are much less significant under a time limited noise source. To ensure a correct read, it is essential that Q8 remains low, and since M9 gates the supply and dampens the positive feedback that causes the cell to flip, a noise source with a limited duration would usually be discharged rather than flip the cell. This will be shown in the following subsections.

The SNM for write operations is very robust, as expected, considering the descriptions given in Section 2. Fig. 8 shows the write margin for the write ’0’ and write ‘1’ operations as compared to a standard 8T cell with a 400 mV supply. The margins were measured using a modified version of the bitline sweep method [25]. In the standard method, a voltage noise source is only applied to the ’0’ bitline, as the standard bitcell is designed to be insensitive to the ‘1’ bitline (in order to maintain the read constraint). However, for an asymmetric cell, and for the QSRAM in particular, this argument does not hold up, and so a noise source was added to both bitlines. Note that this methodology should also be used when measuring an 8T cell sized without consideration of half-select scenarios, such as the reference cell used here, as the influence of the ‘1’ bitline should be taken into consideration. Fig. 8 shows the margins across standard process corners. The QSRAM cell displays substantially higher margins at all corners, other than the best-case fast nMOS—slow pMOS (FS) corner, where the margins are comparable. At this corner, the access transistors of the 8T cell are much stronger than the pull-up pMOS, ensuring a successful write. On the other hand, the 8T loses write-ability at the SF corner, whereas the QSRAM maintains robust write operations under these conditions. Here, the internal feedback weakens the pull-up, which for the 8T cell is stronger than the pull-down through the access transistors, impeding the write margins.

3.2. State space analysis

Several recent works have shown that the standard 6T SRAM bitcell is a non-linear time-variant system, and have developed models to describe the system behavior using state-space analysis [17,18,21,22,32,37]. A similar model can be developed for the QSRAM cell; however, this topology displays a third internal node \( V_{VDD} \), in addition to the standard data nodes \( Q \) and \( QB \). The cell’s transient behavior can be modeled as a system of seven voltage-controlled-current sources and three capacitors (the total lumped capacitance to ground at the three data nodes), as shown in Fig. 9. This model provides the following set of equations:

\[
\begin{align*}
\frac{dV_Q}{dt} &= -I_1(V_Q, V_{QB}) + I_2(V_Q, V_{VDD}) + I_3(V_Q, V_{QB}, V_{VDD}) \\
\frac{dV_{QB}}{dt} &= -I_4(V_{QB}, V_Q) + I_5(V_Q, V_{VDD}) + I_6(V_{QB}, V_Q, V_{VDD}) \\
\frac{dV_{VDD}}{dt} &= I_8(V_{VDD}) - I_9(V_Q, V_{QB}, V_{VDD}) - I_{10}(V_{QB}, V_Q, V_{VDD})
\end{align*}
\]

where \( I_n \) indicates the voltage dependent current through device \( n \) of Fig. 2; \( C_{node} \) indicates the lumped capacitance to ground at nodes \( Q \), \( QB \) and \( V_{VDD} \); and \( V_{node} \) indicates the transient voltages at nodes \( Q \), \( QB \) and \( V_{VDD} \). This system comprises a three-dimensional state vector, \( V \), made up of the node voltages of the three internal nodes \( V_Q, V_{QB}, V_{VDD} \). Interestingly, the circuit settles at two stable points, similar to a standard SRAM cell, with \( V_{VDD} \) approximately equal to the higher of the two data nodes. This simplifies the dynamic analysis of the QSRAM topology, as the cell state can be represented on two-dimensional state plots of \( V_Q \) and \( V_{QB} \), as shown in previous works [21,22,30,32,37] for the standard 6T cell.

Analytically solving the dynamic system of (4) is out of the scope of this manuscript, especially in light of the non-linear multiple parameter dependence of each of the device currents. However, it is clear that state space analysis based on simulation data is necessary to understand the dynamics and behavior of the cell. In the following subsections, we will present the state space map, separatrix, phase portrait, and internal loop gain of the QSRAM bitcell. Note that as far as we know, this is one of the first times such an analysis has been presented for a non-standard (6T or 8T) bitcell.

3.3. Separatrix

To consider dynamic stability, it is essential to consider how the state of a bi-stable circuit transients from one stable equilibrium to the other (i.e., a state flip) under a single event upset (SEU), a read, or a write event [22]. The state space of such a circuit is divided into two stability regions by a stable manifold, better known as the circuit’s “separatrix” [33]. Following an event (e.g., SEU, read, or write access), the circuit will stabilize at one of its two stable states, depending on which side of the separatrix it is at, once the event is over. The separatrices of the QSRAM cell at standard process corners are mapped in Fig. 10, providing an...
interesting insight into the QSRAM cell\(^5\). The figure shows that the separatrix is quite different than that of a standard cross-coupled latch structure, where the manifold is roughly set by the line where \(Q=QB\). For the QSRAM, the region of attraction (ROA) of the hold \('1'\) state is larger than that of the hold \('0'\) state, as the separatrix is slightly above the \(Q=QB\) line; this enables the voltage level of \(Q\) to vary as long as \(QB\) is low. The \('1'\) equilibrium is at the state vector of \(V(V_{\text{G}}V_{\text{QB}}V_{\text{DD}})=(0.185\ \text{V},\ 3\ \text{mV},\ 0.186\ \text{V})\) providing a dynamic margin from the separatrix during either a noise event that discharges \(Q\) or one that charges \(QB\). The \('0'\) equilibrium, on the other hand, is at \(V=(0\ \text{V},\ 0.382\ \text{V},\ 0.383\ \text{V})\)—close to the standard SRAM equilibrium. This equilibrium provides a dynamic margin from the separatrix during either a discharge of \(QB\) or a charge of \(Q\).

In addition to the separatrices of the QSRAM cell, Fig. 10 also includes the cell's stable states at standard process corners, as well as a scatter of 5k Monte Carlo samples of these states. As the separatrix is relatively constant under process variations, the scatter plot provides additional insight to the stability of the cell. All samples distinctively fall onto one side or the other of the separatrix, providing bi-stable ROAs for the circuit. The plot also shows the worst case SF corner with a depleted \(QB\) level, as shown in Fig. 11. The cell's two ROAs and the separatrix are easy to observe from within the plot. Any temporal state above the separatrix has a high \(QB\) voltage, causing \(Q\) to quickly discharge, as shown by the horizontal left-pointing vectors. The state will subsequently progress to the \(V_0\) stable point, as the leakage current replenish \(QB\), while keeping \(Q\) low. The opposite attraction occurs under the separatrix; these states have a high \(Q\) value, providing a strong \(QB\) pull-down through M4. The result is the down-pointing vertical vectors showing a quick discharge of \(QB\). Subsequently, \(Q\) will slowly discharge to its equilibrium state of \(V_1\). For additional clarity, the circuit's two equilibriums \((V_0\) and \(V_1\)) are clearly marked on the figure.

In addition to the hold phase portrait, similar plots can be extracted for the asymmetrical write '0' and write '1' operations, as shown in Fig. 12a and b, respectively. Here we can see that each operation is attracted to a single stable state. The behavior of the write operations can also be observed, corresponding with the descriptions given in Section 2. For the write '0' operation, the \(Q\) node is strongly discharged, with the approach to the high \(QB\) state much less aggressive. For the write '1' operation, \(QB\) is discharged much more abruptly than \(Q\) is charged. Both plots show a single stable equilibrium (clearly marked on the figures), as required by a write operation. Note that the stable points are slightly different than those shown in Fig. 11, as the plots are extracted with the access transistors asserted. The cell will eventually stabilize at its equilibriums following the discharge of WWL.

### 3.5. Loop gain

An additional observation of a circuit's stability should be verified using a bi-stable circuit's small signal loop gain \([21]\). A loop gain plot can be derived by performing small signal analysis at every arbitrary DC operating point across the state space. For stability, the loop gain plot must show gain lower than unity in the areas of the stable states. This corresponds to the circuit's ability to recover from a transient interference. If the gain is larger than 0 dB, the interference will amplify, resulting in loss of state (i.e., cell flip). If the loop gain is less than unity, the state will return to the minimum gain point. The QSRAM cell's loop gain, shown in Fig. 13, is very different than the saddle shape of the standard 6T bit cell's gain. The unity gain level is emphasized in this figure by the gray plane that dissects the plot and the local minima are marked. For clarity, the topographical contours of the 3-D plot are also shown. It is clear that the two minima are well below unity and reside in the immediate area of the cell's equilibriums, verifying the cell's stability. The maximum gain of the cell is achieved in the area of the meta-stable point, corresponding with the high gain of both the feed-forward and feedback networks, as can be seen in Fig. 7a. Additional areas with less than unity gain correspond to the relatively large recovery times under the occurrence of a disturbance that would skew the space to such a region.

### 3.6. Dynamic noise margin

Combining the observations of the preceding sub-sections, we can now discuss the dynamic noise margins of the QSRAM bitcell. A concrete DNM metric was first proposed by Dong et al., \([22]\) as the difference between the access time in read and/or write \((T_{R} \text{ and } T_{W}, \text{ respectively})\) and the time it takes the cell to cross the separatrix during these events \((T_{\text{across,R}} \text{ and } T_{\text{across,W}}, \text{ respectively})\). Specifically, the DNM for read (RDNM) and write (WDNM) are

---

\(^{5}\) Note that due to the abundance of data on the figure, and since the separatrices for all corners are very close to each other, the key attaching each separatrix to its relevant corner was omitted.

\(^{6}\) Note that this cell includes three internal nodes, rendering a three-dimensional state space analysis. However, since \(V_{\text{DD}}\) is connected to either \(Q\) or \(QB\) through a conducting pMOS device at almost every state, the addition of a third axis to these plots provides very little additional insight at the cost of visual complexity and therefore the third state variable was omitted from these plots.
given as:

\[
R_{DNM} = T_{across,R} - T_R \\
W_{DNM} = T_W - T_{across,W}
\]

For cells displaying a non-penetrative read operation, such as the QSRAM cell, RDNM is infinite, as the core of the bitcell is not disturbed. Dynamic stability during a read operation should therefore be considered according to the cell’s dynamic hold margin, as discussed below. WDNM is separate for the write ‘1’ and write ‘0’ operations, and due to its dependence on cell access time \(T_W\) which is controllable, it is more informative to measure the write operations’ \(T_{across,W}\) characteristics. At 400 mV, both \(T_{across,W1}\) and \(T_{across,W0}\) were found to be 0.6 ns at nominal conditions. This can be compared to the \(T_{across,W}\) time of a standard 6T cell, which is 1.4 ns at 400 mV. For the worst case SF process corner, \(T_{across,W1}\) is 0.9 ns and \(T_{across,W0}\) is a much slower 3.2 ns, but this is compared to the infinite \(T_{across,W}\) of the 6T cell, as the cell is non-writeable (WSNM \(< 0\)) at this corner. This concurs with the discussion of Section 2.

The DNM for hold was only briefly introduced in [22], but was later extended by Zhang et al. [24], [34]. During hold, two transient noise models are discussed: the square pulse model, and the exponential model. In the square pulse model, a current noise source is injected into the low data node of a 6T SRAM cell. For a current amplitude of \(I_n\), the required pulse width to drive the state across the separatrix is measured \(T_{crit}\). Since the square pulse model presents a very simplified representation of an actual noise event, the exponential model is proposed to more accurately model an SEU caused by radiation [35]. A third model is achieved by applying the noise current infinitely, and thus measuring the static current noise margin \(ISNM\), which is an alternative to the standard SNM metric.

Due to the symmetry of the 6T bitcell, it is sufficient to solely discuss the current noise injection at the low data node to develop an analytical model, as presented in [17]. However, the QSRAM cell is extremely asymmetric, such that hold DNM must be measured for all cases. Fig. 14 plots \(T_{crit}\) as a function of \(I_n\) for the four cases of a square noise pulse. These cases are: a charging current to QB or a discharging current from Q for the hold ‘1’ state; and a charging current to Q or a discharging current from QB for the hold ‘0’ state. The ISNM is extremely low for this cell.
(approximately 10 pA – shown by the vertical asymptotes of Fig. 14), as the cell is quasi-static and held at its stable state by leakage currents. However, as shown in the figure, the dynamic hold stability of the QSRAM cell is much more robust than the ISNM predictions. The cell is most sensitive in the hold ‘1’ state with noise charging the QB node; however, these are still worst case predictions, as generally a charging pulse to one node will also similarly affect the inverted node, maintaining some (or all) of the difference between them. In order to further test the state stability under a typical cross-talk disturbance, all four scenarios were simulated with a 0 to VDD pulse through the extracted coupling capacitances, and the cell returned to its initial state following each of these interferences.

A further observation of the cell’s behavior during a square pulse noise event is shown in Fig. 15a. This figure plots the trajectory of two noise events in the hold ‘0’ state. The first noise pulse is with T_{noise} > T_{crit}, resulting in a destructive state flip, whereas the second is with T_{noise} < T_{crit} resulting in full state recovery. A similar plot is shown for a write ‘0’ event in Fig. 15b. In this case, when T_{W} > T_{cross,W}, the write operation is successful and the cell reaches the ‘0’ state; however, when T_{W} < T_{cross,W} the cell state never crosses the separatrix and eventually returns to the ‘1’ state.

4. Implementation and measurements

4.1. Bitcell implementation and layout

The proposed 9T QSRAM bitcell was implemented in a commercial 40 nm low-power TSMC technology. Layout of the bitcell was done exclusively with standard process layers, including multiple V_T implants. The chosen layout, shown in Fig. 16, maintains unidirectional polysilicon gates, while reducing area overhead by implementing vertical wells and sharing bus contacts between adjacent bitcells. This implementation includes LVT implants for M9 and M7, as well as HVT implants for M1 and M4. Longer than minimal lengths were used for several of the devices in order to utilize the reverse short channel effect (RSCE) [27] and improve immunity to process variations. The layout was carried out according to standard logic design rules, resulting in a total area 1.044 μm² per bitcell. This is approximately 10% larger than an 8T 2-port cell abiding to the same rule set, but about 2.8X larger than the pushed rule single-port 6T bitcell provided by the
foundry. This ratio could be significantly reduced by employment of pushed rules for this circuit, as well.

4.2. Test chip implementation and architecture

An 8 kbit array of QSRAM bitcells was assembled and fabricated as part of a 40 nm test chip along with other test structures. The array utilized a divided bit-line architecture, separating the readout path into four 64-row sub-sections to enable the integration of a 256-row block. In order to isolate the power consumption of the SRAM core, separate power supplies were used for the periphery and the array. The peripheral circuits were powered with a nominal supply voltage and level shifters were employed to propagate the 400 mV signals to and from the array core. The operation cycle was divided into two phases, such that the low-voltage core signals were only produced at the falling edge of the clock. The data out signals were held by a negative latch during the positive clock phase to eliminate hold violations when passing the data to the digital circuitry. The general architecture of the array is shown in Fig. 17 and the test chip micrograph is shown in Fig. 18. An extensive description of the array architecture and operation cycle can be found in [7], for which the same test setup was used.

4.3. Test chip measurements

The test chips were fabricated with a dual interface enabling internal testing with an on-chip built-in-self-test (BIST) unit, as well as external testing with a standard FPGA for better debugging. Three power supplies were used: 2.5 V for I/O circuits, 1.1 V for digital logic, and 400 mV for the array core. The 400 mV supply was provided by the Agilent B1500a Semiconductor Device Analyzer in order to accurately measure power consumption. This also enabled testing the functionality and power consumption of the chips while sweeping the array core supply voltage. All packaged test chips functioned successfully at the specified voltage (400 mV) and were further tested at voltages up to 1.1 V.

Fig. 19 shows representative waveforms of the chip testing. In Fig. 19a, signals from the FPGA’s digital chip scope show writing and reading a 32-bit word, while the analog behavior of a single bit being written and subsequently read is shown in Fig. 19b. A write operation is carried out when both the write–read (WR) and chip select (CS) signals are high, whereas a read operation occurs at the falling edge of the clock (CLK) after WR goes low and CS goes high.

At the specified voltage of 400 mV, the test chips were operated at 1 MHz. The frequency limit is dominated by the peripheral circuits which were not optimized for performance. Note that read access time is highly dependent on the array architecture (i.e., the number of cells on a bitline) and the sensing scheme. Various techniques can be implemented to improve upon these, as shown in other publications. The static power consumption of the array was measured to be 21 nW for an array filled with ‘1’s, and 45 nW for an array filled with ‘0’s. The power consumption for continuous reads and writes of an equivalent number of ‘0’s and ‘1’s distributed amongst the array was 234 nW for operation at 1 MHz.

Table 1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Node</td>
<td>0.13 µm</td>
<td>0.13 µm</td>
<td>65 nm</td>
<td>65 nm</td>
<td>90 nm</td>
<td>40 nm LP</td>
<td>40 nm LP</td>
</tr>
<tr>
<td>Capacity</td>
<td>2 kb</td>
<td>4 kb</td>
<td>256 kb</td>
<td>256 kb</td>
<td>32 kb</td>
<td>8 kb</td>
<td>8 kb</td>
</tr>
<tr>
<td>Transistor count</td>
<td>6T</td>
<td>10T</td>
<td>2T</td>
<td>10T</td>
<td>8T</td>
<td>9T</td>
<td>9T</td>
</tr>
<tr>
<td>$V_{DD,core}$</td>
<td>193 mV</td>
<td>160 mV</td>
<td>380 mV</td>
<td>350 mV</td>
<td>180 mV</td>
<td>250 mV</td>
<td>400 mV</td>
</tr>
<tr>
<td>Word lines</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Bit lines</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Floating supplies</td>
<td>VVDD, VGND</td>
<td>N/A</td>
<td>N/A</td>
<td>VVDD</td>
<td>N/A</td>
<td>VGND</td>
<td>N/A</td>
</tr>
<tr>
<td>WL boost</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>100 mV</td>
<td>50 mV</td>
<td>33%</td>
<td>None</td>
</tr>
<tr>
<td>Frequency</td>
<td>21.5 kHz @210 mV</td>
<td>620 kHz @400 mV</td>
<td>475 kHz @400 mV</td>
<td>25 kHz @350 mV</td>
<td>581 kHz @300 mV</td>
<td>1 MHz @400 mV</td>
<td>1 MHz @400 mV</td>
</tr>
<tr>
<td>Dynamic power</td>
<td>~1.8 µW @400 mV</td>
<td>0.146 µW @400 mV</td>
<td>3.28 µW @400 mV</td>
<td>3.39 µW @350 mV</td>
<td>1.44 µW @300 mV</td>
<td>285 nW @400 mV</td>
<td>234 nW @400 mV</td>
</tr>
<tr>
<td>Leakage power</td>
<td>~120 nW</td>
<td>~25 µW</td>
<td>~2.5 µW</td>
<td>2.2 µW</td>
<td>363 nW</td>
<td>37–60 nW</td>
<td>21–45 nW</td>
</tr>
</tbody>
</table>
A comparison of several recent sub-threshold SRAM designs is presented in Table 1. This table displays various relevant figures of merit, including technology node and transistor count; minimum operating voltage ($V_{\text{min}}$); number of peripheral buses and access schemes; measured operating frequency; and power metrics.

5. Conclusions

In this paper, we presented the ultra-low leakage 9T quasi-static RAM bitcell and provided an extensive analysis of this topology. Due to its quasi-static nature, the QSRAM bitcell displays somewhat depleted static noise margins; however, through dynamic stability analysis, the functionality of the cell was verified. We showed that the QSRAM bitcell is designated for low-voltage operation due to its robust write-ability and its non-linear reduction of SNM with voltage lowering. Simulations of the QSRAM cell were shown for a specified supply voltage of 400 mV, under which it retained functionality under global and local variations for the chosen 40 nm technology.

The QSRAM bitcell was compiled into an 8 kbit array and fabricated as part of a test chip in a commercial TSMC 40 nm LP process. Measurement results show full functionality for the specified 400 mV operating voltage at 1 MHz consuming 234 nW of power. Its static power figure was shown to typically be 1.5 to 4 times lower than a standard 8T bitcell held at 400 mV.

We would like to thank Mr. Nir Sever, the Zoran Corporation and be 1.5 to 4 times lower than a standard 8T bitcell held at 400 mV.

References