High-Performance SRAM in Nanoscale CMOS: Design Challenges and Techniques

Ching-Te Chuang, Saibal Mukhopadhyay, Jae-Joon Kim, Keunwoo Kim, and Rahul Rao
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, U. S. A.
Phone: +1-914-945-3596, Fax: +1-914-945-1358, e-mail: ctchuang@us.ibm.com

Abstract
This paper reviews the design challenges and techniques of high-performance SRAM in the “End of Scaling” nanoscale CMOS technologies. The impacts of technology scaling, such as signal loss due to leakage, degradation of noise margin due to \( V_T \) scatter caused by process variations and random dopant fluctuation, and long term reliability degradation such as NBTI, are addressed. Design directions and leakage/variation/degradation tolerant SRAM circuit techniques to mitigate various performance and reliability constraints in conventional planar CMOS technology are discussed. Examples are given and merits discussed for cell isolation and strength preservation, thin cell layout, bit-line and word-line leakage mitigation, migration to large signal read-out, unclamped bit-line, dual-supply, dynamic Read/Write supply, floating power-line, header/footer power-gating structures, Read- and Write-assist circuits, leakage/variation detection and compensation techniques, word-line and bit-line pulsing schemes, gate leakage tolerant design, and NBTI tolerant design. Alternative cell structures, such as asymmetrical SRAM, 7T, and 8T SRAMs, which decouple the cell storage node from the word-line and bit-line are discussed. Finally, some design issues and opportunities in emerging technologies such as FD/SOI and multi-gate FinFET are illustrated.

1. Technology Scaling and SRAM Design Challenges
High-performance SRAMs are crucial elements for high-performance microprocessor cache memory and SoC applications. Due to large number of transistor count, the use of the smallest transistors in the cells, and leakage and variation in sub-100 nm CMOS technologies, SRAMs have become a major yield limiter. SRAM is more vulnerable to process variations and \( V_T \) mismatch than logic circuits since minimum size (or sub-groundrule, or groundrule “waivered”) devices are used in the cell, and there is no ”averaging” effect as in logic data paths. As the cell size for 6T SRAM is scaled down to between 0.5 \( \mu m^2 \) to 0.75 \( \mu m^2 \) at 65 nm node, and below 0.5 \( \mu m^2 \) at 45 nm node (Fig. 1) [1], the local random device \( V_T \) mismatch (\( V_T \) scattering) (Fig. 2) due to discrete dopant effect (Fig. 3(a)) [2, 3] further exasperates the problem to impose a fundamental limit on the attainable cell size. At 50 nm \( L_{eff} \) (90/65 nm node), there are only about 200 dopant atoms in the device channel. At 12 nm \( L_{eff} \) (32/28 nm node), the number of dopant atoms in the device channel is expected to reduce to merely 2.6 [2]. The “Sigma” (\( \sigma \)) of \( V_T \) variation has remained relatively flat for technology scaling down to around 130 nm node, yet it is rising quickly starting at 90 nm and is expected to increase by a factor of 4 to 30 nm node (Fig. 3(b)) [3]. This rapid increase in \( V_T \) variation, against the backdrop of lower supply voltage, has rendered the redundancy needed for repair of the projected array size at 45 nm impractical, and even prohibitive for designs below “5-Sigma”, as a “4-Sigma” design requires 33 fixes per Mb of cells (0.57 fixes per Mb of cell for “5-Sigma” design).

While it is possible and circuit techniques have been developed to monitor, detect, and compensate for the global systematic die-to-die variations, it is much more difficult to develop similar schemes for local random variations. In addition to leakage and variation, NBTI (Negative Bias Temperature Instability) [4, 5] has recently emerged as a limiting factor in the scaling of PMOS. At high negative bias and elevated temperature, the PMOS \( V_T \) gradually drifts to become more negative, thus reducing PMOS current drive and affecting cell stability, margin, and \( V_{min} \) (minimum operating voltage) of SRAM. This long term \( V_T \) drift and other gate oxide related \( V_T \) degradation mechanisms must be accounted for over the life span of usage. Furthermore, with high-\( k \) gate materials, PBTI (Positive Bias Temperature Instability) in NMOS must be considered, as high-\( k \) gates exhibit significant charge trapping and thus \( V_T \) shift in NMOS [6, 7].

2. Cell Stability and Disturb
The cell stability, Static Noise Margin (SNM), and \( V_{min} \) of 6T SRAM are limited by leakage, variation, and supply voltage in the physical domain. In the design domain, the major limiting factors are the conflicting Read/Write requirements and cell disturbs.

Read/Write Requirements: To facilitate Read and minimize Read-disturb (Fig. 4), it is desirable to have a strong pull-down NMOS and a small \( \beta \) ratio (strength ratio of access pass-transistor NMOS to pull-down NMOS). However, in order to improve Writeability (Write margin) and Write performance, it is desirable to have a strong access pass-transistor NMOS.

Read Disturb: Conventional 6T SRAM is most unstable (worst SNM) in Read operation (Fig. 4), since the access pass-transistor NMOS and the pull-down NMOS form a voltage divider, and the cell “0” storage node rises during Read (Read-disturb voltage). If the Read-disturb voltage is larger than the “Trip” voltage of the other half-cell inverter, the cell can flip. As shown in Fig. 4, with technology scaling, the variations in the Read-disturb voltage (“Cell down-level”) and inverter Trip voltage (“Cell switch-point”) increase, causing overlap and thus failure starting at 90 nm node [8].

Half-Select Disturb: During a Read or Write operation, the half-selected cells on the same word-line are actually experiencing a Read operation, and thus disturb similar to Read-disturb described above (Fig. 5).


3. Circuit and Design Techniques

Many circuit and design techniques have been developed to mitigate the leakage and variation, to improve the cell stability, noise margin, and to extend the scalability of 6T SRAM. The most notable ones are:

**Thin-Cell Layout:** Starting around 90 nm node, thin-cell layout with uni-directional poly (Fig. 6) have become prevalent across the entire industry [9-14]. The “thin” vertical height reduces bit-line loads for performance and noise immunity, and the uni-directional poly improves the manufacturability and yield.

**Hierarchical Bit-Lines with Short Local Bit-Lines:** The short local bit-lines reduce the bit-line load, charge injection into the cell, leakage, and noise coupling, thus improving the performance, power, and noise margin. The global bit-lines are typically reset to “Low”, thus further reducing the power [15-17].

**Large Signal Domino-Like Sensing:** Large signal read-out schemes with full swings at local and global bit-lines have replaced the traditional small signal differential sensing schemes to ensure stability, combat variability, improve margin, and address reduced cell current problem. In most cases, single-ended large signal Read and differential Write are employed [16, 17].

**Unclamped (Floating) or Weakly Clamped Local Bit-Lines:** The local bit-lines are left unclamped (floating) or weakly clamped to suppress or reduce the leakage from the reset (precharge) devices during the standby mode or for any unselected sub-arrays during active mode. This is achieved by turning off the precharge devices in the case of unclamped bit-lines [16, 17]. In the case of weakly clamped bit-lines, the bit-line precharged voltage level is lowered by, for example, one Vt drop by using NMOS precharge devices in source-follower configuration (as opposed to full rail precharge using PMOS in the clamped bit-line case). Unclamping of bit-lines also reduces the half-select disturb since the voltage level at these bit-lines drift lower due to the current and leakage.

**Dual Supplies:** The supply voltage for cells and performance critical peripheral circuits (VCS) is set at a level 150mV – 200mV higher than the logic supply level (VDD) to ensure cell stability and improve performance, while less-critical peripheral circuits stay at VDD to reduce power consumption [15-17]. Fig. 7 illustrates the voltage domains of the 6.0 GHz SRAM array for CELL Broadband Engine™ in 65 nm PD/SOI technology. The cell, word-line, bit-line precharge signals are at VCS (1.0 V) to improve cell stability, Read performance, and allow writeability to track cell stability. Local bit-lines, evaluation, and global write lines are at VDD (0.8V) to reduce cell stress and lower power. This dual supply scheme has helped to improve Vmin from 0.875 V to 0.80 V.

**Adaptive Read/Write Supply:** The cell supply voltage can be adaptively/dynamically switched based on Read, Write, or Standby. Fig. 8 depicts a dynamic supply scheme using column based header switches [18]. High cell supply voltage is used for Read (VCCCell > VWL), while lower cell voltage (200 mV lower) is used for Write (VCCCell < VWL). The scheme improves noise margin for both Read and Write while minimizing leakage, and significantly reduces the random single bit failure in a 3.0 GHz, 70 Mb SRAM in 65 nm technology. Proper combination of adaptive Read/Write supply and cell body bias improves the Read and Write margin, and has been shown to enable 30% power reduction and 0.3 V Vmin for a 64 Kb low-power SoC SRAM in 90 nm CMOS technology [19].

One variation of adaptive Read/Write supply scheme is the “Floating Power-line Write” where the cell supply voltage for the selected column is switched off by column power switch, thus putting the cell supply node in a floating state. In this scheme, the Write current lowers the cell supply node for the selected column, thus facilitating Write, reducing Write power, and improving Vmin by 100 mV in 32Kb/512Kb SRAMs in 90 nm CMOS [20].

Capacitive coupling technique has also been used to bootstrap cell supply voltage to above VDD during access to achieve higher cell current and improved stability in a pico-Joule class, 1 GHz, 32 KByte x 64-b DSP SRAM [21].

**Power-Gating with Header/Footer:** Power-gating structures are used to reduce the leakage in Sleep/Standby mode or Shut-off mode [21, 22]. Fig. 9 illustrates the scheme used in the 3.4 GHz, 16MB L3 cache in dual-core Xeon processor in 65 nm CMOS technology [22], where NMOS sleep transistor is used in the sub-arrays and PMOS power-gating is used in the periphery.

**Read-Assist and Write-Assist Circuits:** The variation detection and compensation technique is best illustrated in the scheme depicted in Fig. 10 [23]. In this scheme, a process and temperature variation tolerant Read-Assist Circuit (RAC) utilizes Replica Access Transistors (RATs) as variation sensing devices. The RATs have topologically the same layout structure as SRAM access transistors. The selected WL level inversely tracks strength of access transistors, thus compensating the access transistor strength variation.

Capacitive coupling technique has also been employed to facilitate Write. Fig. 11 shows a Write-Assist Circuit (W-AC) with divided array supply scheme [23]. In this scheme, the array supply and a dummy supply (at GND level in Standby or Read) in the selected column are put into floating state during Write. The array supply is then shorted to the dummy supply by an NMOS switch, causing the array supply to be capacitively coupled down to facilitate Write. The capacitive coupling effect can be enhanced by dividing the array supply line into short segments. The capacitive coupling effect provides a fast, large response, and there is no need for additional supply. The scheme enhances Write margin against increasing variation due to scaling, resulting in significantly higher yield even with a much smaller cell in a 4 x 256 Kb SRAM macros in 45 nm LSTP CMOS technology.

**Word-Line and Bit-Line Pulsing:** Proper control of word-line pulse width and temporarily lower bit-line voltage are also effective techniques to improve cell stability [24]. The word-line pulse width must be short enough to limit the Read-disturb; while long enough to ensure adequate ΔVth for sensing during Read, and long enough to maintain enough Write margin. The bit-line voltage is pulled-down by 100 – 300 mV just before word-line turns on to reduce...
Read-disturb during Read, and half-select disturb during Write.

Gate Leakage Tolerant Design: The gate leakage of large devices in word-line driver can be reduced by collapsing the voltage across the gate oxide. In [25], the voltage level at the gate of the word-line driver NMOS device is left floating, and discharged by junction leakage in standby, thus reducing gate leakage.

NBTI Tolerant Design: The bias dependence of NBTI can be exploited to minimize the impact on long term reliability. Fig. 12 shows a NBTI tolerant sense amplifier design for L1 cache and TLB in a 64b microprocessor in 130 nm technology [5]. In the original circuit (Fig. 12(a)), the $V_T$ shift over silicon lifetime due to NBTI causes $V_T$ mismatch of as high as 50 mV between pMOS M3 and M4, resulting in sense delay degradation of 42%. Noting that the NBTI induced $V_T$ shift depends strongly on the gate-source bias and temperature but barely on the drain voltage, the new circuit (Fig. 12(b)) replaces the cross-coupled M3/M4 by T3/T4, for which gates are commonly biased at about 40% $V_{DD}$ during sensing. As the gate bias is identical for T3 and T4, the $V_T$ mismatch is minimized. pMOS T1/T2 are added to speed up the output low-to-high transition. Their $V_T$ mismatch is not critical since they get activated after the critical part of the amplification. The new circuit improves the deteriorated sense delay by 22%.

4. 8T SRAMs

8T SRAM has recently emerged as a promising candidate to replace 6T SRAM in sub-32 nm technologies [26-29]. The 8T SRAM utilizes a cell schematically similar to cells commonly used in register files, by separating the Read word-line and Write word-line and adding 2 stacked NMOS for single-ended Read (Fig. 13). If offers the following advantages:

Read Disturb: By separating the Read word-line and Write word-line and isolating the cell storage node from Read current path, the Read disturb is completely eliminated.

Half-Select Disturb: Half-select disturb still exists during Write operation for 8T SRAM. The half-select disturb can be eliminated by:

(1) Array architecture approach: The array can be floor-planed such that all bits in a word are spatially adjacent, and thus requiring no column select. The constraint for this approach is that bits from different words can not be physically interleaved. This approach has been used in a 5.3 GHz, 0.41 $V_{min}$ 32Kb sub-array in 65 nm PD/SOI technology [28].

(2) Gated Write word-line signal (Byte Write): The local Write word-line select signal is gated, and “on” only for the selected block. This scheme has been implemented in a 6.6+ GHz, 0.4 $V_{min}$ dual-supply, 1.2 Mb SRAM in 65 nm PD/SOI technology [29].

(3) Write-back scheme: The Read word-line is activated even during Write, and all cell data in selected word-line are read out to D-latches. The Data-out is then written back to half-selected cells. The scheme has been used in a 0.42 $V_{min}$ 64 Mb SRAM in 90 nm CMOS technology [27].

Cell Size Scaling: At 90 nm node, the cell size of 8T SRAM is bout 20% larger than 6T SRAM cell. However, in 6T cell, the device width of cell pull-down NMOS cannot be scaled down without degrading the Read performance and Read disturb, while that in the 8T cell can. At 45 nm node, the cell areas cross if the operating voltage is 1.0 V, and the area of the 8T cell becomes smaller at the 32 nm node. If the operating voltage is 0.8 V, the 6T cell area is saturated and cannot be shrunk beyond 45 nm due to the large width of the cell pull-down NMOS needed for Read performance and margin. Furthermore, in 8T cell, high voltage can be used for Read word-line to enhance Read performance without causing Read disturb; and high voltage can be used for Write word-line to facilitate Write and expand Write margin. These voltage control schemes further enable the scalability of 8T SRAM cell as shown in Fig. 13 [27].

5. Asymmetrical SRAMs

The Read-disturb in 6T SRAM can be mitigated/eliminated by either cutting off the inverter feedback path or skewing one of the cell inverter to have a higher Trip voltage, while employing single-ended Read with the other cell inverter.

7T Asymmetrical SRAM: Fig. 14(a) shows the schematics of a 7T asymmetrical SRAM cell [30]. In this scheme, the Read word-line is separated from the Write word-line, and single-ended Read is employed. An extra NMOS is inserted in the left-side cell inverter, which cuts off the feedback loop during Read, thus eliminating the failure due to Read-disturb. The scheme has been implemented in a 64 Kb SRAM in 90 nm CMOS technology and demonstrated 20 ns access at 0.5V. $V_{min}$ of 0.44 V (determined by Write SNM) has been achieved with 9% area overhead.

6T Asymmetrical SRAM: Fig. 14(b) shows the schematics of a 6T asymmetrical SRAM cell [31]. Again, the Read word-line is separated from the Write word-line, and single-ended Read is employed. The left-side cell pull-down NMOS is weakened, thus increasing the Trip voltage of the left-side half-cell inverter to improve the Read margin. Weakening the left-side pull-down NMOS by sizing down the device width alone is limited by the minimum width in a given technology, and normally does not provide sufficient improvement in Read SNM. High-$V_T$ device, thick-oxide device, or mid-gap gate material (which increases $V_T$) can be used in combination with device down-sizing to further improve the Read margin [31].

6. Design Considerations for Emerging Devices

In this section, we discuss the design issues/considerations and some novel schemes exploiting the unique features of emerging devices such as fully-depleted SOI (FD/SOI) devices and multi-gate/FinFET devices.

6.1. Fully-Depleted SOI (FD/SOI) Devices

Due to better short-channel control, very low junction capacitance, and the elimination of hysteretic $V_T$ variation, Ultra-Thin Body Fully-Depleted SOI (UTB FD/SOI) device has emerged as a possible candidate for sub-65 nm technologies.

Random Dopant Fluctuation (RDF) Effect: The local $V_T$ variation due to RDF is expected to be more serious in FD/SOI devices compared with bulk-CMOS or PD/SOI devices. This is primarily due to the stronger dependence of bulk change ($Q_b$) on the body doping density ($N_b$) in FD/SOI device ($Q_b$ vs $N_b$) compared with bulk-CMOS or PD/SOI device ($Q_b$ vs $N_b$). The RDF effect can be reduced
by lowering the body-doping. However, reduction of body-doping increases the leakage current. Proper tuning of back-gate bias or gate workfunction, combined with body-doping reduction, has been shown to be very effective in achieving a given leakage target for robust SRAM applications [32, 33].

**Asymmetrical Vt Shift:** Fig. 15 shows the distribution of I-V characteristics due to RDF effect for PD/SOI and FD/SOI devices. For FD/SOI device, the mean Vt is lower than the nominal Vt expected from continuous doping case. This asymmetry in Vt shift causes a large spread in leakage, and degrades the performance and margin as the silicon film thickness is reduced (Fig. 16). This asymmetrical Vt shift originates from the penetration of drain fringing electric field underneath the channel, which affects the barrier at the source for carrier injection and degrades the Short Channel Effect (SCE). The source carrier injection (hence channel current) depends exponentially on the barrier height, and thus highly non-linear and highly asymmetrical. This effect can be mitigated by using thin Buried Oxide (BOX) structure, which reduces the penetration of drain fringing electric field, thus reducing the asymmetry and leakage spread, at small expense of performance and Write margin [32, 33].

6.2. Double-Gate/FinFET Devices

Quasi-planar multi-gate devices, such as FinFET, hold the promise to extend scaling beyond conventional planar CMOS technologies. The device structure imposes some constraints while, at the same time, offers unique features that can be exploited for SRAM applications [34-37].

**Fin Height (Device Width) Quantization:** The quantization of device width to integer multiple of fin height constrains the opportunity of cell transistor sizing for stability optimization in FinFET SRAM.

**Surface Orientation and Mobility:** In FinFET devices, different surface orientation can be realized by rotating the fin in layout (Fig. 17 and 18(b)). The orientation dependence of mobility can be exploited to optimize the β-ratio of cell transistors to improve the performance and margin. As shown in Fig. 18(a), with optimized orientation, the Read SNM can be improved by 23-35%, and access time by ~22-33%, at expense of containable Write margin loss and area [34].

**Independent Gate Feedback Schemes:** The independent-gate capability in double-gate devices can be exploited to improve performance and margin. Fig. 19(a) shows a 6T SRAM cell with “Yin-Yang” feedback with double-gate devices [35]. The front-gate and back-gate of the cell pull-down NMOS devices are tied together. As such, the “On” NMOS becomes stronger, and the “Off” NMOS becomes weaker, thus reducing the Read-disturb to improve the SNM. An improved version of the “Yin-Yang” feedback cell is shown in Fig. 19(b) [36], where the back-gates of the access pass-transistor NMOS’s are connected to the storage nodes. This scheme has the same Read characteristic as Yin-Yang feedback cell, yet improved Write performance and margin as the cell “High” side access pass-transistor becomes stronger.

**Back-Gated Asymmetrical SRAM:** The back-gating capability of asymmetrical double-gate device can be exploited for asymmetrical SRAM applications. Fig. 20(a) illustrates an example where a down-sized front-gate is used for the left-side pull-down NMOS while the back-gate is biased to GND. The scheme offers 1.9X SNM improvement over symmetrical cell and 1.5X improvement over asymmetrical cell with device sizing only (Fig. 20(b)) [37].

7. Conclusions

We have discussed CMOS technology scaling and resulting leakage/variation/reliability design challenges for high-performance SRAMs. Cell stability, Read/Write requirements, Read-disturb, and Half-select disturb are shown to impose constraints on performance, margin and cell scalability. Examples of circuit techniques to mitigate various power, performance, and reliability constraints are given. Alternative cell structures, such as 8T SRAM and asymmetrical SRAM, and their merits are discussed. Design considerations for emerging devices, such as FD/SOI and double-gate/FinFET are addressed, and novel circuit schemes exploiting unique features of emerging devices are discussed.

Acknowledgement

The authors were partially supported by DARPA Contract HR0011-07-9-0002.

References


Fig. 1. 2005 ITRS Product Function Size Trends [1].

Fig. 2. Global systematic and local random variation in device threshold voltage degrade yield of SRAM design.

Fig. 3. (a) Number of dopant atoms in the channel as function of effect channel length [2], and (b) relative “Sigma” of VT variation as function of technology node [3].
Fig. 4. Scaling and cell stability margin of 6T SRAM [8].

Fig. 5. Half-select disturb (illustration for Write operation).

Fig. 6. “Thin Cell” layout to improve performance, noise immunity, manufacturability, and yield [9-14].

Fig. 7. Example of SRAM with dual-supply, hierarchical short local bitline, large signal single-ended “ripple-dominio” Read, and differential Write [17].

Fig. 8. Column based header switch dynamic supply scheme to optimize Read/Write performance [18].

Fig. 9. Power-gating with footer switch [22].
Fig. 10. Process and temperature variation tolerant Read-Assist Circuits (RAC) [23].

Fig. 11. Capacitive Write-Assist Circuit (WAC) with divided array supply scheme [23].

Fig. 12. NBTI tolerant TLB sense amplifier circuit [5].

Fig. 13. Cell size scaling of 6T and 8T SRAM [27].

Fig. 14. (a) Asymmetrical 7T SRAM cell [30], and (b) asymmetrical 6T SRAM cell [31].

Fig. 15. Random Dopant Fluctuation (RDF) effect in (a) PD/SOI, and (b) FD/SOI device [32, 33].
Fig. 16. SRAM in FD/SOI: (a) leakage vs. Si film thickness, and (b) statistical distribution of Read-disturb voltage and cell inverter Trip voltage due to RDF effect [32, 33].

Fig. 17. Modification of surface orientation in FinFET [34].

Fig. 18. Optimized multi-oriented FinFET 6T SRAM cells: (a) Access Time vs. Read SNM, and (b) (100) PUP, (110) AX, (100) PD on (a) (100) wafer [34].

Fig. 19. Double-gate SRAM cells: (a) Yin-Yang feedback cell [35], and (b) improved Yin-Yang feedback cell [36].

Fig. 20. (a) Back-gated asymmetrical 6T SRAM cell in asymmetrical double-gate technology, and (b) Read SNM compared with symmetric 6T SRAM cell [37].