Energy Efficient Volatile Memory Circuits for the IoT Era

5

Jaydeep P. Kulkarni, James W. Tschanz, and Vivek K. De

This chapter addresses the challenges involved in designing energy efficient embedded Static Random Access Memory (SRAM) circuits for the IoT era. It discusses memory design for wide voltage range operation using 6 Transistor (6T), 8T and 10T bitcells and novel circuit assist techniques. In addition, it discusses future memory designs using emerging nano-wire FET, Tunnel FET, III-V FET, and monolithic 3-D technologies.

5.1 Introduction and Challenges in Embedded Memory Design for IoT

5.1.1 Introduction

Technological advances and form factor driven cost reduction have resulted in tremendous growth of computing devices. As shown in Fig. 5.1, the number of internet-connected devices deployed worldwide is projected to grow to 50 billion by 2020.

If continued, this trend might result in tens of billions of computing systems consisting of personal computers, desktop machines, smartphones, wearables and many units connected to the internet; collectively known as the Internet of Things (IoT). This dramatic growth in compute devices results in a data explosion (known as 'data deluge') and would require millions of Zettabytes of memory in order to perform this large scale of computing (Semiconductor Industry Association 2015). Therefore, memories play a critical role in future energy efficient computing systems. This chapter addresses the challenges involved in designing energy efficient embedded memory circuits operating across a wide voltage range, and also presents design examples using emerging device technologies.

5.1.2 SRAM Scaling Trends

Aggressive scaling of transistor dimensions with each technology generation has resulted in increased integration density and improved device performance. The SRAM bitcell area has been scaled by ~0.5× over each process generation as shown in Fig. 5.2. This area scaling is achieved by various lithographic and circuit innovations such as thin-cell layouts, high-K metal gate technology, tri-gate geometry, leakage-reduction techniques, and low voltage assist techniques (Wang et al. 2012; Yoshinobu et al. 2003).

J.P. Kulkarni (✉) • J.W. Tschanz • V.K. De
Circuit Research Lab, Intel Corporation, Hillsboro, OR, USA
e-mail: jaydeep.p.kulkarni@intel.com

M. Alioto (ed.), *Enabling the Internet of Things*, DOI 10.1007/978-3-319-51482-6_5

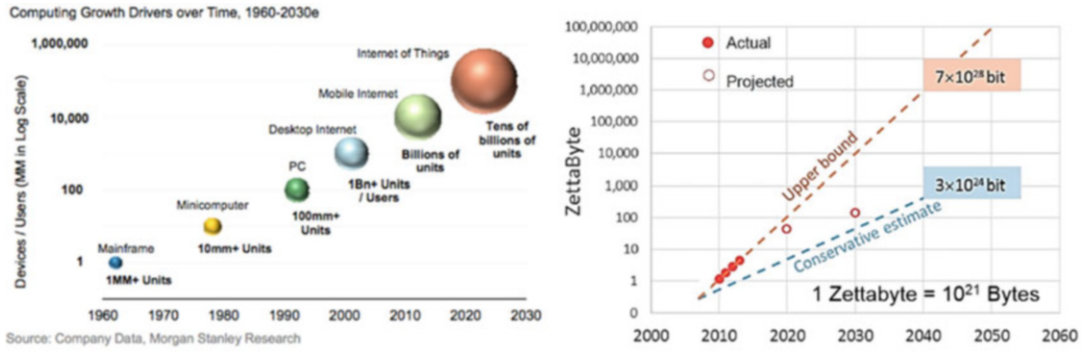


Fig. 5.1 Projected growth of IoT devices and the required memory capacity

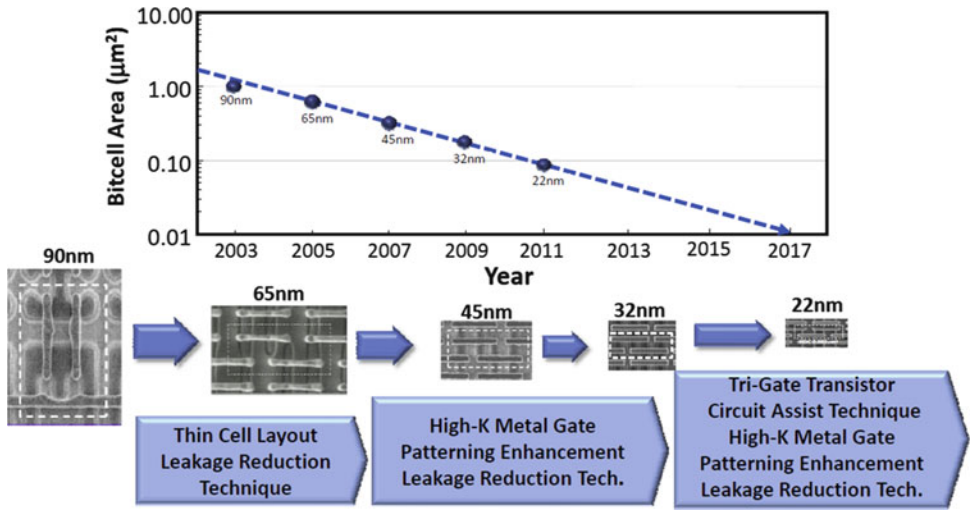


Fig. 5.2 6T Static Random Access Memory (SRAM) bitcell scaling trend

5.1.3 SRAM Design Octagon

Energy-efficient SRAM design in nanoscale technology needs to address several process, design, architecture, and reliability issues as shown in Fig. 5.3. Process technology issues relate to the restricted design rules, diffusion-notch free layout, proximity effects, stress effects, and threshold voltage (V_T) optimization. It impacts the 6T bitcell sizing/area, read/write stability, 6T+ bitcells, and single/multi-port bitcell requirements. These process optimizations and design requirements govern the array efficiency and bit density.

Circuit techniques focus on lowering the minimum successful operating supply voltage (also known as V_{MIN}) using various read/write assist techniques. Another important aspect is leakage reduction during active as well as idle mode. This is particularly important for IoT designs that in general have very strict power budgets and at the same time can exhibit very low array activity due to duty-cycled or burst-mode styles of operation.

Reliability is also an important aspect of modern SRAM design. The effects of bias temperature instability (BTI), time-dependent dielectric breakdown (TDDB), hot carrier injection (HCI), random telegraph noise (RTN), erratic bits, and radiation-induced soft errors limit the operating voltage range and lifetime of the SRAM array.

Architectural techniques such as redundancy, parity protection, error correcting codes (ECC),

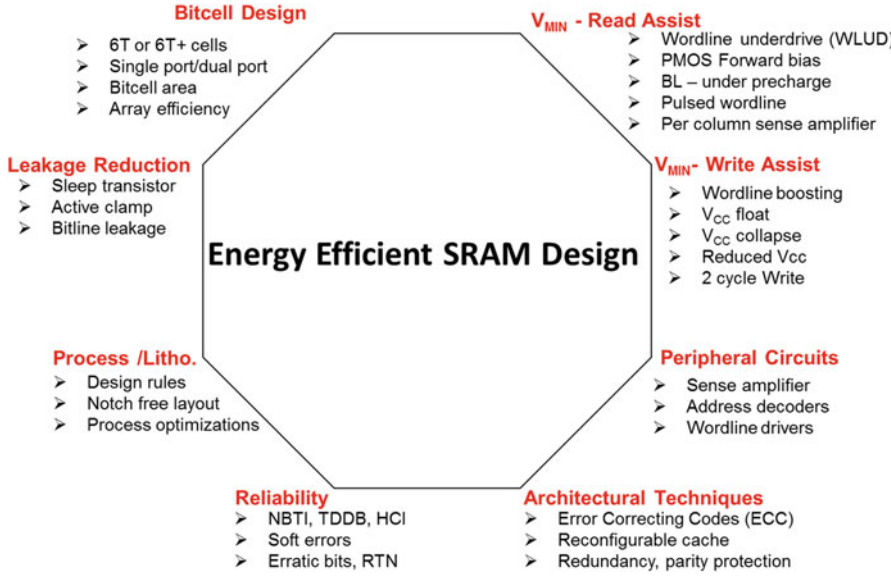


Fig. 5.3 Energy-efficient SRAM design requirements

and reconfigurable caches are used along with the circuit techniques to address these reliability issues. In this chapter, we focus on the circuit techniques for enabling energy-efficient SRAM operation across a wide voltage range specifically needed for IoT applications.

5.1.4 SRAM Parametric Failures and V_{MIN}

Supply voltage scaling has remained the major focus of energy-efficient memory design. However, as the supply voltage is reduced the sensitivity of the circuit parameters to process variations increases. Nanoscale SRAM bitcells having minimum-sized transistors are vulnerable to inter-die as well as intra-die process variations. Intra-die process variations include random dopant fluctuation (RDF), line edge roughness (LER), and other manufacturing variations that may result in a threshold voltage mismatch between adjacent transistors in a memory cell (Bhavnagarwala et al. 2001). Coupled with inter-die and intra-die process variations, supply voltage scaling is limited due to various memory failures (i.e. read failure, retention failure, access time failure, and write failure) also known as minimum operating voltage (V_{MIN}) (Mukhopadhyay et al. 2005; Khellah et al. 2008).

The 'de facto' memory bitcell used in modern SRAM designs is a 6-transistor cell consisting of a cross-coupled inverter pair. SRAM cells are sized to satisfy the conflicting design requirements of read stability and write-ability. For a read stability optimized bitcell as shown Fig. 5.4, pass gate transistors (AXL, AXR) are sized weaker than the pull-up and pull-down transistor.

During a read operation, when a wordline (WL) is asserted, the voltage at the node storing logic 0 (Node V_L in left Fig. 5.4) rises above V_{SS} due to voltage divider action between the pass-gate AXL and pull-down NL transistors. At nominal supply voltage, the V_L node voltage rise during read operation is not significant enough to flip the bitcell contents (shown in blue color). However due to process variations, especially for a bitcell operating at lower supply voltages, the voltage rise at V_L node can be higher than the trip point of the other inverter and can flip the V_R node, resulting in a read failure event (shown in red color). Read failure can also occur during a dummy-read scenario (half-select) in a column-interleaved design.

For the write operation, the design requirement is that the bitcell nodes should flip easily. In a write-optimized bitcell, the pass gate is

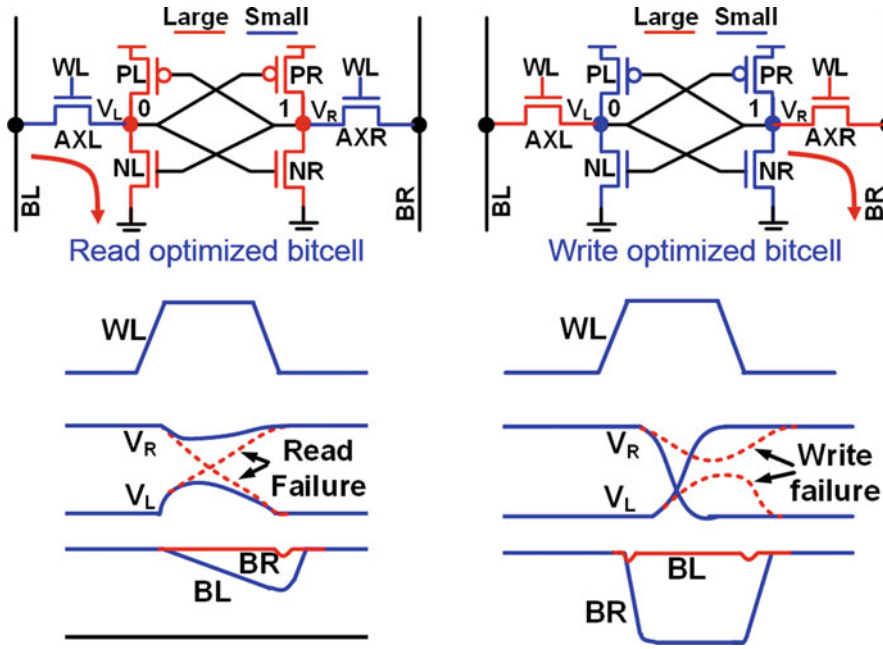


Fig. 5.4 Read and write failure events in a 6T SRAM bitcell

stronger than the pull-up and pull-down devices. At nominal supply voltage, when write data is applied on the bitline (BL) and bitline complement (BR), the pass gate connected to the grounded bitline (BR) pulls down the bitcell node (V_R in right Fig. 5.4) below the trip point of the other inverter (V_L) resulting in the flip of the storage bit (node V_L) and a successful write operation. However due to process variations especially at lower supply voltage, the pass gate AXR can be weaker than the pull-up device PR and may not lower the V_R node voltage below the switching threshold of the other side inverter resulting in a write failure event. Write failure can also happen if the wordline pulse is not long enough for the bitcell to flip the internal nodes.

5.2 6T SRAM Circuit Techniques

To improve the operating voltage range of SRAM arrays in the presence of process variations and to satisfy the conflicting design requirement of read vs. write stability, various circuit assist techniques have been explored. The key idea is to bias different nodes of the 6T bitcells (bitcell-V_{CC}/V_{SS}, WL, BL) appropriately to favor a read or a write operation.

5.2.1 Read Assist Techniques

As explained in the earlier section, a read failure is caused by increase in the access transistor drive strength compared to the cross-coupled inverter pair due to low voltage operation and/or process variations. The read assist circuits try to reinforce access transistors to be weaker than the cross-coupled inverter pair. The node biasing techniques include raising the bitcell-V_{CC} and/or lowering bitcell-V_{SS} of the cross-coupled inverter pair. These techniques also include wordline underdrive (WLUD) or operating the entire bitcell at a higher supply voltage compared to the peripheral circuits (Mann et al. 2010; Khellah et al. 2008) (Figs. 5.5 and 5.6).

5.2.2 Write Assist Techniques

Contrary to a read operation, a write failure is caused by decrease in the access transistor drive strength compared to the cross-coupled inverter

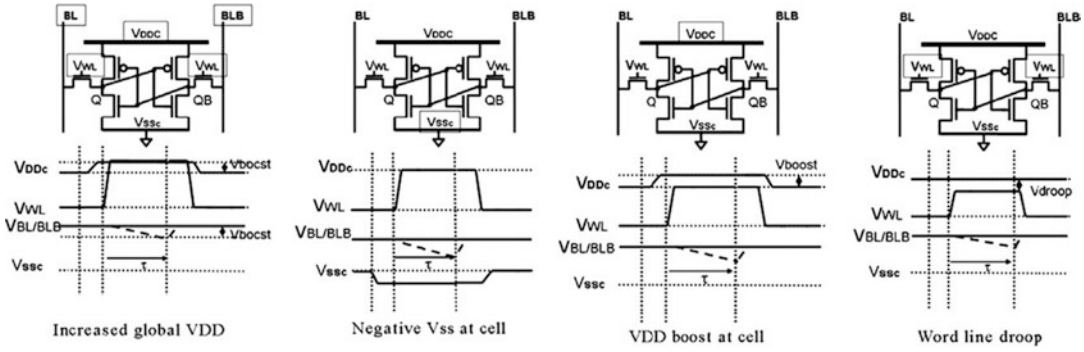


Fig. 5.5 6T SRAM read assist techniques

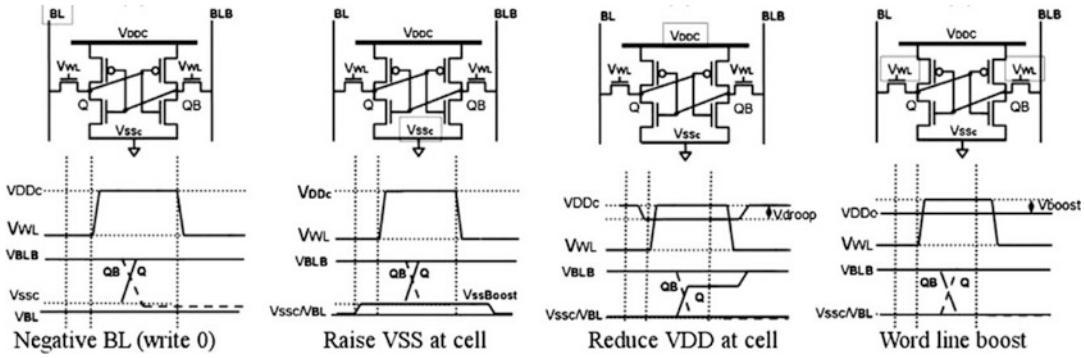


Fig. 5.6 6T SRAM write assist techniques

pair due to low voltage operation and/or process variations. The write assist circuits try to reinforce the access transistor to be stronger than the cross-coupled inverter pair. The node biasing techniques include lowering the bitcell-V_{CC} and/or raising the bitcell-V_{SS} of the cross-coupled inverter pair. Write assist techniques also include wordline boosting or negative bitline approaches (Mann et al. 2010; Khellah et al. 2008).

5.3 8T SRAM Circuit Techniques

5.3.1 Benefits of Decoupled Read/Write Operation

The fundamental conflicting design constraints of read-stability vs. write-ability in the 6T SRAM bitcell limits low voltage operation and requires power and area hungry V_{MIN} assist techniques. The operating voltage can be reduced substantially by adding a dedicated read port using two extra transistors. This 8T bitcell (Fig. 5.7) is commonly used in microprocessor cores for performance-critical low-level caches and multi-ported register-file arrays.

The 8T bitcell offers fast read and write operation, dual-port capability, and generally lower V_{MIN} than the 6T SRAM bitcell. By using a decoupled single-ended read port with domino-style hierarchal read bit-line, the 8T cell features a fast read evaluation path without causing access disturbance that limits read-V_{MIN} in the 6T bitcell. Using the 8T cell in a half-select-free architecture eliminates pseudo-reads during partial writes, hence enabling write-V_{MIN} optimization independent of read. It is well known that both with-in-die (WID) and die-to-die (D2D) device parameter variations are getting worse

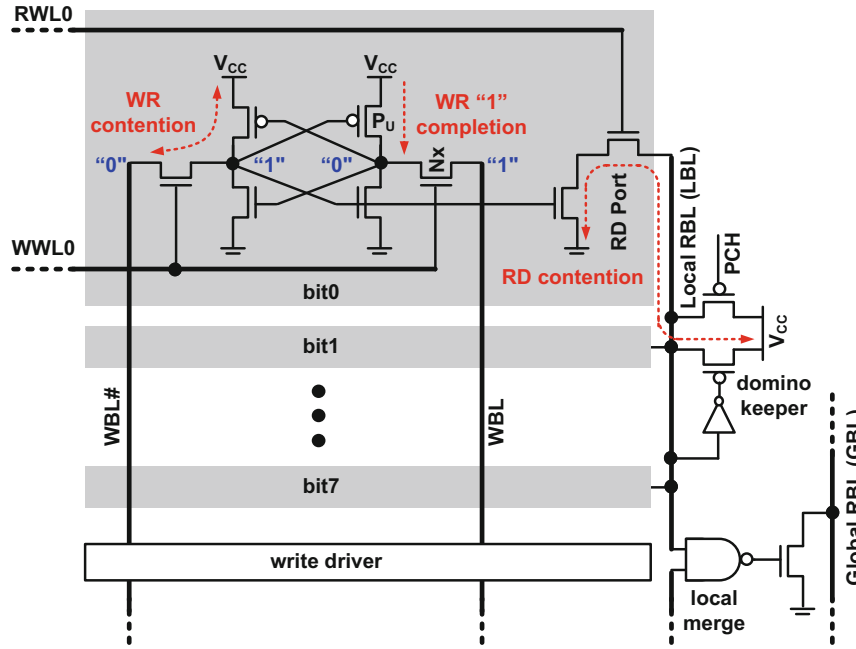


Fig. 5.7 8T bitcell array organization with non-interleaved columns

with feature size scaling. Unfortunately, the typical approach of sizing up the 8T read and write ports to mitigate process variation has limited V_{MIN} returns.

In the read case, using larger NMOS read port helps when reading a "1" by reducing contention with the PMOS domino keeper on the local bit-line (LBL). To compensate for degraded noise margin that comes with a larger (and leakier) read port, the PMOS keeper needs to be relatively upsized for good reading of a "0", resulting in diminishing read-V_{MIN} returns with continued upsizing. Similarly, contention between PMOS pull-up (P_U) and the NMOS pass (N_X) impacting write-V_{MIN} is reduced by sizing up N_X. However, P_U needs to be relatively sized up as well since at a given point a weak P_U will limit the completion of write within a given WL pulse. Only 10% total V_{MIN} reduction is attained through optimal cell upsizing at a cost of ~25% increase in array area (Raychowdhury et al. 2010). Therefore, read and write assis circuit that can achieve V_{MIN} reduction at a minimal area impact are necessary.

5.3.2 Wordline Boosting as V_{MIN} Assist Technique

Wordline boosting is an effective technique for lowering the read and write V_{MIN} of 8T bitcell arrays with minimal area and power overheads. The key idea is to selectively boost critical V_{MIN}-limiting nodes of the 8T bitcell, allowing the majority of the array peripheral circuits and remaining logic block to operate at much lower voltage thereby lowering the overall chip V_{MIN}.

There are various ways to achieve a boosted voltage for the wordline, such as charge-pump based boosting, capacitive coupling, and using separate high-V_{CC} rail. All three schemes rely on one or more combinations of (1) read wordline (RWL) boosting, (2) write wordline (WWL) boosting, and/or bitcell boosting. Boosting RWL and bitcell-V_{CC} enable larger read "ON" current without forcing a larger PMOS keeper. Boosting WWL helps write-V_{MIN} for two reasons—improving contention without upsizing N_X (or lowering its V_{TH}), and improving completion by writing a "1" from the

other side. Unlike V_{CC} collapse, WWL boosting does not degrade the dynamic retention margin of unselected cells on the same column.

5.3.2.1 Charge Pump Based Wordline Boosting

In this approach, an embedded charge pump is used to generate a higher voltage to boost RWL and WWL (V_{BOOST} generation) (Raychowdhury et al. 2010). Figure 5.8 shows a 2 KB 8T SRAM bank with a locally-integrated charge pump (CP) and associated wordline level shifter circuits. The CP itself is divided into ten identical units placed in the layout slices created by the level shifters in the global IO of the 2 KB macro. The boosting ratio is adjusted based on the built-in read-ability and write-ability sensor. The ideal boosting ratio (V_{BOOST}/V_{CC}) under no load current (I_{LOAD}) is $2V_{CC}$.

The actual boosting ratio is lower, however, as determined by I_{LOAD} from all active and inactive level shifters, boosting clock (BCLK) frequency (F_{BCLK}), and boosting capacitance (C_{CP}). To minimize the load current requirement on the charge pump design, a two-stage level shifter is used as the wordline driver. Unlike the conventional (DCVS) level shifter, where a "0"--to-V_{BOOST} transition is all supplied by the V_{BOOST} rail, the two-stage level shifter performs this transition in two steps. In the first step, "0"-to-V_{CC} is supplied by M_{P1} at which point M_{P2} kicks in to supply the remaining V_{CC}-to-V_{BOOST}. This significantly reduces the maximum charge pump load current allowing the boosting ratio to increase to ~1.6×, compared to only 1.1× using DCVS level shifter.

Figure 5.9 shows measurement results from a 45 nm test chip achieving V_{MIN} improvement of

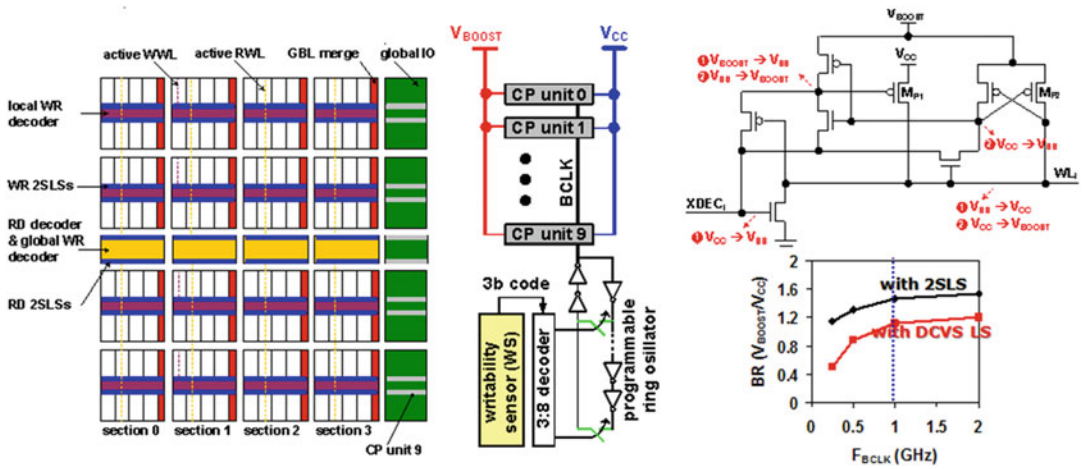


Fig. 5.8 8T SRAM array with distributed charge pump and two-step level shifters

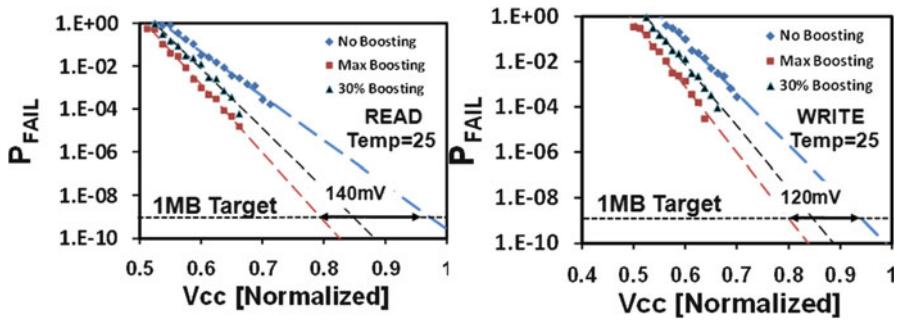


Fig. 5.9 Charge pump based boosting: Measured bit failures vs. V_{CC} extrapolated to 1 MB array

140 mV (120 mV) for read (write) for a 1 MB target array size. The area overhead due to distributed charge pump and two-stage level shifters is significant (around ~25%).

5.3.2.2 Capacitive Coupling with Self-Induced-Collapse

The adoption of charge pump based boosting requires careful design of the charge pump and two-stage level shifters. Furthermore, dynamically turning the charge pump off (or putting it in a drowsy mode) during inactive periods is necessary for net power savings. As an alternative to the charge bump, the capacitive coupling (CC) scheme achieves write-wordline (WWL) boosting using write-bitline (WBL) to WWL coupling. This eliminates the need for a power-hungry charge pump as well as complex level shifters. The basic idea is to take advantage of the large G-S/D capacitance to create a boosted voltage on the WWL without using any charge pump, level shifter or high voltage supply (Kulkarni et al. 2010). There are two locations on the WWL where this capacitance can be found (Fig. 5.10). The first is at the WWL interface to the PMOS/NMOS devices of the final WWL driver (C1), while the second is at the WWL interface to the cell's NMOS WR pass devices (C2 and C3).

To enable use of the first capacitance, the input of the WWL driver is asserted normally (high-to-low) to create a $0 \rightarrow V_{CC}$ transition on the WWL. After a short delay, the input is de-asserted (goes from low to high) turning off the top PMOS (but without turning on the bottom NMOS)-effectively floating the WWL. This in turn excites the G-to-D capacitance from the PMOS transistor of the WL driver creating about 3–5% coupling to the floating WWL. To enable use of the second capacitance, both WBL and WBLx are pre-discharged and, depending on data polarity; one of the bit-lines makes a $0 \rightarrow V_{CC}$ transition. This rising transition on the bitline and the internal bitcell node is capacitively coupled to the floated WWL, boosting it by ~20% of V_{CC}. This scheme is scalable to any number of bits per WWL with the scaling of the WWL driver size and the per-bit coupling capacitance.

A beneficial side effect of pre-discharging both the WBL and WBLx prior to the write operation is a self-induced collapse (SIC) of the

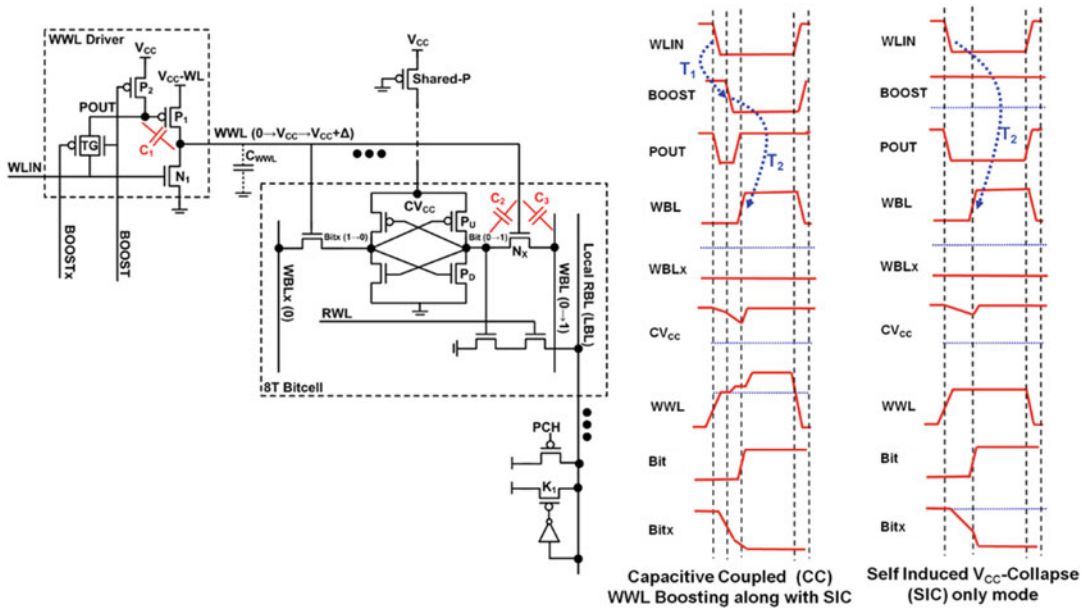


Fig. 5.10 Capacitive coupled WWL boosting with self-induced-collapse (SIC)

virtual bitcell voltage when the WWL is asserted. While SIC is inherent with the CC boost technique, it can also be used alone as a low-overhead write-V_{MIN} reduction technique, and is an effective alternative when WWL boosting violates gate-oxide reliability limits at high voltage. The SIC-only mode can be enabled by keeping the boost signal high (Fig. 5.10), ensuring that the WWL is not floated or boosted during the write operation.

Figure 5.11 shows the measured write failure rate versus supply voltage for CC boost and SIC-only modes. The slope of the write P_{FAIL} curve is governed by the SIC magnitude, WWL boost ratio, and how late the WBL transitions with respect to WWL activation. Extrapolating P_{FAIL} data to 1 MB array size demonstrates 140 mV reduction in write-V_{MIN} for CC boost and 80 mV reduction for SIC-only mode at optimal timing settings at 1.6 GHz operating frequency.

At lower frequencies, V_{MIN} savings increase to 180 mV for CC boost and 130 mV for SIC-only mode (Fig. 5.12). Both SIC and CC boost incur an increase in array power when run at the nominal voltage as baseline due to additional switching of the WBLs, bitcell V_{CC}, and overhead circuitry. However, both techniques enable V_{MIN} scaling beyond the baseline. Total array power savings when operating at lower V_{MIN} are 12% for SIC and 27% for CC boost (Fig. 5.12).

5.3.2.3 Dual-V_{CC} Design

Dual-V_{CC} based boosting selectively increases the voltage of critical nodes in an 8T-bitcell while incurring no array area overhead. A separate voltage $V_{BOOST} \leq V_{MAX}$, supplied externally or generated locally from a fixed high input voltage rail (V_{IN}) using a step-down voltage regulator (VR), is used to "boost" selected read/write wordlines (R/WWLs) and cell-V_{CC} (during read only) as shown in Fig. 5.13 (Kulkarni et al. 2013).

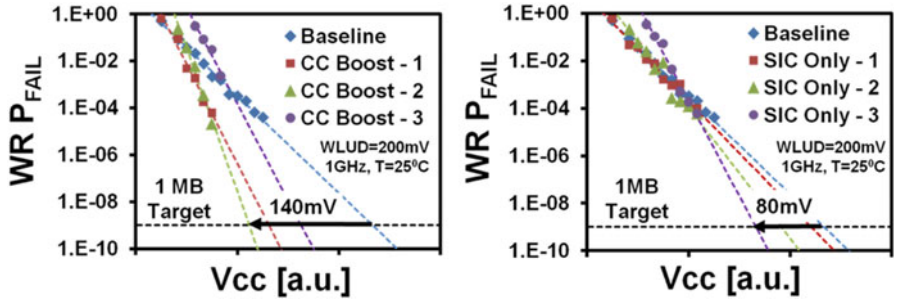


Fig. 5.11 Measured bit failure rate vs. V_{CC} for capacitive coupling and self-induced collapse technique

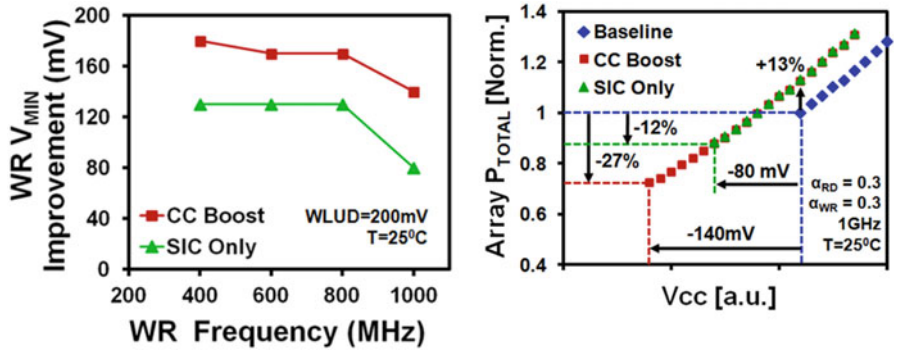
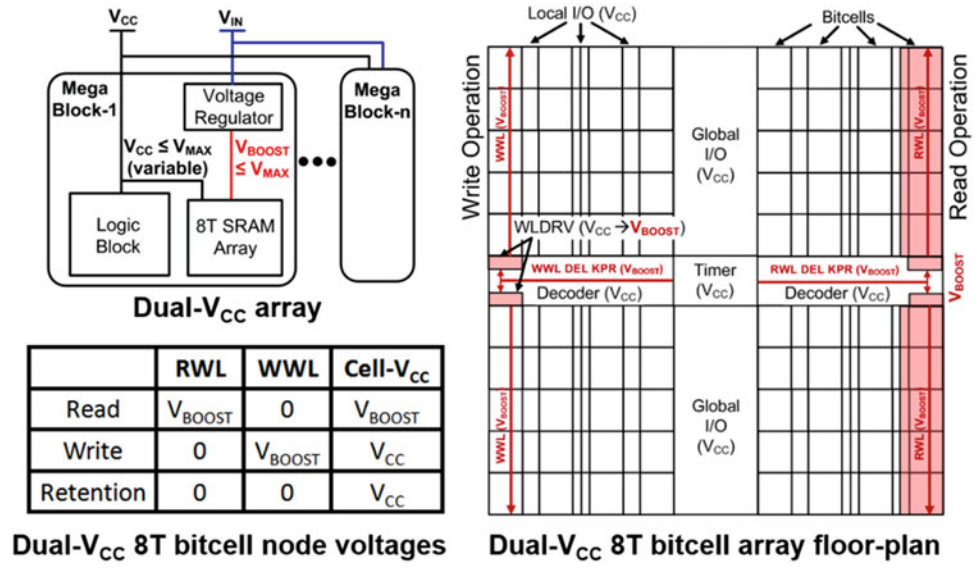


Fig. 5.12 Measured V_{MIN} improvement with frequency and power measurement



	RWL	WWL	Cell-V_{CC}
Read	V_{BOOST}	0	V_{BOOST}
Write	0	V_{BOOST}	V_{CC}
Retention	0	0	V_{CC}

Dual-V_{CC} 8T bitcell node voltages

Fig. 5.13 Dual-V_{CC} approach and array floorplan

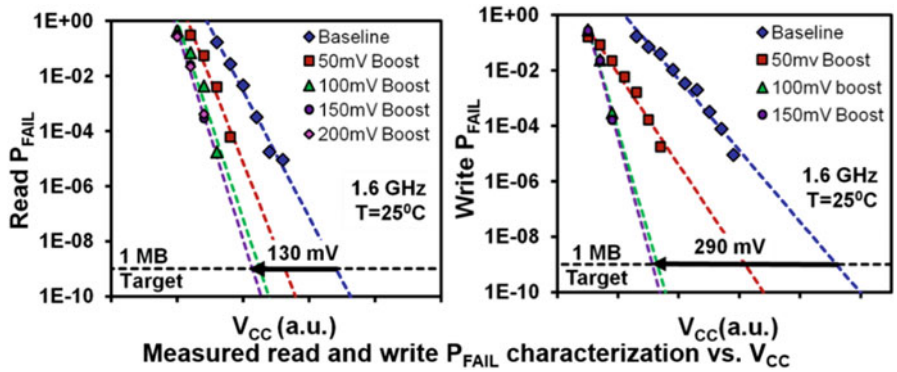


Fig. 5.14 Measured read and write failures vs. V_{CC} for varying boosting levels

All remaining array circuits such as R/WWL pre-decoder, pre-charge logic, local and global bitline (LBL/GBL) sensing, timer, and column- I/O drivers are connected to the variable $V_{CC} \leq V_{MAX}$ that is shared with core logic operating across a wide voltage range. By decoupling the V_{MIN}-limiting 8T bitcell from remaining array and core logic, overall chip V_{MIN} can be reduced, thus improving energy efficiency. During a read operation, selected RWL and associated bitcells are switched to Vboost to enable overdrive of the read-port transistor stack. This alleviates keeper contention and also improves bitline evaluation delay compared to the baseline single-V_{CC} design. During a write operation, selected bitcells remain at V_{CC} while the WWL is boosted to mitigate contention between the pass NMOS and pull-up PMOS in the bitcell. WWL boosting also aids write completion by passing a strong "1" through the pass NMOS. A dynamic level-shifting NAND WL decoder replaces the static single-V_{CC} NAND implementation while fitting in the same area.

Measurement results from a 22 nm tri-gate CMOS process show 130 mV read-V_{MIN} improvement and 290 mV write-V_{MIN} improvement when extrapolated to a 1 MB target array size at 1.6 GHz (Fig. 5.14). As boosting

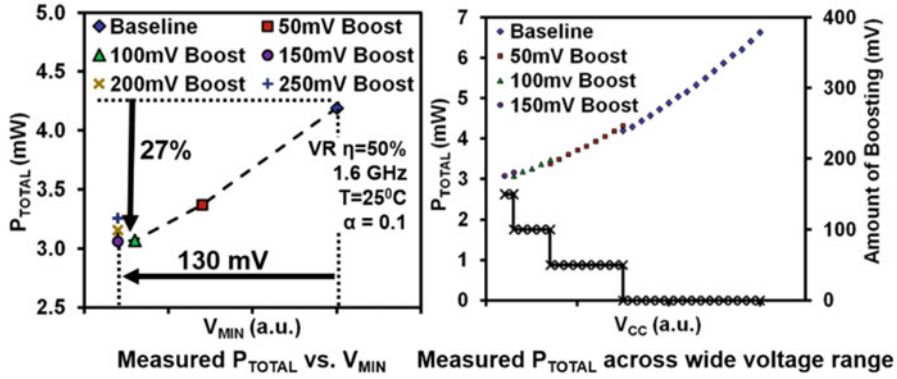


Fig. 5.15 Measured total power vs. V_{MIN} and total power vs. supply voltage

magnitude is increased, the array V_{MIN} is progressively reduced and achieves an optimum of 130 mV improvement for 150 mV of boosting operating at 1.6 GHz, resulting in 27% lower power. Any further boosting does not improve the array V_{MIN} as it is now limited by peripheral circuits, and results in an increase in power dissipation. Operation of the dual-V_{CC} 8T bitcell SRAM across a wide voltage range is achieved by gradually increasing V_{BOOST} value as V_{CC} is scaled down (Fig. 5.15).

5.3.3 Adaptive and Resilient Techniques

In modern microprocessor design, the register file operating voltage (V) and frequency (F) are limited by the delay of the precharge-evaluate read critical path. Furthermore, additional V/F guardbands are applied to account for worst-case dynamic variations such as voltage droops, temperature fluctuations, and aging-induced degradation. However, since most systems usually operate at nominal conditions, the fixed V/F guardbands for infrequent dynamic variations significantly limit the best-achievable performance and energy efficiency. To reduce these guardbands in flip-flop based static CMOS logic units, replica-based approaches such as Tunable Replica Circuits (TRC) have been proposed (Tschanz et al. 2009). In this approach, a set of replica circuits are calibrated to match the critical path pipeline stage delay, and timing errors due to dynamic variations are detected by double-sampling the TRC outputs. The key requirement is that the TRC must always fail before the critical path fails. These circuit techniques are used in conjunction with architecture features to support instruction replay (to recover from a timing error) or dynamic clocking techniques that prevent failure of the critical paths.

The alternative in-situ approach for timing error detection uses Error Detection Sequentials (EDS) in the critical paths of the pipeline stage. Timing errors are detected by a double-sampling mechanism using a flip-flop and a latch (Bowman et al. 2009; Ernst et al. 2003). These error detection techniques, however, cannot be directly used for two-phase precharge-evaluate domino read critical paths in high-performance RF arrays since the data outputs are valid only during the evaluate phase. For error detection in RF arrays, replica-based techniques such as Tunable Replica Bits (TRB) have been proposed (Raychowdhury et al. 2011). In this approach, a set of replica memory bits are tuned at test time so that in the presence of dynamic variations the TRBs fail before the worst-case memory bit fails.

An alternate approach implements in-situ Timing Margin Detector (TMD) and Timing Error Detector (TED) circuits for domino read paths in 8T-bitcell arrays (Kulkarni et al. 2015). TMD circuit enables voltage and frequency adaptation to low-frequency voltage variations, temperature and aging, as well as to excessive

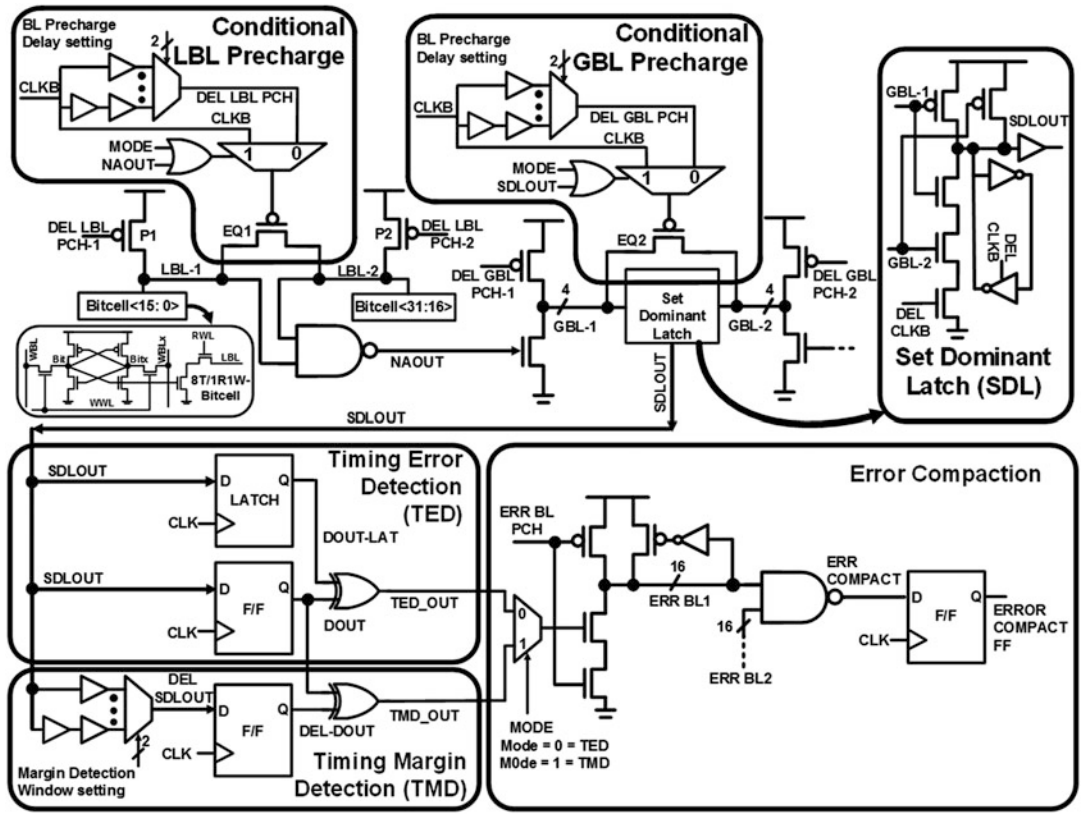


Fig. 5.16 In-situ timing margin and timing error detection along with error compaction circuits for high-performance adaptive and resilient domino register file design

persistent timing errors produced by certain data access patterns. The timing margin is detected by double-sampling the array read output and its delayed version at the same clock edge as shown in Fig. 5.16.

TED circuits enable resiliency to timing errors triggered by local high-frequency voltage droops and nominally random data access in the presence of within-die (WID) delay variations. The timing errors are detected by double-sampling the read output with a flop+latch comparison (Fig. 5.16). The sensing errors in the precharge/evaluate domino read path are converted into timing errors using conditional delayed bitline precharge without affecting the subsequent precharge operation. An error compaction circuit combines the error output from every bit slice to generate a one-bit error-compact signal for the entire register file array. TMD/TED techniques incur 6–13% area overhead and 0.2–0.3% power overhead for a 4 KB sub-array, but this area overhead can be amortized over a larger array size.

Figure 5.17 shows the read timing error measurements at 440 mV under voltage and temperature fluctuations. The read failure measurements are extrapolated to 10^{-6} to estimate the required frequency guardbands for a 1 Mb target array size. With 10% voltage droop, the operating frequency is lowered by 8%. With temperature variation from 100 to 25 °C, the frequency further reduces by 7%. The aging guardband is estimated to be 3% based on a representative ring oscillator delay degradation.

Therefore a baseline design with process, voltage, temperature and aging guardband would operate at 860 MHz at 440 mV. As frequency is pushed higher using the TMD + TED scheme,

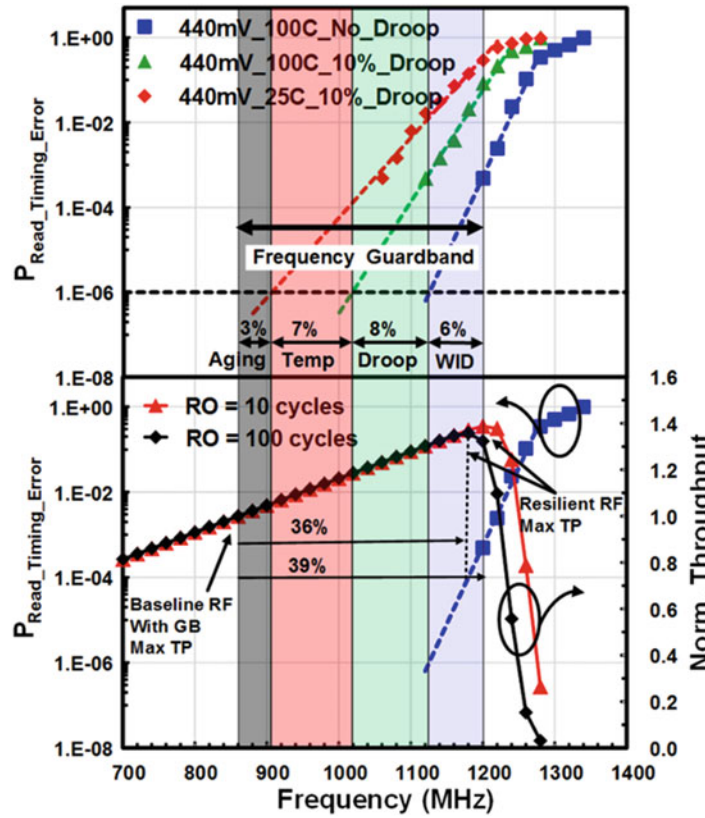


Fig. 5.17 Measured read failures vs. frequency and throughput improvement with adaptive and resilient approach

throughput first increases proportionally, and then peaks when the error rate and the corresponding recovery overheads become too large. As the recovery overhead increases, throughput improvement further reduces. The maximum frequency can be improved by 36–39% using the TMD/TED based V/F adaptation and error recovery via replay, depending on the number of recovery overhead cycles. With TED, frequency can be pushed by 6%, thus compensating for WID variation impacts across random bitcell access patterns.

These 22 nm tri-gate CMOS testchip measurement results demonstrate that the throughput gain is eventually limited by the replay overheads. Maximum achievable throughput gain observed is 21%. For a given throughput (0.9 GOPS) the energy efficiency is improved by 65–67% with a peak energy efficiency 409 GOPS/W.

Figure 5.18 shows the unified framework for the adaptive and resilient static CMOS logic block along with the domino register file operating at the same V_{CC} and frequency. The TRC + EDS address the dynamic variations in the static CMOS logic block. For the domino RF array, TMD addresses the slow variations and generates an output to the V/F adaptation controller. The adaptation controller modulates the clock generator and/or voltage regulator to adapt the voltage and/or frequency applied to the design. On the other hand, TED addresses the fast dynamic variations that occur too quickly to be mitigated by V/F adaptation. The TED circuits give an output to the error response controller which initiates a replay mechanism at increased supply voltage or reduced clock frequency. An error-rate tracker is also included to monitor the error rate. If the error rate exceeds a certain threshold, a signal (ERTe) is sent to the adaptation controller which adapts voltage and frequency to minimize the overheads due to frequent replay mechanisms. In this way,

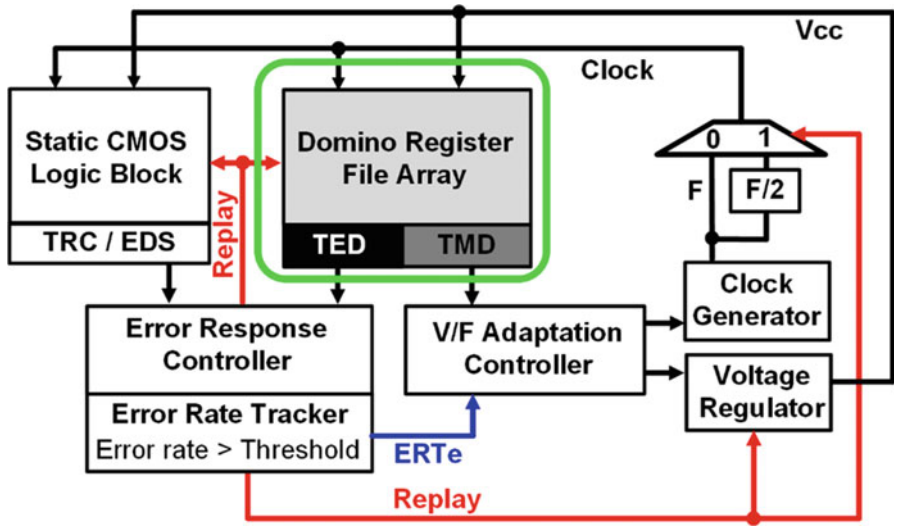


Fig. 5.18 Unified framework for adaptive and resilient logic+register file array operating on same voltage and frequency domain

energy-efficient high-performance 8T SRAM based register files featuring in-situ timing margin and timing error detection circuits can enable an adaptive and resilient framework for the entire block operating on the same voltage and frequency domain.

5.4 10T SRAM for Sub-threshold Operation

Most IoT systems exhibit a burst mode of operation followed by long idle periods. For always-ON blocks performing continuous sensing, minimizing the state retention leakage becomes an important design consideration. For such always-ON blocks, ultra-low voltage memory operation is critical for extending the battery lifetime. Various SRAM bitcells have been explored for sub-threshold region operation.

5.4.1 Sub-threshold SRAMs

For sub-threshold operation, various 6T, 8T bitcells with decoupled read and write operations have been proposed (Zhai et al. 2007; Chang et al. 2007; Verma et al. 2007). Single-ended 10T bitcells are similar to the single-ended 8T bitcell except for the read port configurations as shown in Fig. 5.19. Additional transistors are used to control for the read bitline leakage (Calhoun et al. 2006; Kim et al. 2007). Noguchi et al. have proposed a single-ended transmission gate 10T bitcell (Noguchi et al. 2008) in which the bitcell contents are buffered using an inverter and then transferred to the read bitline whenever the bitcell is accessed. The use of a transmission gate eliminates domino-style read-bitline sensing. Thus read bitline does not require precharge and keeper transistors. Also if the same data is accessed in consecutive read cycles, the charging/discharging of the read-bitline is reduced. A differential 10T bitcell with two separate ports for read-disturb-free operation has also been reported (Noguchi et al. 2008). Chang et al. have proposed a read-disturb-free differential 10T bitcell which is suitable for bit-interleaved architecture (Chang et al. 2008). A similar 10T cell with a column-assist technique is also reported (Okumura et al. 2009). However, series-connected write access transistors degrade the write-ability of the bitcell and require write-assist circuits such as wordline boosting for a successful write operation.

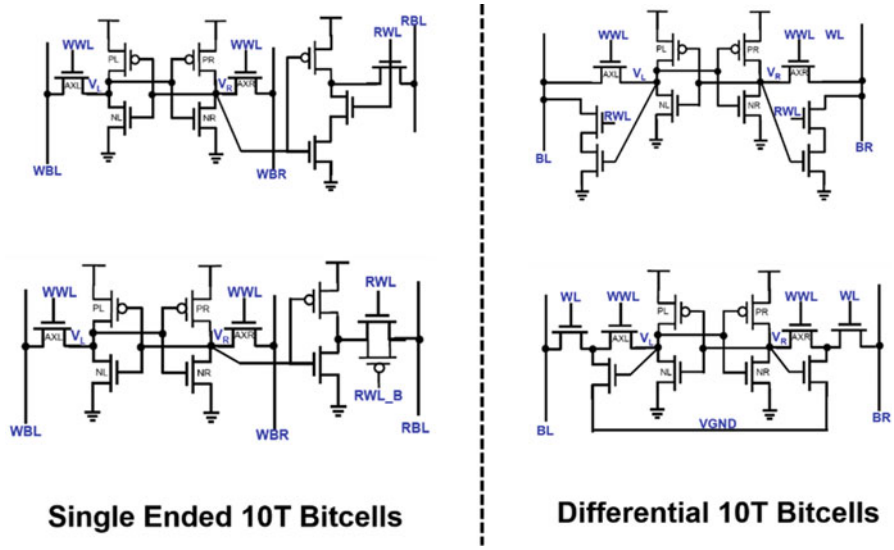


Fig. 5.19 10T sub-threshold SRAM bitcell topologies

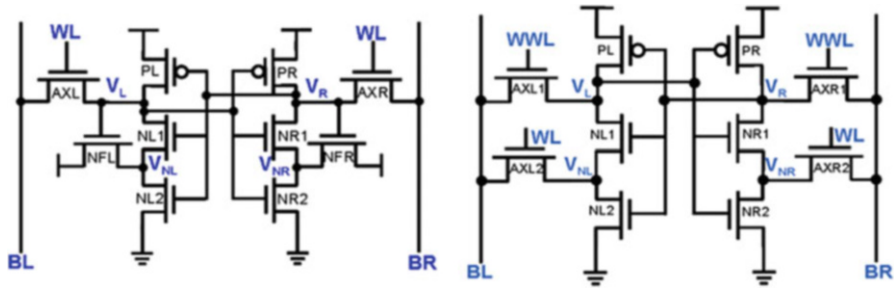


Fig. 5.20 10T Schmitt Trigger-1 (*ST-1 on left*) and Schmitt Trigger-2 (*ST-2 on right*) SRAM bitcell topologies

5.4.2 Schmitt Trigger SRAMs

To improve the stability of the cross-coupled inverter pair for ultra-low voltage state retention in always-On domains of an IoT system, Schmitt Trigger based SRAM topologies have been proposed (Kulkarni et al. 2007; Kulkarni and Roy 2012). A Schmitt trigger is used to modulate the switching threshold of an inverter depending on the direction of the input transition. In the Schmitt trigger SRAM bitcells, the feedback mechanism is used only in the pull-down path.

Figure 5.20 (left side) shows the schematics of Schmitt Trigger-1 (ST-1) bitcell. ST-1 bitcell utilizes differential sensing with 10 transistors, 1 wordline (WL) and 2 bit-lines (BL/BR). Transistors PL-NL1-NL2-NFL form one ST inverter while PR-NR1-NR2-NFR form another ST inverter. Feedback transistors NFL/NFR raise the switching threshold of the inverter during the $0 \rightarrow 1$ input transition enabling the Schmitt Trigger action. The positive feedback from NFL/NFR adaptively changes the switching threshold of the inverter depending on the direction of the input transition. During a read operation, (with $V_L = 0$ and $V_R = 1$, for example) due to voltage divider action between the access transistor and the pull down NMOS, the voltage of V_L node rises. If this voltage is greater than the switching threshold (trip point) of the V_R

inverter, the contents of the cell can get flipped resulting in a read failure event. In order to avoid a read failure, the feedback mechanism should increase the switching threshold of the inverter PR-NR1-NR2. Transistor NFR raises the voltage at node V_{NR} and increases the switching threshold of the inverter storing '1'. Thus Schmitt trigger action is used to preserve the logic '1' state of the memory cell.

Figure 5.20 (right side) shows the schematics of the Schmitt Trigger-2 (ST-2) bitcell utilizing differential sensing with 10 transistors, two wordlines (WL/WWL) and two bit-lines (BL/BR). The WL signal is asserted during read as well as the write operation, while WWL signal is asserted during the write operation. During the retention mode, both WL and WWL are OFF. In the ST-2 bitcell, feedback is provided by separate control signal (WL) unlike the ST-1 bitcell, wherein feedback is provided by the internal nodes. In the ST-1 bitcell, the feedback mechanism is effective as long as the storage node voltages are maintained. Once the storage nodes start transitioning from one state to another state, the feedback mechanism is lost. To improve the feedback mechanism, a separate control signal WL is employed for achieving stronger feedback. This enables robust operation compared to 6T and ST-1 bitcells operating at very low supply voltages.

5.4.3 Static Memory Arrays

Another approach to achieve very low voltage operation is to design the SRAM bitcell to avoid contention during the write operation, and to employ hierarchical read bitline to avoid leakage due to unselected bits in order to achieve high I_{ON}/I_{OFF} ratio. Worst-case process variations make it difficult to satisfy both read and write conditions at sub-threshold voltages. A latch-based write scheme with C^2MOS tri-state inverters can be more effective for sub-threshold operation (Wang and Chandrakasann 2005).

The read operation of the memory in sub-threshold can be challenging due to bitline leakage, where the leakage through the pull-down devices causes the dynamic bitline to drop. Because clock speed is not the key metric in this application, a sense-amplifier-based read-bitline for fast read accesses is not needed. In a conventional dynamic bitline design, the keeper/precharge transistor needs to be upsized to mitigate the leakage due to unselected bitcells. However it reduces effective read current due to lower I_{READ}/I_{OFF} ratio (Fig. 5.21).

Instead, a hierarchical read-bitline which segments the bitline by using a 2-to-1 multiplexers can be used. Segmenting reduces parallel leakage for each level of the hierarchy and the effect of process variations is mitigated.

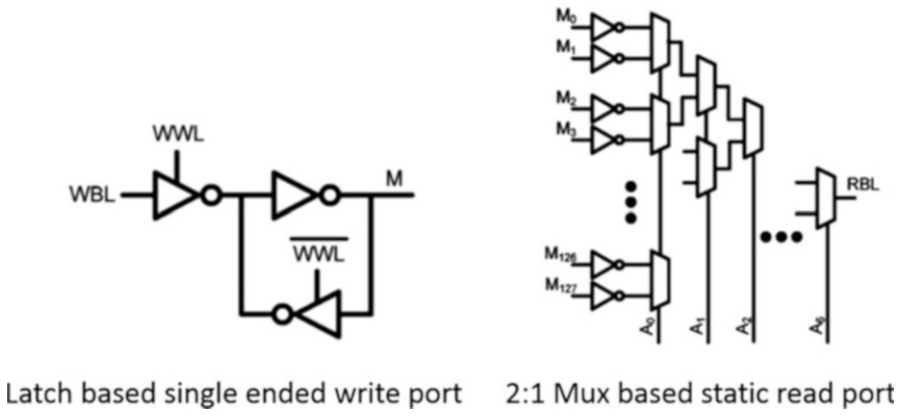


Fig. 5.21 Static memory design using latch bitcell and hierarchical mux based read path

These multiplexers can be designed to avoid stacked devices and sneak leakage paths by inserting inverters between each level of hierarchy.

5.5 SRAMs Using Emerging Technologies

Many alternative technologies such as nanowire Field Effect Transistors (FET), Tunnel FETs, and III–V semiconductor FETs having superior Ion/Ioff characteristics compared to the silicon MOSFTs have been explored for energy efficient low voltage SRAM designs.

5.5.1 Nanowire FET

Nanowire transistors with gate all around structure exhibit significant reduction in V_T variation with lower channel dopant concentration as well as superior short-channel control. Figure 5.22 shows nano-wire FET TEM cross-section and 0.039 μm^2 6T SRAM bitcell SEM image (Chen et al. 2009). Nano-wire FET is fabricated using Nano Injection Lithography (NIL) technique which doesn't use photoresist as well as any masks. This patterning technology utilizes gas-phase reaction activated with finely-controlled electron beam to deposit the desired nanometer scale hard mask for subsequent etching process. Mask-less lithography reduces the number of process steps, while photoresist-free technology is more immune to light/electron interference, thus resulting in less proximity effects and better spacing resolution. A Dynamic V_{DD} Regulator (DVR) is used to improve the bitcell stability. DVR increases bitcell voltage during read operation and lowers the read disturb voltage.

5.5.2 Tunnel FET

The Tunneling Field Effect Transistor (TFET) is a quantum mechanical device in which electron transport is governed by the quantum tunneling across the source-channel junction instead of thermionic barrier modulation as in MOSFETs. Figure 5.23 shows 0.069 μm^2 hybrid 6T SRAM bitcell schematic and SEM fabricated in 65 nm process technology (Nirschl et al. 2005). Pull-down (PD) NMOS devices are formed using TFETs. Pass gate (PG) transistor requires bi directional current conduction capability due to read-0 vs. write-0 scenarios. Hence TFETs are not used for PG but only for PD devices.

An all-TFET SRAM design could be challenging due to the unidirectional current conduction in TFETs, complicating its use as a SRAM access transistor. Dual-port 8T SRAM bitcells

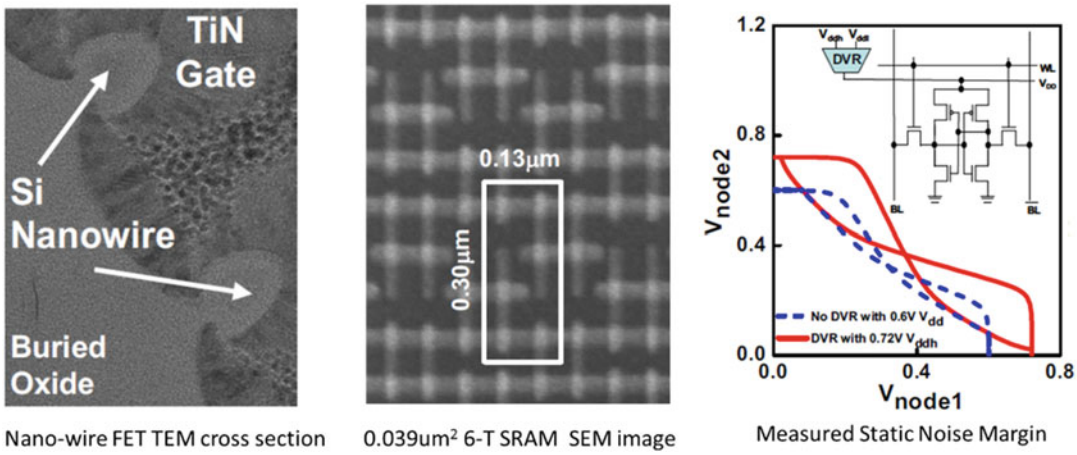


Fig. 5.22 Nanowire FET structure, 6T SRAM bitcell SEM image and read stability analysis

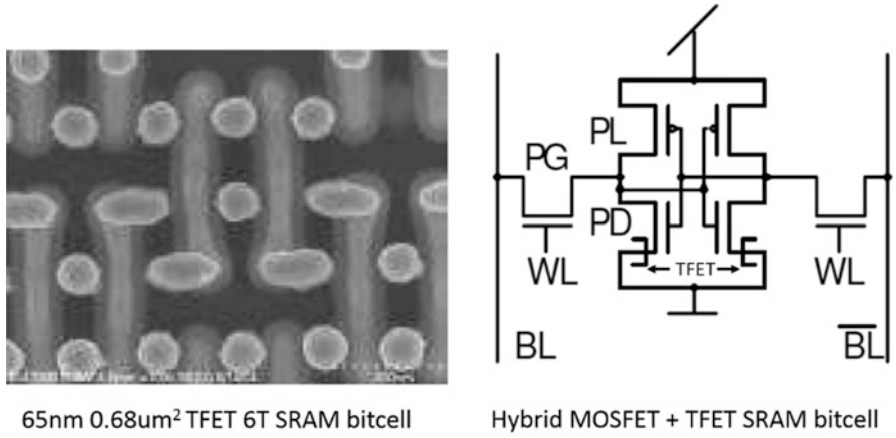


Fig. 5.23 TFET based 6T SRAM schematics and SEM image

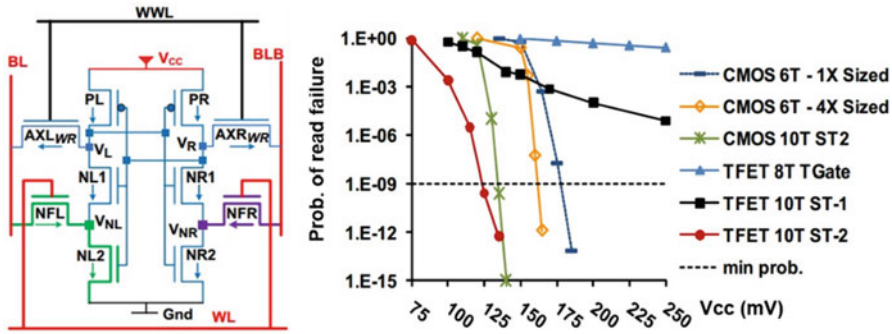


Fig. 5.24 TFET based Schmitt Trigger-2 bitcell and read stability comparison

with two TFET access transistors having opposite current directionality are used for read and write operation. Schmitt Trigger-2 bitcell topology can be realized with TFETs as shown in Figure 5.24 (Saripalli et al. 2011). Statistical simulations show extreme low voltage read operation for TFET based ST-2 bitcell compared to the CMOS based 6T, 10T and TFET dual port 8T bitcells. Therefore, TFET devices with better Ion/Ioff at low supply voltages could be a suitable candidate for ultra-low voltage SRAM designs in always-ON modules of an IoT system.

5.5.3 III–V Quantum Well FET

Compound semiconductor-based quantum well FETs are a promising candidate to realize high-performance yet low-voltage SRAM designs. Figure 5.25 shows Indium Antimonide (InSb) semiconductor having highest electron mobility used to form NFET device. This could be used to form the read port of the hybrid 8T SRAM bitcell (Kulkarni and Roy 2008). The higher Ion current in InSb can be leveraged to improve the read bitline delay up to 60% compared to the Silicon CMOS based 8T SRAM.

5.5.4 Monolithic 3-D Integration

The demand for higher SRAM density is continually growing due to increased design complexity (application processor, modem, and graphics on the same die) and also due to increased die cost in dimensionally-scaled advanced CMOS

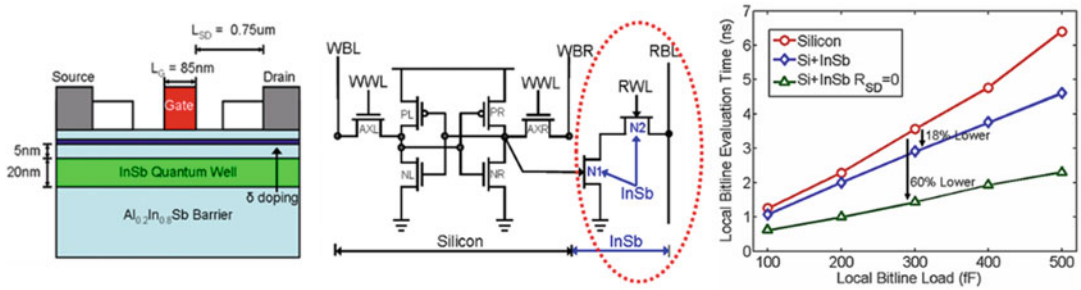


Fig. 5.25 InSb based quantum well FET structure, hybrid Si+InSb 8T SRAM bitcell and bitline delay comparison

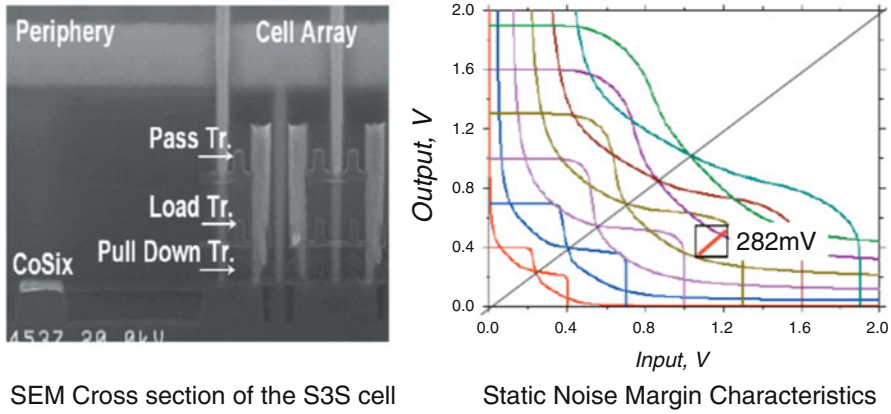


Fig. 5.26 Monolithic SRAM bitcell SEM image and read stability measurements

technologies. Towards this goal, many approaches for monolithic 3-D integration of SRAM bitcells have been reported. Figure 5.26 shows the scanning electron microscope (SEM) cross section of Stacked Single crystal Silicon (S^3) SRAM cell fabricated using 65 nm process technology achieving 0.16 μm^2 bitcell area showing almost 3× better bitcell density compared to the conventional 65 nm planar bitcell (Jung et al. 2005). Load PMOS and pass gate NMOS transistors are stacked on the planar pull-down transistors in different layers. The Static Noise Margin (SNM) achieved is 282 mV at $V_{CC} = 1.2$ V operation.

5.6 Summary

In this chapter we have shown that there are a wide variety of circuit and device techniques that can be applied for energy-efficient, low-voltage memory arrays. These techniques improve the dynamic operating range of the memory by enhancing the periphery to apply optimum access voltages (dynamic assist, dual-V_{CC} arrays), or optimizing the bitcell itself for low-voltage operation (10T bitcells, static memory arrays), or by embedding resiliency techniques to dynamically detect and respond to variations. These techniques differ widely in terms of area overhead, complexity, ease of integration, and power and voltage reduction that can be achieved. Therefore it is important to analyze the unique requirements for each memory array in a particular application before selecting the best implementation for that array (Table 5.1).

In this section we give a high-level summary of the key metrics for the techniques which have been detailed in this chapter. In addition to these circuit techniques, emerging technologies (such as nanowires, tunnel FETs, and quantum well FETs) can also provide power/performance

Table 5.1 Qualitative comparison of various 6T, 8T, 10T SRAM V_{MIN} and guardband reduction techniques

Technique	Ref.	Array type	Key goal	Overhead[a]	Example results[a]
Read/write assist	Khellah et al. (2008) and Mann et al. (2010)	6T SRAM	V_{MIN} reduction	Varies	Varies
Wordline boosting: charge pump	Raychowdhury et al. (2010)	8T SRAM	V_{MIN} reduction	~25% area	120–140 mV V_{MIN} reduction
Wordline boosting: capacitive coupling	Kulkarni et al. (2010)	8T SRAM	V_{MIN} reduction	5–11% area	140 mV write V_{MIN} reduction
Dual-Vcc design	Kulkarni et al. (2013)	8T SRAM	V_{MIN} reduction	Extra supply rail	130 mV V_{MIN}, 27% lower power
Subthreshold SRAMs	Verma et al. (2007); Calhoun et al. (2006); and Kim et al. (2007)	10T SRAM bitcells	V_{MIN} reduction	~50% area	Sub-500 mV operation
Schmitt trigger SRAM	Kulkarni et al. (2007) and Kulkarni and Roy (2012)	10T SRAM bitcells	V_{MIN} reduction	~100% area	~100 mV lower V_{MIN}
Static memory	Wang and Chandrakasann (2005)	Latch-based bitcell	V_{MIN} reduction	>2–3× area	Sub-500 mV operation
Resiliency: tunable replica bits	Raychowdhury et al. (2011)	6T, 8T, 10T	V_{MIN}, guardband reduction	~5% area	9% V_{MIN}, 7.5% power reduction
Resiliency: in-situ timing monitors	Kulkarni et al. (2015)	8T	V_{MIN}, guardband reduction	6–13% area	~36% frequency increase

[a]Note: overheads and results for these techniques are strongly dependent on process technology, array design characteristics, optimization target, etc. The numbers shown here are meant to be only representative examples

improvements at the low voltages required for many IoT devices.

5.7 Trends and Perspectives

The IoT domain presents both a challenge and opportunity for advances in low-power, low-voltage memory arrays. Power requirements for many IoT devices are extremely low, requiring memory arrays with low standby power as well as low access (read and write) power. Voltage scaling of traditional SRAM arrays has slowed or even stopped with technology scaling, which points to the need for advanced circuit techniques for energy efficient and reliable on-die memory for the IoT space. At the same time, new and emerging workloads of special importance to IoT-for example, always-on audio and visual recognition workloads, cognition, etc.-require ever-increasing amounts of memory storage and bandwidth. Clearly, advancements in memory design are needed to enable these usages for IoT.

The techniques discussed here also point to the need for cross-layer optimization to reap maximum benefit for these advanced memory techniques while minimizing cost. IoT SoCs can contain hundreds of embedded memory arrays, each with its own unique size, arrangement, access pattern, and power/performance requirements. Detailed architecture and system simulations are needed early in the design phase to determine the optimum method for implementing each memory array. Dynamic assist or multi-voltage techniques require area-efficient and energy-efficient power delivery circuits such as on-die voltage generators, charge pumps, and dynamic power management. Resiliency techniques that detect timing margins or timing errors show the promise to drastically reduce array operating voltages, but require architecture support (such as instruction replay) to obtain the full benefit. Looking further towards the future, dramatic gains in memory energy efficiency can be obtained by over-scaling the voltage and allowing a small number of bits to fail. Traditionally, these failures are either

avoided or handled via error-correction techniques such as ECC for large memory arrays. However, there are a class of emerging workloads such as convolutional neural networks that can tolerate infrequent errors in certain computations. These types of workloads, coupled with the necessary resiliency and monitoring techniques to ensure memory operation within the desired bit error range, could provide the next big gains in energy efficient operation.

Acknowledgments Authors would like to thank Muhammad Khellah, Arijit Raychowdhury, Keith Bowman, Carlos Tokunaga, Dinesh Somasekhar, Bibiche Geuskens, Tanay Karnik, and Shekhar Borkar for insightful technical discussions. Authors would also like to thank Greg Taylor, Richard Forand and Matthew Haycock for encouragement and support. This research was, in part, funded by the U.S. Government (DARPA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

A. Bhavnagarwala, X. Tang, J. Meindl, The impact of intrinsic device fluctuations on CMOS SRAM cell stability. IEEE J. Solid State Circ. **36**(4), 658–665 (2001)

K. Bowman, J. Tschanz, C. Wilkerson, S.-L Lu, T. Karnik, V. De, S. Borkar, Circuit techniques for dynamic variation tolerance, in *Proceedings of the 46th Annual Design Automation Conference* (2009), pp. 4–7

B.H. Calhoun, A.P. Chandrakasan, A 256 kb Sub-threshold SRAM in 65 nm CMOS, in *Proceedings of the International Solid State Circuits Conference* (2006), pp. 628–629

L. Chang, Y. Nakamura, R.K. Montoye, J. Sawada, A.K. Martin, K. Kinoshita, F.H. Gebara, K.B. Agarwal, D.J. Acharyya, W. Haensch, K. Hosokawa, D. Jamsek, A 5.3 GHz 8T-SRAM with operation down to 0.41 V in 65 nm CMOS, in *Proceedings of the VLSI Circuit Symposium* (2007), pp. 252–253

I. Chang, J.-J. Kim, S. Park, K. Roy, A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS, in *Proceedings of the International Solid State Circuits Conference* (2008), pp. 628–629

H.-Y. Chen, C.-C. Chen, F.-K. Hsueh, J.-T. Liu, C.-Y. Shen, C.-C. Hsu, S.-L. Shy, B.-T. Lin, H.-T. Chuang, C.-S. Wu, C. Hu, C.-C. Huang, F.-L. Yang, 16 nm functional 0.039 μm^2 6T-SRAM cell with nano injection lithography, nanowire channel, and full TiN gate, in *Proceedings of International Electron Device Meeting (IEDM)* (2009), pp. 958–960

D. Ernst, N.S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, T. Mudge, Razor: a low-power pipeline based on circuit-level timing speculation, *IEEE/ACM MICRO-36* (2003), pp. 7–18

S.-M. Jung, Y. Rah, T. Ha, H. Park, C. Chang, S. Lee, J. Yun, W. Cho, H. Lim, J. Park, J. Jeong, B. Son, J. Jang, B. Choi, H. Cho, K. Kim, Highly cost effective and high performance 65 nm S^3 (Stacked Single-crystal Si) SRAM technology with $25F^2$, 0.16 μm^2 cell and doubly stacked SSTFT cell transistors for ultra high density and high speed applications, *Symposium on VLSI Technology* (2005), pp. 220–221

M.M. Khellah, A. Keshavarzi, D. Somasekhar, T. Karnik, V. De, Read and write circuit assist techniques for improving Vccmin of dense 6T SRAM cell, in *Proceedings of the International Conference on Integrated circuit Design and Technology* (2008), pp. 185–189

T.-H. Kim, J. Liu, J. Keane, C.-H. Kim, A high-density subthreshold sram with data-independent bitline leakage and virtual ground replica scheme, in *Proceedings of the International Solid State Circuits Conference* (2007), pp. 330–331

J.P. Kulkarni, K. Roy, Technology circuit co-design for ultra-fast InSb quantum well transistors. IEEE Trans. Electron Dev. **55**, 2537–2545 (2008)

J.P. Kulkarni, K. Roy, Ultralow-voltage process-variation-tolerant schmitt-trigger-based SRAM design. IEEE Trans. VLSI Syst. **20**(2), 319–332 (2012)

J.P. Kulkarni, K. Kim, K. Roy, A 160 mV robust schmitt trigger based sub-threshold SRAM. IEEE J. Solid State Circ. **42**(10), 2304–2313 (2007)

J. Kulkarni, B. Geuskens, T. Karnik, M. Khellah, J. Tschanz, V. De, Capacitive-coupling wordline boosting with self-induced VCC collapse for write V_{MIN} reduction in 22-nm 8T SRAM, *International Solid State Circuits Conference (ISSCC)* (2010), pp. 234–235

J.P. Kulkarni, M. Khellah, J. Tschanz, B. Geuskens, R. Jain, S. Kim, V. De, Dual-Vcc 8T-bitcell SRAM array in 22 nm Tri-gate CMOS for energy efficient operation across wide dynamic voltage range, *VLSI Circuit Symposium (VLSI Symp)* (2013), pp. C126–C127

J. P. Kulkarni, C. Tokunaga, P. Aseron, T. Nguyen Jr., C. Augustine, J. Tschanz, V. De, A 409 GOPS/W adaptive & resilient domino register file in 22 nm Tri-Gate CMOS featuring in-situ timing margin & error detection for tolerance to within-die variation, voltage droop, temperature & aging, *International Solid State Circuits Conference (ISSCC)* (2015), pp. 82–83

R.W. Mann, J. Wang, S. Nalam, S. Khanna, G. Braceras, H. Pilo, B.H. Calhoun, Impact of circuit assist

methods on margin and performance in 6T SRAM. Solid State Electron. **54**(11), 1398–1407 (2010)

S. Mukhopadhyay, H. Mahmoodi, K. Roy, Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS. IEEE Trans. Comput. Aided Des. **24**(12), 1859–1880 (2005)

H. Noguchi, S. Okumura, Y. Iguchi, H. Fujiwara, Y. Morita, K. Nii, H. Kawaguchi, M. Yoshimoto, Which is the best dual-port SRAM in 45-nm process technology?—8T, 10T single end, and 10T differential, *IEEE International Conference on Integrated Circuit Design and Technology* (2008), pp. 55–58

Th. Nirschl, St. Henzler, J. Fischer, A. Bargagli-Stoffi, M. Fulde, M. Sterkel, P. Teichmann, U. Schaper, J. Einfeld, C. Linnenbank, J. Sedlmeir, C. Weber, R. Heinrich, M. Ostermayr, A. Olbrich, B. Dobler, E. Ruderer, R. Kakoschke, K. Schrüfer, G. Georgakos, W. Hansch, D. Schmitt-Landsiedel, The 65 nm tunneling field effect transistor (TFET) 0.68 μm^2 6T memory cell and multi-V_{th} device, *35th European Solid-State Device Research Conference* (2005), pp. 173–176

S. Okumura, Y. Iguchi, S. Yoshimoto, H. Fujiwara, H. Noguchi, K. Nii, H. Kawaguchi, M. Yoshimoto, A 0.56-V 128 kb 10T SRAM using Column Line Assist (CLA) scheme, in *Proceedings of the International Symposium on Quality Electronics Design (ISQED)* (2009), pp. 659–663

A. Raychowdhury, B. Geuskens, J. Kulkarni, J. Tschanz, K. Bowman, T. Karnik, S.-L. Lu, V. De, M. Khellah, PVT & Aging Adaptive Word-Line Boosting for 8T SRAM Power reduction, *International Solid State Circuits Conference (ISSCC)* (2010)

A. Raychowdhury, B. Geuskens, K. Bowman, J. Tschanz, S.-L. Lu, T. Karnik, M. Khellah, V. De, Tunable replica bits for dynamic variation tolerance in 8T SRAM arrays, IEEE Journal of Solid State Circuits, **46**(4), (2011)

V. Saripalli, S. Datta, V. Narayanan, J.P. Kulkarni, Variation-tolerant ultra low-power hetero-junction tunnel FET SRAM Design, *7th International Symposium on Nanoscale Architectures (NANOARCH)* (2011), pp. 45–52

Semiconductor Industry Association (SIA) and Semiconductor Research Corporation (SRC) report on "Rebooting the IT revolution" (2015)

J. Tschanz, K. Bowman, S. Walstra, M. Agostinelli, T. Karnik, V. De, Tunable replica circuits and adaptive voltage-frequency techniques for dynamic voltage, temperature, and aging variation tolerance, *Digest of Technical Papers, Symposium on VLSI Circuits* (2009), pp. 112–113

N. Verma, A.P. Chandrakasan, 65 nm 8T Sub-Vt SRAM employing sense-amplifier redundancy, in *Proceedings of International Solid State Circuits Conference* (2007), pp. 328–329

A. Wang, A. Chandrakasann, A 180-mV subthreshold FFT processor using a minimum energy design methodology. IEEE J. Solid State Circ. **40**(1), 310–319 (2005)

Y. Wang, Robust SRAM Design in Nanoscale CMOS Circuit and Technology, *ISSCC Forum on embedded memories* (2012)

N. Yoshinobu, H. Masahi, K. Takayuki, K. Itoh, Review and future prospects of low-voltage RAM circuits. IBM J. Res. Dev. **47**(5/6), 525–552 (2003)

B. Zhai, D. Blaauw, D. Sylvester, S. Hanson, A sub-200 mV 6T SRAM in 0.13 μm CMOS, in *Proceedings of the International Solid State Circuits Conference* (2007), pp. 332–333