Challenges and Directions for Low-Voltage SRAM

Masood Qazi, Mahmut E. Sinangil, and Anantha P. Chandrakasan
Massachusetts Institute of Technology

Editor's note:
SRAMs capable of operating at extremely low supply voltages—for example, below the transistor threshold voltage—can enable ultra-low-power battery-operated systems by allowing the logic and memory to operate at the same optimal supply voltage. This review article presents SRAM techniques including new bit cells, novel sensing schemes, and read/write assist circuits for ultra-low-power applications.

—Chris H. Kim, University of Minnesota

SRAM IS THE most common embedded-memory option for CMOS ICs. As the supply voltage of low-power ICs decreases, it must remain compatible with the operating conditions. At the same time, increasingly parallel architectures of such low-power systems demand more on-chip cache to effectively share information across parallel processing units. Finally, supply voltage scaling improves the energy consumed by SRAM and dramatically reduces its leakage power.

Achieving low-voltage operation in SRAM faces a confluence of challenges, originating from process variation, and related to bit cell stability, sensing, architecture, and efficient CAD methodologies. The trend toward increased quantity of embedded SRAM in scaled technology compounds the specific need of SRAM in low-power systems. Integrating more memory on chip provides an effective means to use silicon because of memory's lower power density, layout regularity, and performance and power benefits from reduced off-chip bandwidth. As a result, the ever-increasing integration of embedded SRAM continues.[1]

Current solutions to low-voltage SRAM have shown promise. As the data in Figure 1 reveals, a voltage-scalable 8-Kbyte SRAM operates from 0.25 V to 1.2 V.[2] By employing some of the design innovations discussed in this article, this memory design minimizes energy per access and achieves a 50× reduction in leakage power by trading off performance. Indeed, low-power systems benefit from SRAMs that function at very low voltage in the state of the art, but such design solutions of low-voltage SRAM significantly impact area and performance.[3] Reducing this area overhead and further improving the metrics of energy per accessed bit and leakage power will enable new opportunities for low-power electronics in mobile platforms. Wearable electronics, portable medical monitors, and implantable medical devices are some of the applications requiring the storage of significant quantities of information (e.g., patient data), low-access energy caches, and a long operating lifetime from a battery.

In this article, we discuss the challenges to embedded-SRAM design, with particular emphasis on the factors that limit the minimum operating supply voltage V_{min}. We also explore various design solutions and discuss open areas of investigation.

Challenges

The workhorse of embedded memory is SRAM based on the 6T (six-transistor) cell, shown in Figure 2a. From this cell, a subarray is assembled by tiling memory cells into a grid with wordlines (WLs) running horizontally and bitlines (BLs) running vertically, as in Figure 3. Each memory cell is associated with one or more WLs and one or more BLs. Nominally, the bit cell supplies and well biases are globally connected to static voltage sources. During read, the WL voltage V_{WL} is raised, and the memory cell discharges either BLT (bitline true) or BLC (bitline complement), depending on the stored data on nodes Q and bQ. A sense amplifier converts the differential signal to a logic-level output. Then, at the end of the read cycle, the BLs return to the positive supply rail. During write, V_{WL} is raised and the BLs are forced to

0740-7475/11/$26.00 © 2011 IEEE Copublished by the IEEE CS and the IEEE CASS

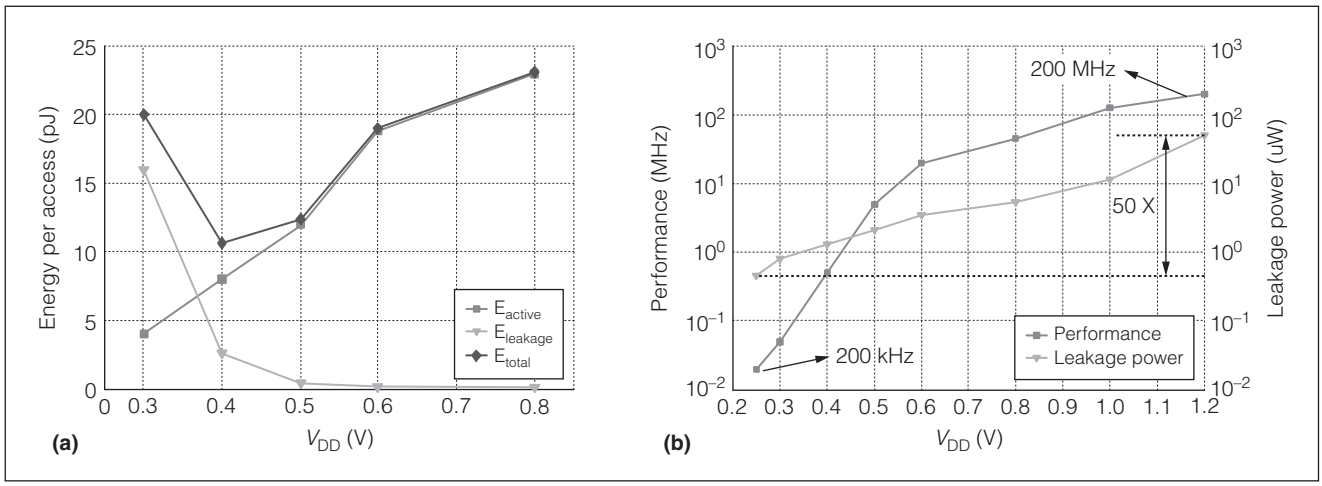

Figure 1. Scaling SRAM supply voltage improves both energy consumption (a) and leakage power (b) at the expense of performance.[2]

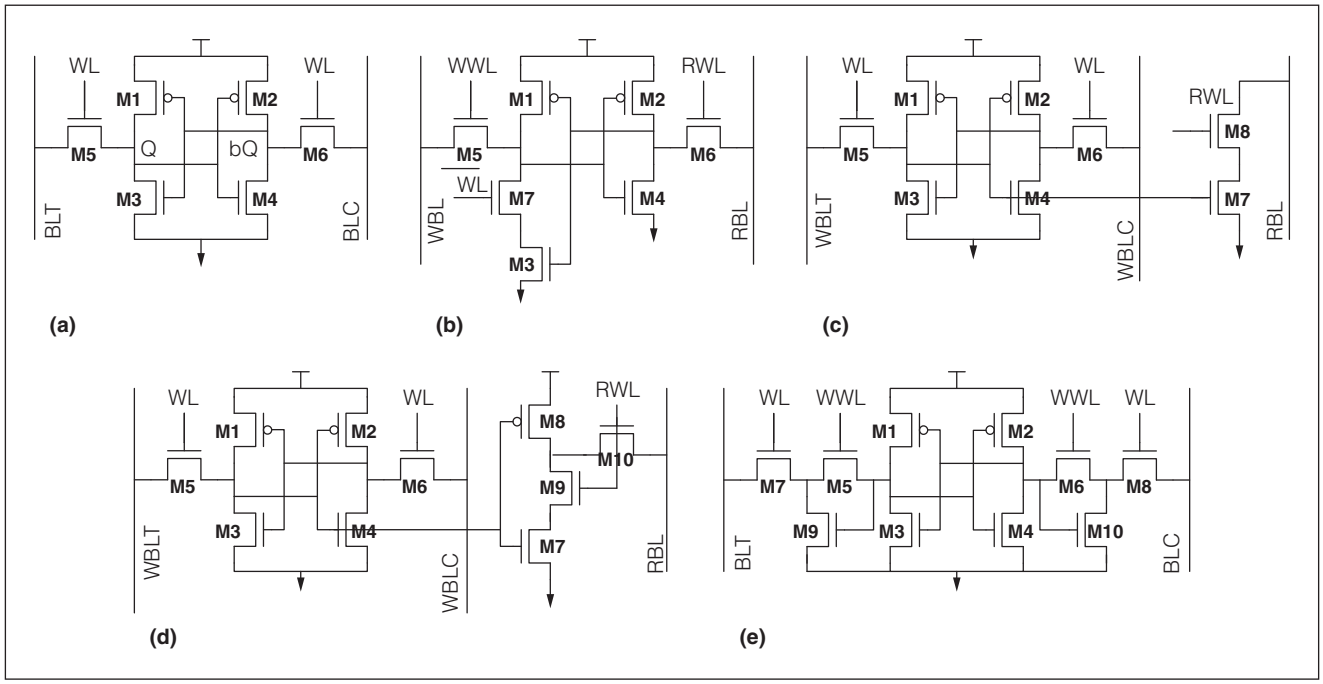

Figure 2. Several SRAM bit cell options: conventional 6T (six-transistor) (a), 7T (Source: Takeda et al.[4]) (b), 8T (Source: Chang et al.[5]) (c), 10T single-ended (Source: Calhoun and Chandrakasan.[6]) (d), and 10T differential (Source: Chang et al.[7]) (e).

either V_{DD} or V_{SS} (depending on the data), overpowering the contents of the memory cell. During hold, V_{WL} is held low and the BLs are left floating or driven to V_{DD}.

The stability of the 6T memory cell can be verified by examining its butterfly curve, which contains the voltage transfer characteristics (VTCs) of the two inverters formed by devices M1, M3, M5 and devices M2, M4, M6, respectively, in Figure 2a. The input-output relations from bQ to Q and from Q to bQ are plotted on the same set of axes, assuming the BLs are driven by DC voltage sources, as in Figures 4a through 4c. During read or hold, three roots of intersection are desired, indicating bistability. During write, only one root of intersection is desired, so that the cell will deterministically flip to one of the two data states, as set by the BL polarity. The severity

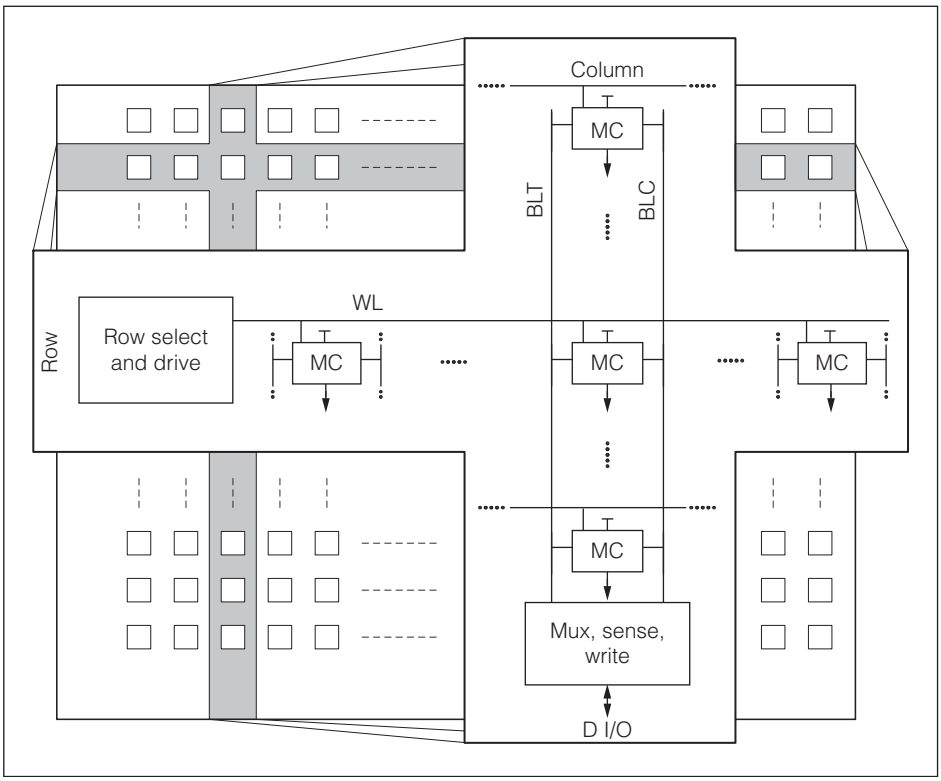

Figure 3. General architecture for an SRAM subarray.

of local variation among devices in the memory cell, primarily from threshold voltage fluctuation in transistors, threatens the capability to preserve stability across 10^6 to 10^9 memory cells within one die.[9]

The effect of this variation on SRAM stability is evident in the transition from Figures 4a-4c to Figures 4d-4f. Reducing V_{DD} from 0.9 V to 0.6 V in the simulation of an SRAM cell in 32-nm predictive technology[8] reveals a dramatic degradation of read and write butterfly curves. Some outlier read butterfly curves will fail to preserve a 0 when device M5 becomes too strong, relative to M3, and the trip point of the opposite inverter shifts toward 0 V from a weakened M2 and a strengthened M4. Also, some outlier write butterfly curves will fail to successfully write from 1 to 0 under complementary conditions of a strong M1, weak M5, and strong M4. The root cause is related to the fact that the current through a transistor is proportional to $(V_{GS} - V_{TH})$ when $V_{GS} > V_{TH}$, and is approximately proportional to $10^{(V_{GS}-V_{TH})/100mV}$ when $V_{GS} < V_{TH}$. Thus, the impact of V_{TH} fluctuation reduces with increasing power supply but becomes intolerable at low supply voltages.

At 0.6 V, the hold butterfly curves remain robust. Because the pass transistors, M5 and M6, are off, the logic-level outputs of the inverters easily reach the supply rails, guaranteeing a wide separation of the bistable points from the metastable point. The hold characteristics will eventually fail when the supply is reduced to the subthreshold regime, where even inverters cannot drive a strong 0 or 1.

Single-event-upsets from radiation also corrupt data in SRAM. When alpha particles from packaging materials or neutrons from space penetrate a silicon wafer, they can generate a charge that perturbs the state nodes of a memory element, causing it to flip. This failure rate increases with a reduction of supply voltage because of the decrease in stored charge on internal nodes.[10] To address these soft errors, an SRAM can be protected with an error-correcting code (ECC), requiring extra memory bits in each word and adding latency to both write access (for encoding) and read access (for detection and correction). If more than 1 bit must be corrected, ECC complexity increases significantly. Therefore, multibit errors from soft-error phenomena are avoided by interleaving multiple words onto the same physical row.[11] For example, a row of 128 adjacent bits comes from eight 16-bit words interleaved. Therefore, physical multibit errors show up as multiple single-bit errors in different words.

Assuming the bit cell preserves its state under all the aforementioned contexts, sensing remains a challenge. The separation in current between a worst-case *on* BL and a worst-case *off* BL poses a fundamental barrier to the sensing margin. This margin can be improved by shortening the BLs at the expense of area efficiency. For single-ended sensing, an accessed cell storing 0 can produce a false 1 on the output of the sense amplifier before a cell storing 1 can produce a true 1 on the output. For differential sensing, the voltage difference between the 0 BL and the 1 BL might not overcome the offset of the associated sense amplifier. Timing variation in the periphery worsens this problem. SRAMs with sense amplifiers

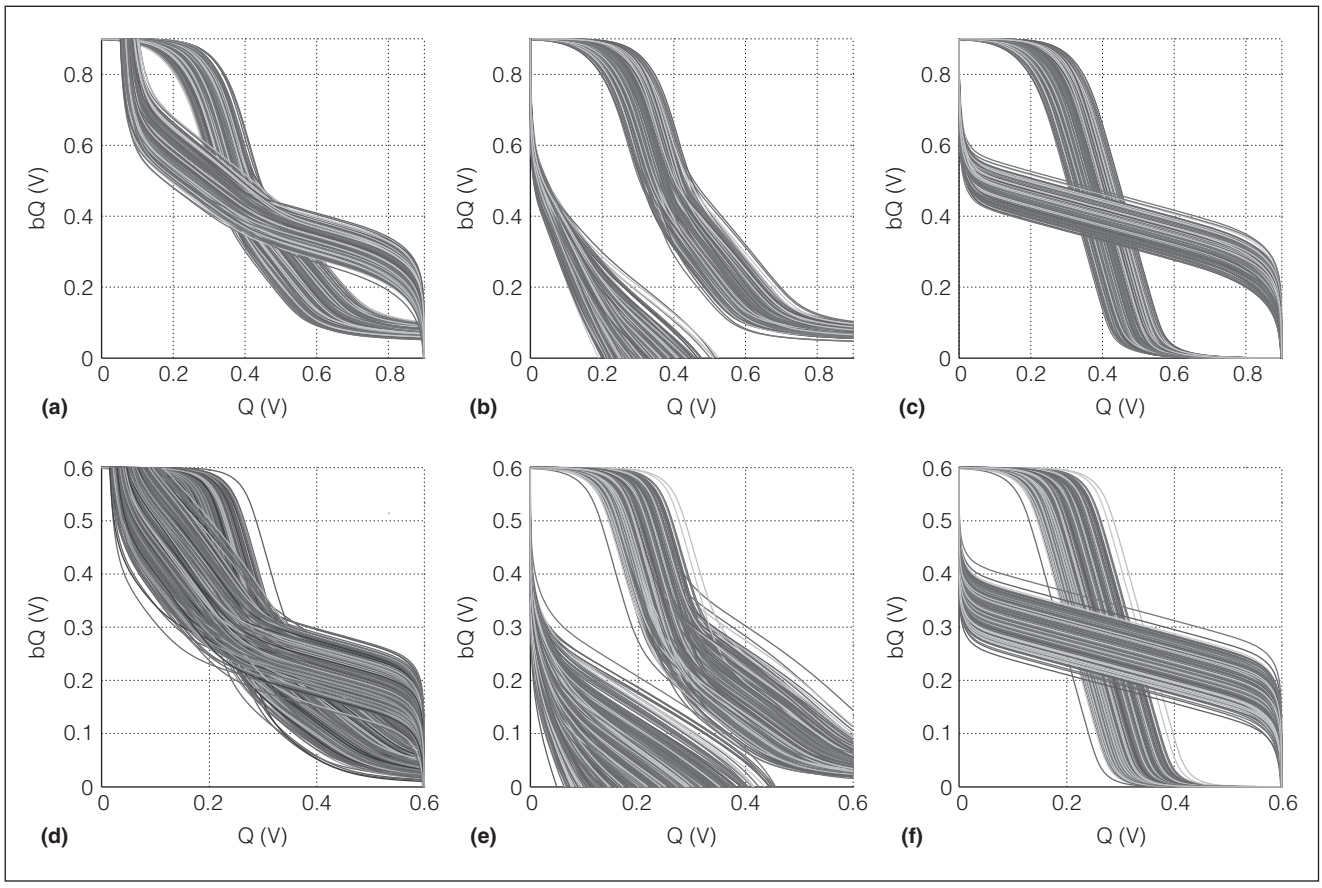

Figure 4. Statistical butterfly curves simulated with local variation at the nominal process corner in 32-nm predictive technology[8]: read (a), write (b), and hold stability (c) at 0.9 V; and at 0.6 V [(d)–(f) respectively].

require a separate signal that bounds the maximum BL signal delay to ensure correct evaluation. This condition is difficult to preserve across process corners, operating temperatures, and low supply voltages.[12]

Finally, a low-power memory that degrades overall bit density in terms of Mbits/mm^2 will increase system cost and potentially reduce the range of applicability because of economic factors. For area efficiency, longer BLs and WLs are chosen. But, beyond a certain point, the penalty of power and delay motivates the partitioning of the memory into multiple subarrays. Within a subarray, the area of a sense amplifier is typically amortized over multiple memory columns to improve efficiency. Therefore, a 32-bit word memory is not restricted to a WL of only 32 columns. This amortization is becoming increasingly important because larger sense amplifiers can mitigate the problem of offset from variation in scaled CMOS. As with soft-error immunity, bit interleaving is an important feature, which in this case enables the optimization of an SRAM architecture for area efficiency. Bit interleaving is made possible by 6T SRAM because read-stable memory cells in unselected columns can have their BLs floating at V_{DD}, corresponding to a dummy read condition that, by construction, will not upset the data.

In all cases, SRAM eventually fails as the supply voltage decreases: bit cells become unreadable or unwritable, radiation-induced soft errors accelerate, sense amplifiers go bad, timing-control signals deviate, and the BL signal vanishes. The limit on the minimum operating supply voltage V_{min} is critical. Therefore, circuit design techniques to lower V_{min} are actively under development.

Solutions

Creating a low-power memory from 6T SRAM has been difficult. Various circuit design solutions try to solve the problems highlighted in the previous section. *Circuit assists* (modifications to the peripheral circuitry that drive WLs, BLs, and power supplies) have been employed to expand the operating margin.

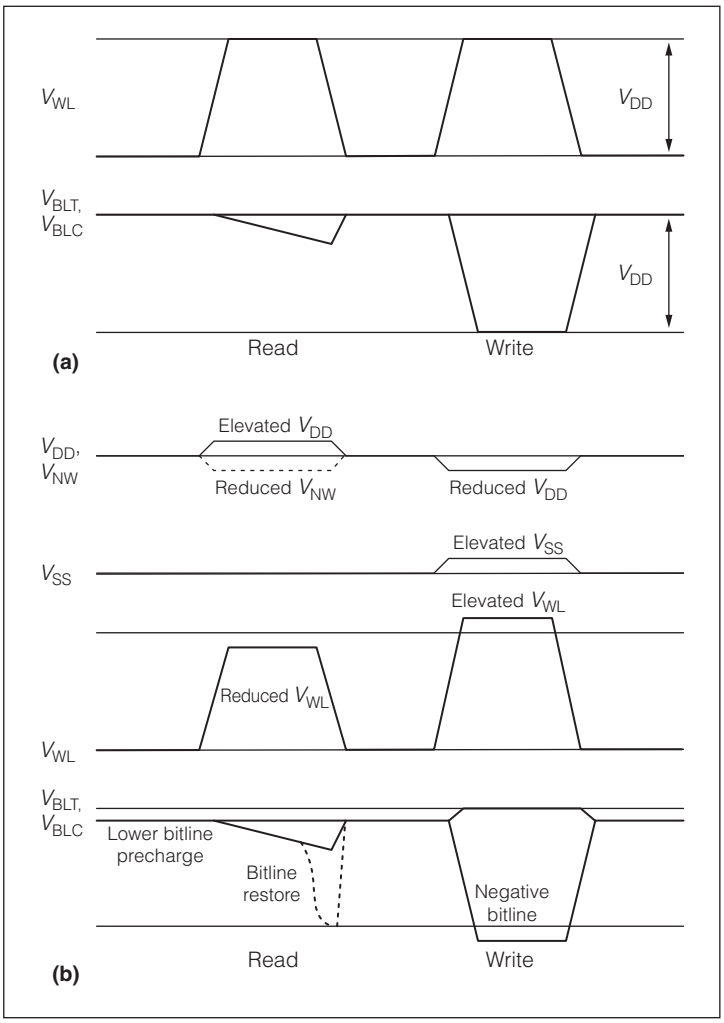

Figure 5. Standard digital operating waveforms for 6T SRAM (a) and various dynamic peripheral-assist possibilities (b).

Another class of approaches abandons the 6T cell, adding transistors appropriately to bypass the problem of the doubly ratioed contention between write and read stability. In conjunction with bit cell stability, new sensing techniques have better handled global and local process variation in the signal path. For standby, techniques to aggressively scale the supply voltage to the statistical limit on hold stability have greatly reduced power. Finally, the circuit optimization for these designs must capture the statistical worst-case realization of local mismatches at the worst-case global corner. Therefore, CAD tools and methods for fast statistical SRAM design have emerged.

Peripheral assists

Because the constraints on relative bit cell device strengths depend on the operation mode, dynamic peripheral circuit assists have been developed to widen the statistical operating margin, thereby reducing V_{min}. Figure 5 illustrates the possible modulations of the different cell terminal voltages—V_{DD}, V_{SS}, V_{BLT} and V_{BLC}, V_{WL}, V_{NW} (n-well for PMOS body), V_{PW} (p-well for NMOS body)—to improve either read or write. Experimental techniques have been developed for each of these terminals (see Table 1). For example, the design by Zhang et al. reduces V_{DD} below the WL and BL levels on a column of bit cells when they are written,[13] but raises V_{DD} above the WL and BL levels on the same bit cells when they are under read stress.

Many of the assists, particularly those modulating horizontal signals, are incompatible with bit interleaving. To recover bit interleaving, a read-modify-write scheme can be employed, in which every column has a sense network so that a read operation can precede every write operation. Then, unselected columns are written back with the original data, while selected columns are written with new data. Unfortunately, this scheme degrades performance, area efficiency, and power.

Static biasing has also been employed. For example, the WL voltage can be suppressed to enhance read stability.[14] Nho et al. have extended this static-biasing technique to adaptively suppress the WL voltage only to the dies that require it.[22] Yamaoka et al. have employed static body bias on both PMOS and NMOS devices in the cell array to recenter the global shift in device strength in order to rebalance the write and read failure rates.[24] Alternatively, a conservative approach to yield adds an extra off-chip power supply for the bit cells (and optionally the WLs) to improve the stability and performance yield of SRAM to the point at which the CMOS logic limits yield.[23]

Finally, investigation of the dynamics of read disturbance reveal that it's possible to produce a functional memory from bit cells that exhibit failing butterfly curves under read stability. Pilo et al. directly connect sense amplifiers to every BL pair so that a full-logic level is restored during a read operation.[20] Therefore, bits that have flipped will be written back to the correct data if enough correct signal is generated before data corruption. Cosemans, Dehaene, and Catthoor buffer short local BLs to long global BLs so that the local BLs collapse before the peak cell current, and hence peak cell disturbance, is established.[25]

Table 1. Description of various bit cell assists.

Terminal	Description	Limitations
Dynamic modulation of vertically routed V_{DD}	Switch between supplies to obtain $V_{DD} - V_{WL} > 0$ for bit cells under read stress, and $V_{DD} - V_{WL} < 0$ for bit cells under write operation.[13]	■ Settling time ■ Up to two extra power supplies ■ Extent of assist limited by hold margin
	Float V_{DD} and charge share with grounded dummy line to reduce cell supply voltage for write.[14]	■ Assist setting must be determined at design time ■ Extent of assist limited by hold margin
Dynamic modulation of horizontally routed V_{DD}	Fully collapse V_{DD} during write for only the adjacent bits in a word.[6]	■ Bit cell area impact ■ Incompatible with bit interleaving
Dynamic modulation of V_{WL}	During read, cut off wordline (WL) pulse before unstable bit flip completes.[15]	■ WL timing control ■ Incompatible with bit interleaving
	Swap V_{WL} and V_{DD} between two supply levels: $V_{WL} < V_{DD}$ during read, $V_{WL} > V_{DD}$ during write to maximize the extent of assist.[16]	■ Requires extra supply and multiplexing in both row and column periphery ■ Incompatible with bit interleaving
Dynamic modulation of horizontally routed V_{SS}	Raise footer on 2T read stack for unselected rows to eliminate bitline (BL) off current.[17]	■ Reduced performance at nominal voltage ■ Peripheral area overhead ■ Bit cell area impact
Dynamic modulation of globally routed V_{SS}	Elevate V_{VSS} during idle time or write, but discharge to 0 V during a read cycle.[18]	■ Reduced performance at nominal voltage ■ Separate supply level must be generated
Dynamic modulation of vertically routed V_{SS}	Generate negative V_{VSS} from peripheral charge pump to improve read current and read stability.[19]	Increased leakage and disturbance to neighboring cells in the same column
Dynamic modulation of BL	Generate negative BL voltage from peripheral charge pump to improve write ability.[19]	Extent of assist must be determined at design time and is limited by disturbance to neighboring cells in the same column
	During read, restore every BL to logic level for writeback of disturbed cells.[20]	Increased active power and sense-amplifier area
	Lower BL precharge with supply[21] or current pulse[15] for read stability.	Uncertainty associated with optimal magnitude of assist
Static V_{WL} setting	Reduce V_{WL} with loading NMOS transistors to track global conditions.[14]	■ Reduced performance and limited adaptability to global process corners ■ Small area overhead
	Adaptively choose V_{WL} underdrive by monitoring global conditions with replica memory cells.[22]	■ Added test flow complexity ■ Small power and area overhead
Static dual supply rails	Employ a separate bit cell supply V_{CS} and periphery supply V_{DD} to improve stability and performance.[23]	■ System complexity ■ Test-flow overhead
Static NMOS and PMOS body bias	Recenter the balance between read and write stability over variable global process corners.[24]	■ Detecting global conditions ■ Limited influence of body bias

Alternative SRAM bit cells

Figures 2b through 2e show alternative bit cells that bypass the stability constraints of the 6T cell. Adding one transistor (along with an associated WL control signal) in series with one of the pull-down NMOS devices greatly reduces read and write instability by breaking the feedback in the cross-coupled inverters.[4] The resulting 8-Kbyte 7T design functions

Table 2. Comparison of bit cell characteristics.

Property	6T cell	7T cell	8T cell	Single-ended 10T cell	10T cell differential
Area overhead compared to 6T	0%	13%	30%, 50% (with write and read assists)	70% (with write assist)	110% (with write assist)
Stability limitation	Read and write	Write	Write or hold (with assist)	Hold (with assist)	Hold (with assist)
Sense scheme	Differential or single-ended	Single-ended	Single-ended	Single-ended	Differential
Column interleaving	Yes	No	No	No	Yes

at a ratio V_{DD}/V_{TH} of around 1.4. By adding two transistors, the read can be entirely decoupled from the write operation in an 8T cell by sensing the data through a separate read stack controlled by a separate read wordline (RWL).[5] The remaining 6T portion of the cell is optimized for write, resulting in an overall lower V_{min}.

Further leakage and energy savings are enabled by operating an SRAM in or near the subthreshold region. Thus, an 8T SRAM design by Verma and Chandrakasan contains a write assist in which a horizontally routed V_{DD} line is collapsed during write.[17] As a result, the bit cell array V_{min} is limited by the hold margin. For memory applications unconstrained by performance that must always retain state (e.g., wireless sensor nodes for environmental monitoring), leakage is the primary concern. Therefore, operating at the minimum possible supply, which is determined by hold stability, will require a solution like this 8T SRAM design.

Furthermore, the V_{SS} line for the 2T read stack is routed horizontally and driven high for unselected rows, thereby eliminating the off-state leakage on the BL. One of the drawbacks is the area overhead related to routing four signals—V_{DD}, V_{SS}, V_{WL}, and V_{RWL}—horizontally on the memory cell's pitch. The 10T cell described by Calhoun and Chandrakasan does not require a read assist, as in the 8T subthreshold design, but it must contain two extra transistors to mitigate BL leakage from unselected cells.[6]

Because of the required WL control signals or horizontally routed power supply assists, none of the aforementioned cells (7T, 8T, 10T) can interleave bit cell columns. In response, some memory designs propose the read-modify-write scheme for these alternative bit cell topologies.[16] Alternatively, the 10T cell described by Chang et al. permits bit interleaving and exhibits superior sense margin with a differential read path based on a DCVSL (differential cascade voltage-switch-logic level) structure at the column periphery.[7] This cell enables column interleaving because of a NAND-type structure in the pass-gate that conducts only if it receives both a row-select and column-select signal. There is a performance degradation from stacked transistors that requires boosted WL voltages, but BL leakage is reduced at the same time.

Finally, most of the alternatives to the 6T cell require single-ended sensing instead of differential sensing. This property imposes an additional constraint that the weakest memory cell in the chip must overpower the strongest *off* BL elsewhere in the chip. Furthermore, a global mechanism to define a midpoint between the two data states must be established—whether it be strobe timing, a voltage input to a pseudodifferential amplifier, or an implicit conversion of a dynamic BL to static voltage levels. The challenge of separating *on* and *off* BLs in a single-ended sensing scheme could ultimately limit V_{min} instead of bit cell stability. Table 2 summarizes the characteristics of the various bit cell options we've discussed.

Sensing innovations

For low-voltage operation, bit cell assists are not sufficient. Hence, new sensing approaches, beyond those available for operation at higher voltages, have emerged. Several solutions have addressed the difficulties of single-ended sensing.

The 8T SRAM described by Verma and Chandrakasan employs sense-amplifier redundancy by selecting a backup sense amplifier in case the original one does not work (see Figure 6a).[17] Sinangil, Verma, and Chandrakasan have also employed reconfigurability by choosing one of two gate-input differential

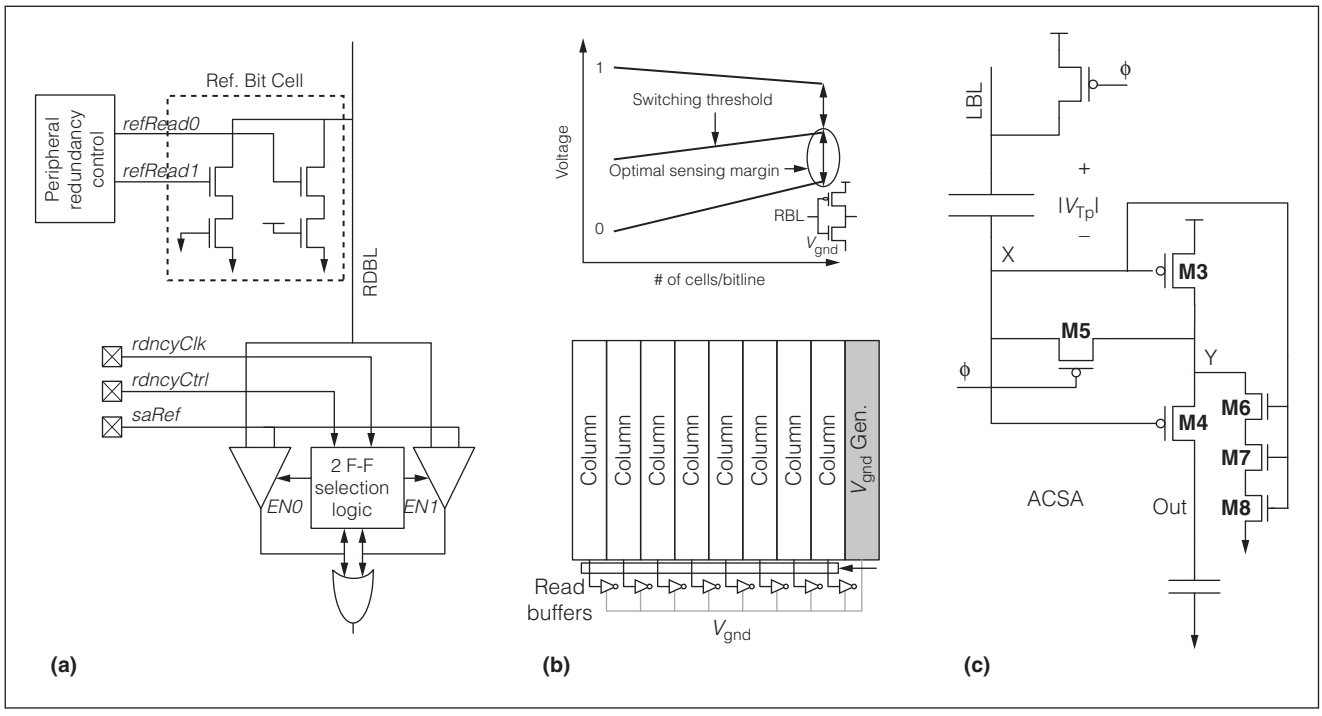

Figure 6. Low-voltage sensing techniques for SRAM: reconfigurability (Source: Verma and Chandrakasan.[17]) (a), replica biasing (Source: Kim et al.[26]) (b), and offset compensation (Source: Qazi et al.[27]) (c). (ACSA: AC-coupled sense amplifier.)

sense amplifiers.[2] For high-speed operation, the BL common-mode voltage is closer to V_{DD} and is better sensed through an amplifier with NMOS inputs. For low-speed, low-voltage operation, the significant amount of BL droop produces a signal window closer to V_{SS} that is better served by a sense amplifier with PMOS inputs. Alternatively, Cosemans, Dehaene, and Catthoor employ a single sense amplifier with redundant voltage references to tune each sense amplifier.[25] Thus, the effective offset is reduced at the cost of testing complexity.

Another type of sensing strategy involves using replica circuits to determine optimum bias conditions. The subthreshold SRAM design based on a 10T cell described by Kim et al. modifies read stack devices M7 through M10 so that unselected cells in a common BL have the same type of parasitic leakage current pulling up on the BL regardless of the data state.[26] As a result, it's possible to observe the voltage generated on the BL when a selected cell pulls the BL low through a replica column, as illustrated in Figure 6b. This voltage provides the virtual ground voltage to sensing inverters for functional columns. As a result, the trip point of the sensing inverter automatically adjusts to the midpoint between the BL's logic-high and logic-low voltages.

A third class of sensing innovation relates to offset compensation (see Figure 6c). Variation-tolerant sensing networks are critical to work within the diminishing separation between an *on* and *off* BL at low voltages. For the single-ended case (in Figure 7a), this diminishing separation is illustrated by the light gray delay histograms in Figures 7b and 7c, which correspond to sensing a long BL of 256 cells at 1 V and 0.55 V, respectively, with a dynamic PMOS inverter. In the latter case, the distribution of the false 1 overlaps with the distribution of the true 1, making it impossible to capture the data of all bits correctly.

The dark gray histograms in Figures 7b and 7c show the result of sensing the same long BL with the AC-coupled sense amplifier (ACSA) of Figure 6c, which is described in detail elsewhere.[27] Not only does the worst-case delay of the true 1 decrease, but also the separation between true 1 and false 1 widens and preserves a sampling window at 0.55 V for 90% yield of a 64-Kbyte array. The ACSA works by storing the variable threshold of the amplifying PMOS M3 in series with the BL signal while also suppressing the variation of the output PMOS M4 by

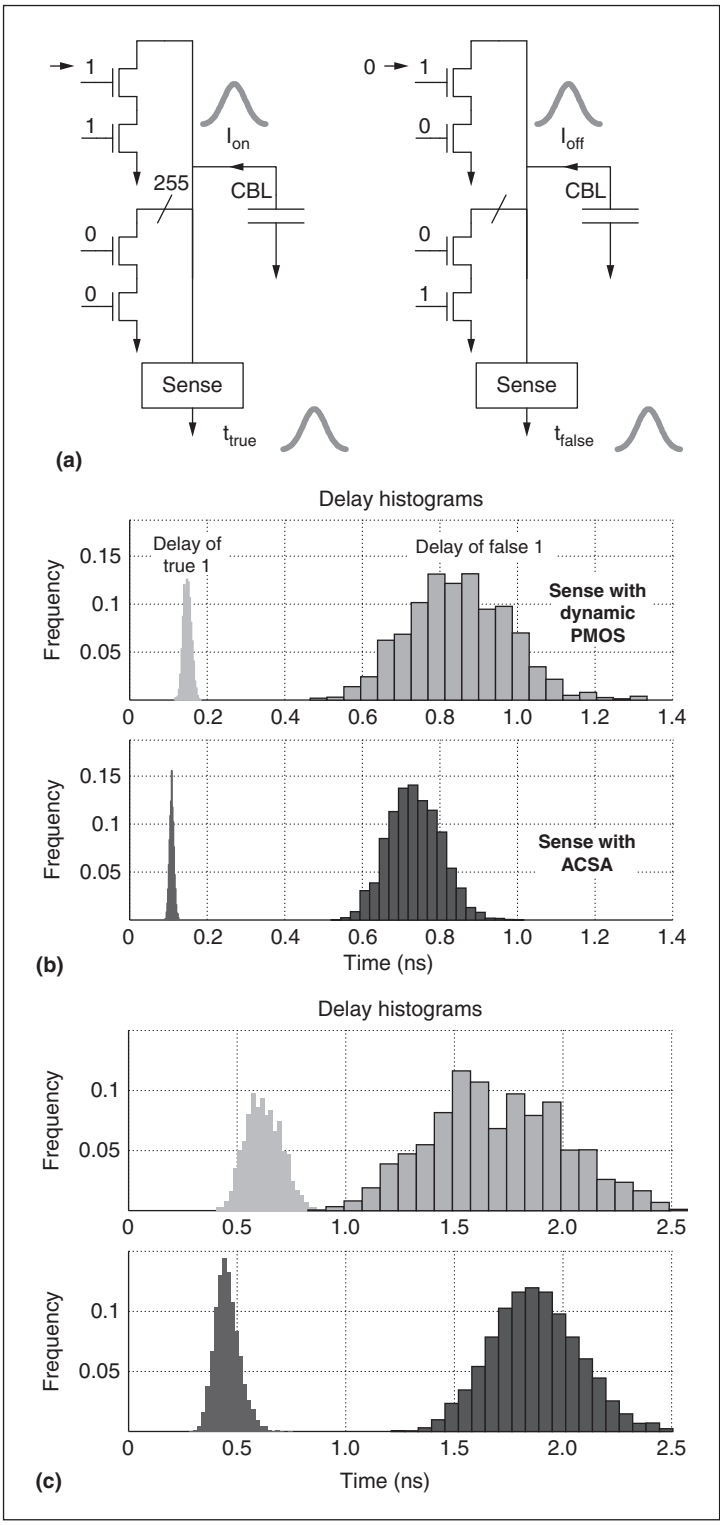

Figure 7. The statistical separation of bitline on and off states (a) corresponds to the dynamic sampling window between the slowest true 1 and the fastest false 1 at 1 V (b) and at 0.55 V (c) under global conditions (corner, temperature) of fast NMOS, fast PMOS; 85°C.

driving its gate-to-source voltage at a rate equal to $(1 + g_m r_o)$ times the BL signal development, where g_m is the transconductance of M3, and r_o is the output resistance of M3. Our measurements show that when employing offset compensation and avoiding variable timing-control signals to activate the sense network, the memory operates down to 0.57 V. As a limitation of single-ended sensing, this technique requires a midpoint reference in the form of a timing signal that falls within the sampling window.

Data retention voltage for standby power

For idle memory banks, lowering the supply voltage to the limit of hold stability enables a dramatic reduction of standby power as both gate leakage and subthreshold leakage (through the phenomenon of drain-induced barrier lowering) scale down. This limit, known as the data-retention voltage (DRV), results from the effects of the local variation depicted in Figures 4c and 4f. Moreover, this relation changes with process corner, temperature, and transistor end-of-life degradation.[28]

Because of the uncertainty related to the DRV, current approaches conservatively place PMOS or NMOS diodes in series with the SRAM power supply to reduce the SRAM array leakage. Recent memory designs have emphasized both accurately setting the DRV with active regulation[15] and applying retention bias with fine granularity to maximize the number of memory cells held in retention mode. For example, the SRAM design described by Pilo et al. dynamically biases individual subarrays out of retention on a cycle-by-cycle basis.[29]

Researchers have proposed techniques to predict the DRV in order to aggressively reduce the standby supply voltage. Takeyama et al., for example, bias the array to twice the threshold voltage of bit cell devices[30] (a sufficient condition for retention) by observing the thresholds through replica memory cell devices. Qazi et al. determine the DRV from a smaller sample of 256 sensor cells by accelerating the failure rate with skewed supplies and analytically estimating the true failure rate with statistical techniques.[27]

SRAM design methodologies for yield

SRAM specifications such as performance and V_{min} are subject to the constraint of functionality, which is difficult to guarantee at process-variation extremes. This type of problem can be treated using

the Monte Carlo method, which samples memory cell parameters over multiple trials and interprets the frequency of failing trials as the failure probability. However, the required number of simulation runs for a high-confidence estimate of a failure probability p is given by $N_{MC} = 100/p$. For a multi-megabit memory, this can require hundreds of millions of simulation runs.

Therefore, accelerated simulation techniques have been developed to more efficiently use computational resources. For instance, Singhee and Rutenbar have evaluated the characteristics of a 6T SRAM bit cell.[31] Their approach runs fewer simulations than in the standard Monte Carlo approach by blocking out the realization of variation parameters close to the nominal case. The resulting simulations produce more data in the tails of probability distributions, which are then analytically modeled. Another set of approaches have focused on the application of *importance-sampling simulation* to IC design. Rather than counting the number of failing realizations relative to passing realizations, failure probability is observed by coarsely determining how much skew the nominal realization needs to become a failing realization. Then, the results of a short, skewed Monte Carlo simulation trial are analytically unbiased to estimate the true failure rate.

For example, Qazi et al. have evaluated the read access yield of the SRAM critical path.[32] The circuit's 12D parameter space of local threshold voltage fluctuation is explored through a two-stage process of statistical sampling. First, the general direction of skew is identified through a modest number of simulation trials on the surface of a generalized sphere in the parameter space. Next, the estimated skew is refined through targeted local sampling, which gravitates toward failure mechanisms of increasing likelihood, as indicated by the joint probability density function. Finally, an importance-sampling simulation is run with distributions, whose means are shifted according to the skew until the estimator settles to a tolerable level of relative error (sample variance). (The estimator is obtained from a well-known formula on the basis of the specific parameter realizations in each trial, along with the skew.) This technique matches the results of the nominal Monte Carlo method, with 650× fewer total Spice simulations at a failure of 10^{-4}, and it extends to far lower failure probabilities with increasing speedups.

The basic intuition behind importance sampling comes from the analytical formulation of the failure as the integral of the probability density of variation parameters over the failure region. This integral is dominated by a small subregion of interest, so a Monte Carlo simulation approach can converge more quickly if most random samples are drawn from this region. In the example of the SRAM read path, the circuit designer can quickly reconcile the interaction of the variable BL signal and sense-amplifier offset to predict overall chip yield. More importantly, as the designer explores solutions to enable lower voltage operation, the iteration time between circuit modification and yield determination reduces exponentially.

Moreover, postfabrication techniques to recover yield in SRAM have long existed. Some techniques are based on bypassing faulty memory cells, determined during initial product test, with redundant rows and columns.[33] Others use ECC, primarily to fix transient faults, but also hard defects.[11] Such techniques will increase in relevance to the stressful operating conditions of low-voltage SRAM.

As semiconductor process technology continues scaling, transistor mismatches and process fluctuations will worsen. Circuit designers are challenged to address the need for low-voltage SRAM design through novel circuit techniques in the periphery and in the bit cell while minimizing area overhead or performance penalty. Although different cell topologies providing many advantages over conventional 6T bit cells have been proposed, the 8T bit cell has attracted considerable attention from academia and industry because of its compact layout implementation.

As bit cell design, circuit assists, and sensing techniques continue to develop, the circuit designer will employ improved statistical CAD methodologies. The capability of the current framework of circuit design tools cannot cope with the statistical simulation of circuits containing hundreds of randomly fluctuating devices, both within a die and from die to die. In fact, satisfactory methods for selecting global process conditions for analysis are essential. Current approaches typically cycle through all permutations of global conditions by brute force and must continue to do so for lack of more-efficient methodologies, yet die-to-die variation has just as important an impact on SRAM yield.[22]

Finally, a low-power SRAM does not minimize its power as an isolated unit but rather enables a low-power system. Various circuit components—I/O, digital core, analog interfaces, and memory—will not have the same optimum supply voltage. For example, logic and memory will exhibit different minimum energy points because of the different ratios of switched capacitance to idle device width. Therefore, the minimization of system power will require a balance among energy-efficient DC-DC converters, level-shifting circuitry, power grid routing, noise immunity, and system complexity. Whether it dominates the die area, requires the highest supply voltage, or must remain always on for retention, SRAM will continue to play a critical role and must be mindfully integrated into this balance. ■

Acknowledgments

We acknowledge the funding support of both DARPA and the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity. We also thank Texas Instruments and IBM for chip fabrication.

■ References

1. N.A. Kurd et al., "Westmere: A Family of 32nm IA Processors," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 10), IEEE Press, 2010, pp. 96-97.
2. M.E. Sinangil, N. Verma, and A.P. Chandrakasan, "A Reconfigurable 8T Ultra-Dynamic Voltage Scalable (U-DVS) SRAM in 65 nm CMOS," *IEEE J. Solid-State Circuits,* vol. 44, no. 11, 2009, pp. 3163-3173.
3. J. Kwong et al., "A 65 nm Sub-V_t Microcontroller with Integrated SRAM and Switched Capacitor DC-DC Converter," *IEEE J. Solid-State Circuits,* vol. 44, no. 1, 2009, pp. 115-126.
4. K. Takeda et al., "A Read-Static-Noise-Margin-Free SRAM Cell for Low-VDD and High-Speed Applications," *IEEE J. Solid-State Circuits,* vol. 41, no. 1, 2006, pp. 113-121.
5. L. Chang et al., "Stable SRAM Cell Design for the 32 nm Node and Beyond," *Proc. Symp. VLSI Tech.,* IEEE Press, 2005, pp. 128-129.
6. B.H. Calhoun and A.P. Chandrakasan, "A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation," *IEEE J. Solid-State Circuits,* vol. 42, no. 3, 2007, pp. 680-688.
7. I.J. Chang et al., "A 32 kb 10T Sub-threshold SRAM Array with Bit-Interleaving and Differential Read Scheme in 90 nm CMOS," *IEEE J. Solid-State Circuits,* vol. 44, no. 2, 2009, pp. 650-658.
8. W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for Sub-45 nm Design Exploration," *IEEE Trans. Electron Devices,* vol. 53, no. 11, 2006, pp. 2816-2823.
9. K.J. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," *Proc. IEEE Int'l Electron Devices Meeting* (IEDM 07), IEEE Press, 2007, pp. 471-474.
10. T. Karnik et al., "Scaling Trends of Cosmic Ray Induced Soft Errors in Static Latches beyond 0.18 μ," *Proc. Symp. VLSI Circuits,* 2001, pp. 61-62.
11. T. Suzuki et al., "A Sub-0.5-V Operating Embedded SRAM Featuring a Multi-bit-Error-Immune Hidden-ECC Scheme," *IEEE J. Solid-State Circuits,* vol. 41, no. 1, 2006, pp. 152-160.
12. K. Osada et al., "Universal-V_{dd} 0.65-2.0-V 32-kB Cache Using a Voltage-Adapted Timing-Generation Scheme and a Lithographically Symmetrical Cell," *IEEE J. Solid-State Circuits,* vol. 36, no. 11, 2001, pp. 1738-1744.
13. K. Zhang et al., "A 3-GHz 70-mb SRAM in 65-nm CMOS Technology with Integrated Column-Based Dynamic Power Supply," *IEEE J. Solid-State Circuits,* vol. 41, no. 1, 2006, pp. 146-151.
14. S. Ohbayashi et al., "A 65-nm SoC Embedded 6T-SRAM Designed for Manufacturability with Read and Write Operation Stabilizing Circuits," *IEEE J. Solid-State Circuits,* vol. 42, no. 4, 2007, pp. 820-829.
15. M. Khellah et al., "A 4.2GHz 0.3mm2 256kb Dual-Vcc SRAM Building Block in 65nm CMOS," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 06), IEEE Press, 2006, pp. 2572-2581.
16. Y. Morita et al., "An Area-Conscious Low-Voltage-Oriented 8T-SRAM Design under DVS Environment," *Proc. IEEE Symp. VLSI Circuits,* IEEE Press, 2007, pp. 256-257.
17. N. Verma and A.P. Chandrakasan, "A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy," *IEEE J. Solid-State Circuits,* vol. 43, no. 1, 2008, pp. 141-149.
18. M. Yamaoka et al., "A 300 MHz 25μA/Mb Leakage On-Chip SRAM Module Featuring Process-Variation Immunity and Low-Leakage-Active Mode for Mobile-Phone Application Processor," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 04), vol. 1, IEEE Press, 2004, pp. 494-495, 542.

19. M. Yabuuchi et al., "A 45nm 0.6V Cross-Point 8T SRAM with Negative Biased Read/Write Assist," *Proc. Symp. VLSI Circuits,* IEEE Press, 2009, pp. 158-159.
20. H. Pilo et al., "An SRAM Design in 65-nm Technology Node Featuring Read and Write-Assist Circuits to Expand Operating Voltage," *IEEE J. Solid-State Circuits,* vol. 42, no. 4, 2007, pp. 813-819.
21. A. Bhavnagarwala et al., "Fluctuation Limits & Scaling Opportunities for CMOS SRAM Cells," *Proc. IEEE Int'l Electron Devices Meeting* (IEDM 05), IEEE Press, 2005, pp. 659-662.
22. H. Nho et al., "A 32nm High-k Metal Gate SRAM with Adaptive Dynamic Stability Enhancement for Low-Voltage Operation," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 10), IEEE Press, 2010, pp. 346-347.
23. J. Pille et al., "Implementation of the Cell Broadband Engine in 65 nm SOI Technology Featuring Dual Power Supply SRAM Arrays Supporting 6 GHz at 1.3 V," *IEEE J. Solid-State Circuits,* vol. 43, no. 1, 2008, pp. 163-171.
24. M. Yamaoka et al., "65nm Low-Power High-Density SRAM Operable at 1.0V under 3σ Systematic Variation Using Separate Vth Monitoring and Body Bias for NMOS and PMOS," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 08), IEEE Press, 2008, pp. 384-385, 622.
25. S. Cosemans, W. Dehaene, and F. Catthoor, "A 3.6 pJ/Access 480 MHz, 128 kb On-Chip SRAM with 850 MHz Boost Mode in 90 nm CMOS with Tunable Sense Amplifiers," *IEEE J. Solid-State Circuits,* vol. 44, no. 7, 2009, pp. 2065-2077.
26. T.-H. Kim et al., "A 0.2 V, 480 kb Subthreshold SRAM with 1 k Cells per Bitline for Ultra-Low-Voltage Computing," *IEEE J. Solid-State Circuits,* vol. 43, no. 2, 2008, pp. 518-529.
27. M. Qazi et al., "A 512kb 8T SRAM Macro Operating down to 0.57V with an AC-Coupled Sense Amplifier and Embedded Data-Retention-Voltage Sensor in 45nm SOI CMOS," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 10), 2010, pp. 350-351.
28. H. Qin et al., "SRAM Leakage Suppression by Minimizing Standby Supply Voltage," *Proc. 5th Int'l Symp. Quality Electronic Design* (ISQED 04), IEEE Press, 2004, pp. 55-60.
29. H. Pilo et al., "A 450ps Access-Time SRAM Macro in 45nm SOI Featuring a Two-Stage Sensing-Scheme and Dynamic Power Management," *Proc. IEEE Int'l Solid-State Circuits Conf.* (ISSCC 08), 2008, pp. 378-379, 621.
30. Y. Takeyama et al., "A Low Leakage SRAM Macro with Replica Cell Biasing Scheme," *Proc. IEEE Symp. VLSI Circuits,* IEEE Press, 2005, pp. 166-167.
31. A. Singhee and R.A. Rutenbar, "Statistical Blockade: A Novel Method for Very Fast Monte Carlo Simulation of Rare Circuit Events, and Its Application," *Proc. Design, Automation and Test in Europe Conf.* (DATE 07), IEEE CS Press, 2007.
32. M. Qazi et al., "Loop Flattening & Spherical Sampling: Highly Efficient Model Reduction Techniques for SRAM Yield Analysis," *Proc. Design, Automation and Test in Europe Conf.* (DATE 10), IEEE CS Press, pp. 801-806.
33. H. McIntyre et al., "A 4-MB On-Chip L2 Cache for a 90-nm 1.6-GHz 64-Bit Microprocessor," *IEEE J. Solid-State Circuits,* vol. 40, no. 1, 2005. pp. 52-59.

Masood Qazi is pursuing a PhD in electrical engineering at the Massachusetts Institute of Technology (MIT). His research interests include IC design for semiconductor memories. He has an MEng in electrical engineering and computer science from MIT. He is a member of IEEE.

Mahmut E. Sinangil is pursuing a PhD in electrical engineering at MIT. His research interests include low-power digital-circuit design in embedded memories and video coding. He has an SM in electrical engineering from MIT. He is a member of IEEE.

Anantha P. Chandrakasan is the Joseph F. and Nancy P. Keithley Professor of Electrical Engineering and the director of the Microsystems Technology Laboratories at MIT. His research interests include low-power digital IC design, wireless microsensors, ultrawideband radios, and emerging technologies. He has a PhD in electrical engineering and computer sciences from the University of California, Berkeley. He is a Fellow of IEEE.

■ Direct comments and questions about this article to Masood Qazi, 50 Vassar St., Room 38-107, Cambridge, MA 02139; mqazi@mit.edu.