

A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation

Benton Highsmith Calhoun, *Member, IEEE*, and Anantha P. Chandrakasan, *Fellow, IEEE*

Abstract—Low-voltage operation for memories is attractive because of lower leakage power and active energy, but the challenges of SRAM design tend to increase at lower voltage. This paper explores the limits of low-voltage operation for traditional six-transistor (6 T) SRAM and proposes an alternative bitcell that functions to much lower voltages. Measurements confirm that a 256-kb 65-nm SRAM test chip using the proposed bitcell operates into sub-threshold to below 400 mV. At this low voltage, the memory offers substantial power and energy savings at the cost of speed, making it well-suited to energy-constrained applications. The paper provides measured data and analysis on the limiting effects for voltage scaling for the test chip.

Index Terms—Low-voltage memory, sub-threshold SRAM, voltage scaling.

I. INTRODUCTION

SUBTHRESHOLD digital circuit design has emerged as a low-energy solution for applications with strict energy constraints. Analysis of sub-threshold designs has focused on logic circuits (e.g., [1]). SRAMs comprise a significant percentage of the total area and total power for many digital chips [2]. SRAM leakage can dominate total chip leakage, and switching highly capacitive bitlines and wordlines is costly in terms of energy. Lowering V_{DD} for SRAM saves leakage power and access energy. Also, for system integration, SRAM must become capable of operating at sub-threshold voltages that are compatible with sub-threshold combinational logic. Overcoming the difficulties of operating an SRAM in sub-threshold requires both circuit and architectural innovations. The benefits are significant, however, since low-energy SRAM is essential for enabling ultra-low-energy systems. This paper describes an SRAM capable of operating in the sub-threshold region.

Previous low-power memories show a trend of lower voltage operation. Exploiting dynamic voltage scaling (DVS) for SRAM is one motivation for designing a voltage-scalable memory. A 0.18- μm 32-kB four-way associative cache offers DVS compatibility from 120 MHz, 1.7 mW at 0.65 V to 1.04 GHz, 530 mW at 2 V [3]. Although DVS can provide power reduction for active memories, most previous approaches apply voltage scaling primarily to idle blocks by lowering V_{DD} (e.g., [2], [4]–[6]), raising ground (e.g., [7]–[10]), or both (e.g.,

Manuscript received May 23, 2006; revised August 1, 2006. This work was supported by DARPA and by Texas Instruments Incorporated. Chip fabrication was provided by Texas Instruments.

B. H. Calhoun is with the Electrical and Computer Engineering Department, University of Virginia, Charlottesville, VA 22904-4743 USA (e-mail: bcalhoun@virginia.edu).

A. P. Chandrakasan is with the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Digital Object Identifier 10.1109/JSSC.2006.891726

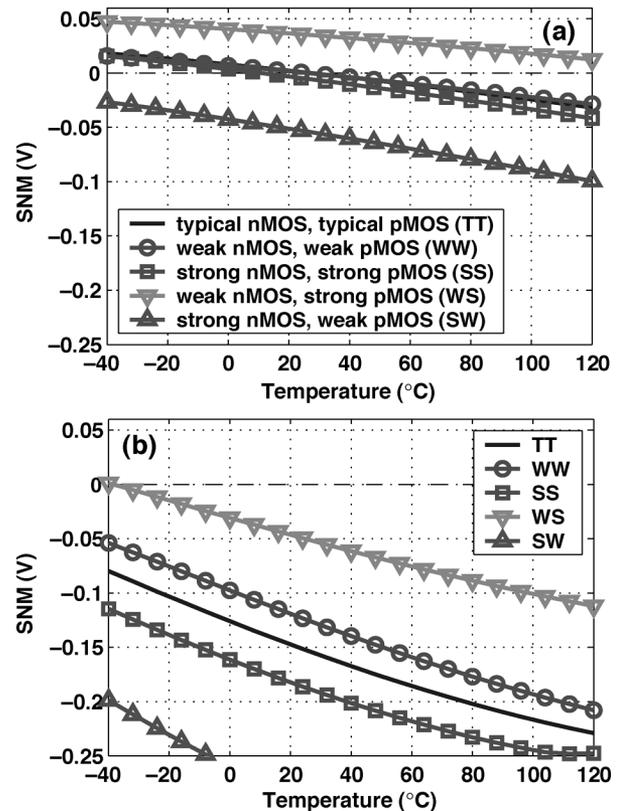


Fig. 1. SNM for write access versus temperature and process corner (TT, WW, SS, WS, and SW) at $V_{DD} = 0.3$ V (a) and $V_{DD} = 0.6$ V (b). Negative SNM indicates successful write.

[11]). Implementations of SRAM using lower V_{DD} in standby are available [5] along with software policies to determine when to enter the lower leakage mode [2]. Voltage scaling for SRAM promises to continue, leading to sub-threshold storage modes and even sub-threshold operation for SRAMs operating in tandem with sub-threshold logic.

One issue for deeply voltage scaled SRAM is soft error rate (SER). Soft errors occur when an alpha particle or cosmic ray strikes a memory node and causes data loss. Since bitcell storage capacitance decreases with scaling and voltage scaling further reduces the stored charge, SER is a concern for sub-threshold memory. Fortunately, there are methods for handling soft errors. Studies of soft errors have shown that multi-cell errors from a single strike only occur in two to three adjacent cells along a wordline [12]. Thus, physically interspersing bits from different words can prevent multi-errors from occurring in a single word [12]. Coupling this with error correcting codes can dramatically reduce SER [8].

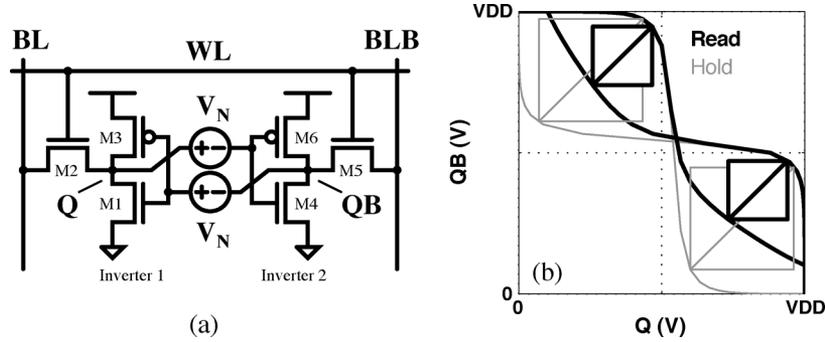


Fig. 2. Schematic for 6 T bitcell showing voltage noise sources for finding SNM [15] (a), and example Hold and Read SNM butterfly plots (b).

Clearly, previous efforts have explored many options for voltage scaling. However, none have yet pushed voltage scaling into the sub-threshold region during active operation.

II. SIX-TRANSISTOR SRAM BITCELL IN SUB-THRESHOLD

Predictions in [13] suggest that process variations will limit standard 90-nm SRAMs to around 0.7 V operation due to degraded Read Static Noise Margin (SNM) and reduced write margin. Small transistors combine with random and systematic process variations to cause a large spread in Read SNM that leads to destructive read errors for bits at the tail of the distribution. Standard write operation depends on a ratio of currents, and process variations make this ratio difficult to maintain as V_{DD} decreases, leading to write errors. These practical problems limit traditional six-transistor (6 T) bitcells and architectures to higher V_{DD} , above-threshold operation. Reports in the literature of 65-nm SRAMs confirm this voltage barrier. A 65-nm SRAM built in a dynamic-double-gate SOI (D2G-SOI) process functions to 0.7 V and is predicted to fail below 1.0 V for bulk CMOS [14]. A bulk CMOS 65-nm SRAM reports a minimum operating voltage of 0.7 V [9]. Our results confirm that SNM degradation and inability to write are the two primary obstacles to sub-threshold SRAM functionality, where they are exacerbated by the exponential impact of V_T variations.

A. Write Operation

Proper write operation depends on sizing the access nMOS to win the ratioed fight with the pMOS inside the bitcell to write a “0”. For a successful write, the bitcell becomes monostable, forcing the internal voltages to the correct values. If the cell retains bistability then the write does not occur, and the SNM is positive on the cell’s butterfly plot. Thus, a negative SNM indicates a successful write (monostability in the cell). For above- V_T operation, stronger nMOS devices (due to mobility) and relatively low dependence of current on V_T make device sizing successful at maintaining the proper ratio of currents for writing the cell. For sub-threshold, the ratio of currents in p/nMOS depends exponentially on V_T . Since process designers generally focus on strong-inversion operation, the sub-threshold pMOS and nMOS current can be imbalanced for typical transistors. Even if the pMOS and nMOS currents are well-balanced at the typical nMOS, typical pMOS (TT) corner, process variation can still create a relative difference in p/nMOS current of an

order of magnitude or more. Furthermore, local variations in V_T from cell to cell can aggravate this problem. For sub- V_T , sizing alone is not a strong knob for fixing this problem because only unreasonable sizing ratios could account for the wide ranges of possible current that arise due to V_T mismatch.

In the 65-nm process for which we are designing, iso-size pMOS devices are stronger in sub- V_T than nMOS by roughly an order of magnitude, which makes write functionality more challenging. Fig. 1 shows the write margin (neg. SNM means successful write) of a 6 T bitcell versus temperature and process corner. At $V_{DD} = 300$ mV in Fig. 1(a), the writing fails for large regions of process corner and temperature. The general trend showing an improvement of write operation (i.e., more negative margin) at higher temperature occurs because the pMOS transistors weaken relative to nMOS as temperature rises. As V_{DD} increases, the write margin improves. Fig. 1(b) shows the write margin at 0.6 V. This voltage is above V_T , so the pMOS has weakened relative to the nMOS because the mobility dominates the differences in V_T . Even at 0.6 V, the write margin is barely negative for the worst-case corner, and this plot does not account for local V_T variation. For these reasons, $V_{DD} = 0.6$ V is the best case voltage for which we can expect traditional write operations to work for a sub-threshold memory in this 65-nm process.

B. Read Operation: Static Noise Margin

Fig. 2 shows a conceptual setup for modeling SNM [15]. Noise sources having value V_N are introduced at each of the internal nodes in the bitcell. As V_N increases, the stability of the cell reduces. Once V_N exceeds the SNM, then the cell loses its bistability and its data. Cell stability during active operation represents a more significant limitation to SRAM operation than during hold. At the onset of a read access, the wordline is “1” and the bitlines are precharged to “1”. The internal node of the bitcell that represents a zero gets pulled upward through the access transistor due to the voltage dividing effect across the access transistor (M_2, M_5) and drive transistor (M_1, M_4), which degrades the Read SNM. Fig. 2 shows example butterfly curves during hold and read that illustrate the degradation in SNM during read.

Process variation makes matters worse by shifting the voltage transfer characteristics (VTCs) of the cell inverters and creating a distribution of SNM for both hold and read. A study of the impact of variations on SNM in sub- V_T appears in [16]. Fig. 3

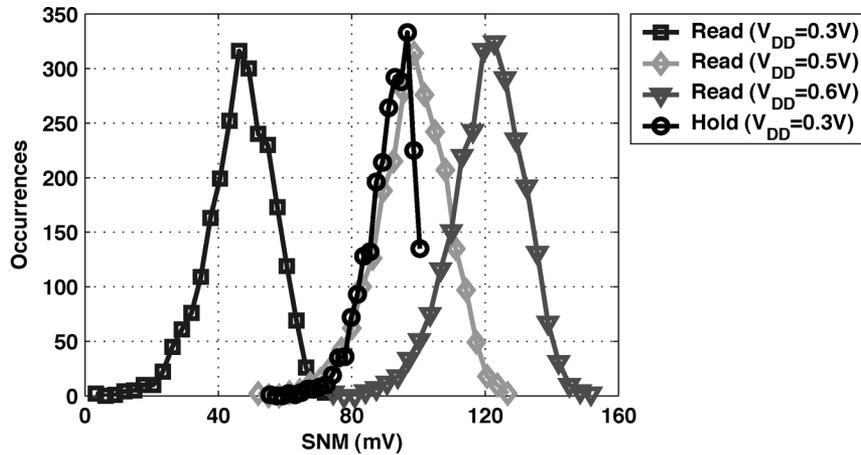


Fig. 3. Distribution of Hold SNM at 300 mV compared with Read SNM distributions at different voltages. Read SNM at 500 mV has the same mean, but it has a larger standard deviation.

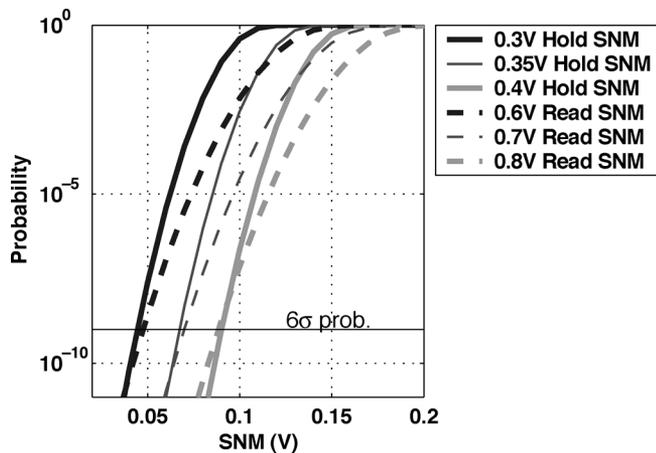


Fig. 4. CDFs of SNM distributions showing that avoiding the Read SNM allows a reduction in V_{DD} by ~ 0.5 for the same 6σ stability.

shows the distribution of the Read and Hold SNMs for a 6 T bitcell at a 300-mV supply voltage. The mean Read SNM is only slightly above half of the mean Hold SNM, and the deviation of the Read SNM is larger than for the Hold SNM. For a multiple megabit memory, numerous cells will have Read SNM less than zero based on this statistical analysis. From this figure, the mean of the Read SNM at 500 mV roughly equals the mean of the Hold SNM at 300 mV. However, it is unclear from this plot how the Hold SNM and Read SNM compare at the worst-case tails. Fig. 4 shows the cumulative distribution function (CDFs) derived from the distributions. For 6σ probability, the Hold SNM for a given V_{DD} roughly equals the Read SNM for twice that V_{DD} in the range of interest. This means that a memory that avoids the Read SNM problem can operate at roughly half of the V_{DD} of a 6 T memory with the same 6σ bitcell stability.

III. A SUB-THRESHOLD BITCELL DESIGN

Previously published works have scaled SRAM V_{DD} into the sub-threshold region during idle, but no SRAM actually operates in this region. The $0.18\text{-}\mu\text{m}$ memory in [1] provides one exception, operating into deep sub- V_T at 180 mV. However, the memory resembles a register file (latch with tristate driver for

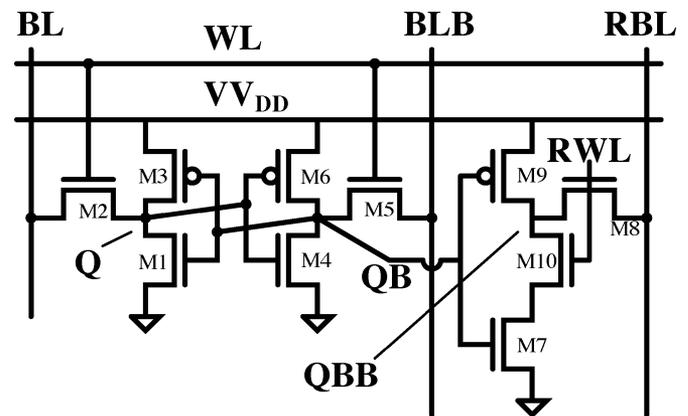


Fig. 5. Schematic of the 10 T sub-threshold bitcell.

writing and muxed outputs) and has an equivalent bitcell size of 18 T. We can use this previous implementation [1] as an end-point in the range of bitcell options that spans from the 6 T bitcell (inoperable below 600–700 mV in 65 nm) to the 18 T bitcell, which will function robustly in sub-threshold since it looks and functions more like combinational logic. In between these two options are many possible bitcell designs that address the obstacles to sub-threshold operation by increasing the number of transistors relative to the 6 T cell. The bitcell that this section describes [17] was selected from among many others because it represents the best trade-off of functionality and area; it is the smallest bitcell from those examined that provides robust sub-threshold functionality.

Fig. 5 shows the schematic of the 10 T sub-threshold bitcell. Transistors M_1 through M_6 are identical to a 6 T bitcell except that the source of M_3 and M_6 tie to a virtual supply voltage rail, VV_{DD} . Write access to the bitcell occurs through the write access transistors, M_2 and M_5 , from the write bitlines, BL and BLB. Transistors M_7 through M_{10} implement a buffer used for reading. Read access is single-ended and occurs on a separate bitline, RBL, which is precharged to V_{DD} prior to read access. The wordline for read also is distinct from the write wordline. One key advantage to separating the read and write wordlines and bitlines is that a memory using this bitcell can have distinct

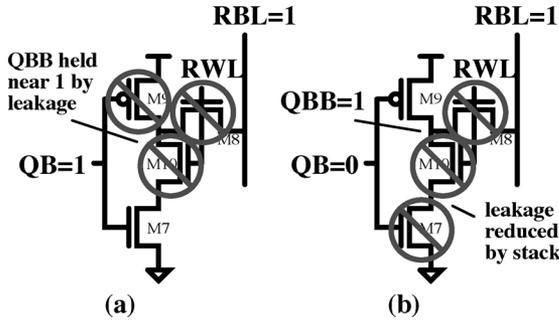


Fig. 6. Schematic of read buffer from 10 T bitcell for both data values. In both cases, leakage is reduced to the bitline and through the inverter relative to the case where M_{10} is excluded.

read and write ports. Since a 6 T bitcell does not have this feature, the 10 T bitcell is in some ways more fairly compared to an 8 T dual-port bitcell (6 T bitcell with two pairs of access transistors and bitlines).

A. Enabling Sub-threshold Read

The 10 T bitcell in Fig. 5 uses transistors M_7 – M_{10} to remove the problem of Read SNM by buffering the stored data during a read access. As described previously, eliminating the Read SNM problem allows this bitcell to operate at half of the V_{DD} of a 6 T cell while retaining the same 6σ stability. A different approach for eliminating the Read SNM in [18] uses a 7 T cell to prevent the higher voltage at the internal node from propagating to the other back-to-back inverter by holding its data dynamically during read accesses. This approach will not work in sub- V_T because the dynamic data is susceptible to leaking away during the long access times.

It is interesting to note that a 9 T bitcell, identical to the bitcell in Fig. 5 but without M_{10} , would eliminate the Read SNM problem while using less area than the 10 T cell. However, M_{10} is valuable to the bitcell because it reduces leakage current and allows more bitcells to share a bitline. Fig. 6 shows the read buffer from the 10 T bitcell for $Q = 0$ (a) and $Q = 1$ (b). When $Q = 0$ and $QB = 1$ Fig. 6(a), M_{10} adds an *off* device in series with the leakage path through M_8 and the path through M_9 , decreasing the leakage through those transistors. Furthermore, since the pMOS in this 65-nm technology generally has higher leakage than the nMOS, the leakage in M_9 holds node QBB near V_{DD} (see Fig. 7), further limiting the leakage through M_8 by making its V_{GS} negative. Even if QBB floats above 0 by only a small amount, the negative V_{GS} in M_8 reduces bitline leakage exponentially. When $Q = 1$ and $QB = 0$ Fig. 6(b), M_{10} reduces leakage through M_7 by the stack effect (note that the stack of devices will also slow down a read access by decreasing read current). Since node QBB is held solidly at V_{DD} , M_8 has $V_{DS} = 0$, so bitline leakage is negligible. In both cases, M_{10} reduces the leakage relative to the 9 T (and 6 T) case. The 10 T only has 16% more leakage than a 6 T cell at the same V_{DD} (9 T has 50% more). This overhead in leakage current is more than compensated by decreasing V_{DD} by several hundred millivolts relative to the 6 T bitcell. In simulation, the 10 T bitcell at

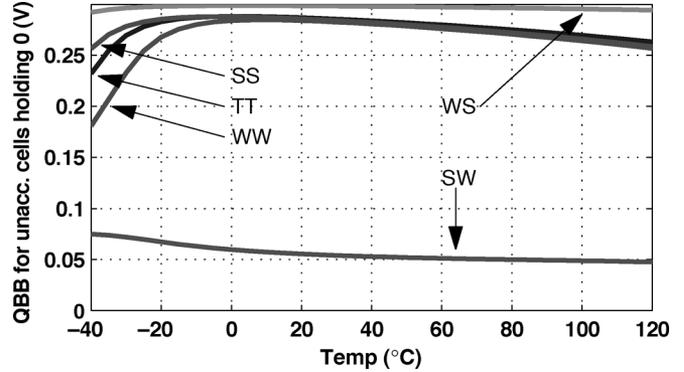


Fig. 7. Simulation of voltage at node QBB in unaccessed 10 T bitcells versus temperature and process corner. Strong pMOS leakage holds QBB near V_{DD} except at the SW corner. Even at SW, QBB is higher than it is for the 6 T cell, lowering bitline leakage.

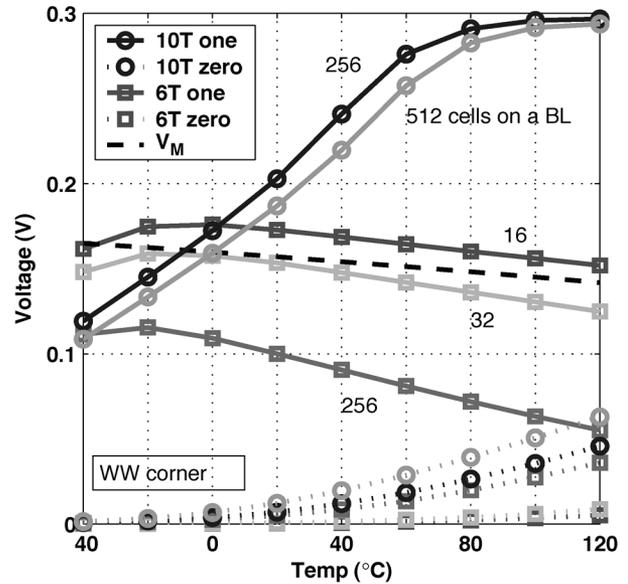


Fig. 8. Simulation showing steady-state bitline voltages. The 10 T bitcell exhibits much better steady-state bitline separation than the 6 T cell. The WW corner is shown at 300 mV.

300 mV consumes 2.25X less leakage power than the 6 T bitcell at 0.6 V [17].

The reduction in sub-threshold leakage through M_8 reduces the impact of leakage from unaccessed cells and gives the additional advantage of allowing more cells on a bitline during read. Leakage from the bitline into the unaccessed bitcells causes undesirable voltage drop that slows differential sensing and that makes single-ended read values difficult to distinguish. Fig. 8 shows the impact of bitline leakage on steady-state voltages (note that the bitline initially is precharged to V_{DD}) while reading a “1” (solid lines) or “0” (dotted lines) at 300 mV. For the same number of cells on a BL, the 10 T bitcell (circles) shows larger bitline separation than the 6 T (or 9 T) bitcells (squares). This figure suggests that “sensing” with an inverter (whose switching threshold, V_M , is shown) should work well from 0 °C to 100 °C even with 256 cells on a bitline for the 10 T cell. In contrast, the 6 T cell (or 9 T bitcell) would allow at most 16 bitcells on a bitline. The bitline that should be “1” stays very

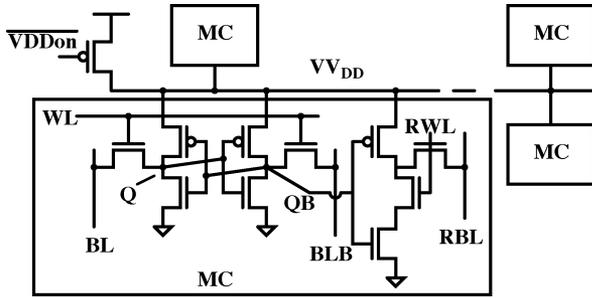


Fig. 9. Schematic of write architecture for a single row using a floating power supply (V_{VDD}). The row is “folded” in layout so that its cells share n-wells, and the entire row is written at once.

close to V_{DD} at high temperatures and then begins to droop at lower temperatures. This occurs because M_{10} inside the unaccessed 10 T bitcells is so successful at reducing sub- V_T current through the access transistors that the sub- V_T current actually drops below the sum of gate currents (which is fairly constant with temperature) into unaccessed cells. If gate leakage was lower (e.g., high-K dielectrics), then sub-threshold leakage into the unaccessed cells is reduced sufficiently such that the bitline will stay very close to V_{DD} . One advantage of more cells on a BL is a reduction in peripheral circuits that offsets some of the area overhead of larger bitcells. For example, an 8 T bitcell (40% larger than 6 T) that allows 256 cells per bitline rather than 16 (same improvement as our cell) actually resulted in a 6% smaller overall array area [19].

B. Enabling Sub-threshold Write

In this 65-nm technology, a 6 T bitcell cannot write in the traditional fashion below around 0.6 V because the nMOS access transistor cannot reliably win the ratioed fight against the pMOS to write a “0”. The technique of weakening the cross-coupled inverters by gating their supply voltage (e.g., [6]) or ground node (e.g., [20]), applied by previous works primarily to improve speed, can dramatically improve write margin. Fig. 9 shows the schematic for a single row using this approach. A single power-supply-gating header switch connects node V_{VDD} to the true power rail. When the bitcell holds its data or during read accesses, $\overline{V_{DDon}} = 0$ so that $V_{VDD} = V_{DD}$. During a write access, the virtual rail floats. For the implementation on the test chip, a conceptual row folds as shown in the figure so that its bitcells can share n-wells, and the entire row is written at once.

Fig. 10 shows the timing associated with a write access using this scheme. First, the write signal goes high to indicate that a write access will occur, and the bitlines are driven with the new data. Next, the decoders drive a global wordline (not shown) which causes the proper local write wordline (WL_{WR}) to go high. Triggered by the local wordline, the $\overline{V_{DDon}}$ signal goes high, allowing node V_{VDD} to float. As the write access transistors discharge the virtual rail, its voltage droops, and Q and QB change to their new values. The logical “1” inside the cell tracks the drooping voltage until $\overline{V_{DDon}}$ goes low again while the *local wordline remains high*, and the virtual rail reconnects to V_{DD} . The feedback inside the bitcell then holds the Q and QB nodes at their correct logical values and amplifies the “1” to full V_{DD} (point (a) in Fig. 10). The plot in Fig. 11 shows the write margin

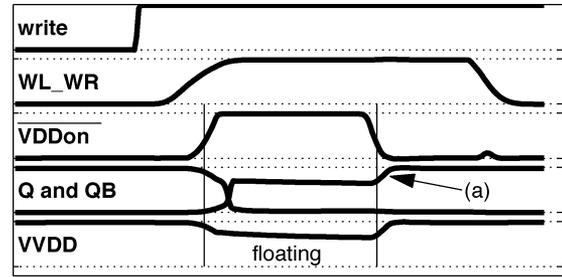


Fig. 10. Timing diagram for write operation. When $\overline{V_{DDon}}$ goes low while WL_{WR} remains asserted, the cell’s feedback restores full voltage levels for the new values of Q and QB (point (a)).

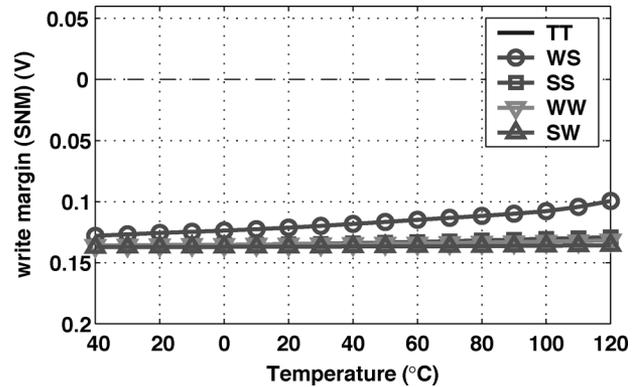


Fig. 11. Write margin (write SNM) versus temperature at 0.3 V for 10 T bitcell with floating V_{VDD} supply. Negative margin for all corners, signifying successful write operation.

for the virtual V_{DD} approach across temperature and process corner at $V_{DD} = 300$ mV. The margin remains negative across all of these ranges, indicating a successful write.

IV. 65-nm SUB-THRESHOLD SRAM TEST CHIP

A. Test Chip Architecture

A 256-kb 65-nm bulk CMOS test chip uses the 10 T bitcell and the architecture shown in Fig. 12. The memory has eight 32-kb blocks with 256 rows and 128 columns each. A single 128-bit DIO bus serves all eight blocks. In this initial instantiation of the sub-threshold memory, only one read or write can occur per cycle, however the 10 T bitcell would allow a read and write access to the same block in one cycle. Such a dual-port instantiation of the memory would require a second DIO bus and additional peripheral logic. A combined global wordline and block select signal assert a local wordline that triggers either WL_{RD} or WL_{WR} . For a write access, $M_P\langle r \rangle$ for the accessed row turns off. The write drivers consist simply of inverters with transmission gates, which turn off when the memory is not writing to minimize leakage on the write bitlines (BL and BLB). The power supply to the WL drivers is routed separately to allow a boosted WL voltage. This technique improves the access speed and increases the robustness to local variations. The read bitline (RBL) is precharged prior to read access, and its steady-state value is “sensed” using a simple inverter, $I_{RD}\langle c \rangle$. Column and row redundancy is a ubiquitous technique in commercial memories used to improve yield. For

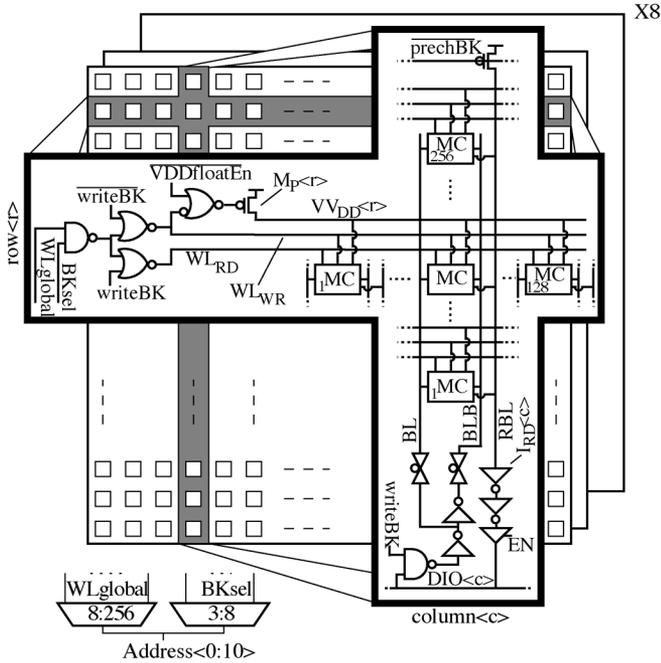


Fig. 12. Architecture diagram of the 256-kb memory on the test chip using 10 T sub-threshold bitcells.

our analysis of the SRAM, we assume the availability of one redundant row and column per block.

The primary goals for this test chip were to test the functionality of the 10 T bitcell in sub-threshold and to explore the limitations of the design. For this reason, all of the peripherals use static CMOS logic for simplicity and for functional robustness in sub-threshold. The large block size was intentionally aggressive to expose limitations in the bitcell and architecture. Integrating 256 bitcells on the bitline (as opposed to 16 for 6 T) pushes the envelope for functionality. The 10 T bitcell layout added 66% area overhead relative to our reference 6 T design, but the overall area penalty will be less due to more bitcells on a bitline, as described in Section III-A. Each row is folded such that a pair of 64-bit physical rows sharing n-wells and a V_{DD} rail makes up one conceptual 128-bit row (c.f. Fig. 9). This folding increases the length of bitlines by roughly 2X and decreases the length of wordlines by roughly (1/2)X. Notice that this is not fundamentally necessary for the write approach to work. The n-wells of two separate rows can be shared and the V_{DD} for each row routed separately. The 10 T architecture does not change the number of WLS or BLs or the number of devices per line relative to the 6 T case, except that it has one fewer read BL. The capacitance of the metal lines themselves will increase somewhat due to the larger bitcell area. Fig. 13 shows a layout shot and die photograph of the test chip (1.89 mm by 1.12 mm, pin-limited).

B. Measurements

Measurements of the SRAM test chip confirm that it is functional over a range of voltages from 1.2 V down well into the sub-threshold region. With the assumption of one redundant row and column per block, read operation works without error to 320 mV and write operation works without error to 380 mV at

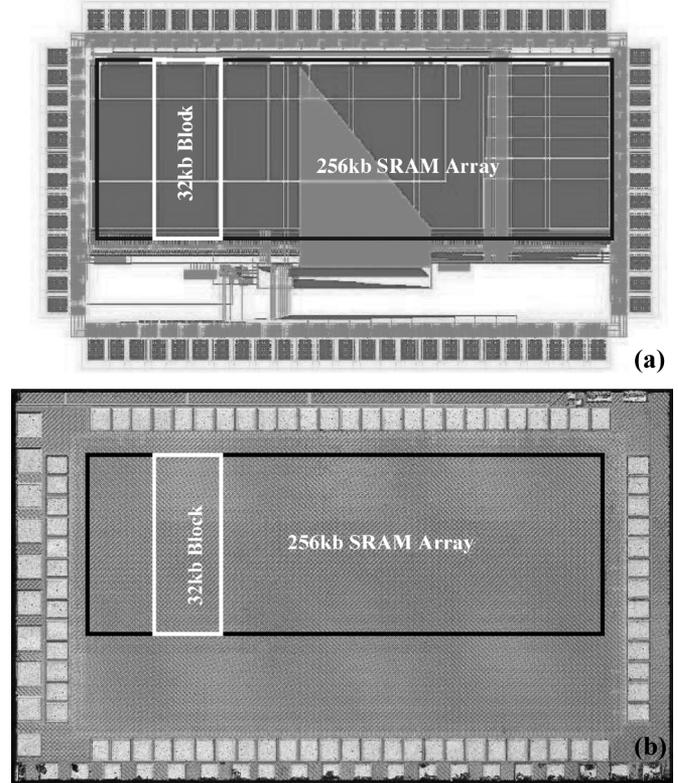


Fig. 13. (a) Annotated layout and (b) die photograph of the 256-kb sub-threshold SRAM in 65 nm. Die size is 1.89 mm by 1.12 mm.

27 °C. We continued to push the supply voltage to even lower values to examine the limits of the implementation. At the low supply voltage of 300 mV, the memory continues to function, but it exhibits bit errors in $\sim 1\%$ of its bits that result from sensitivities in the architecture to local device variation, as described later.

The test chip successfully demonstrates a functional sub-threshold memory that overcomes the problems it was designed to face. First, the bitcell removes the Read SNM problem. Measurements have confirmed that the memory experiences zero destructive read errors at 300 mV. Simulations show that a 6 T memory would experience a high rate of destructive read errors at 300 mV due to degraded Read SNM. Second, whereas a 6 T memory would fail to write below about 600 mV, this memory writes correctly at 350 mV at 85 °C. Third, a 6 T memory would experience problems reading with only 16 bitcells on a bitline. Measurements show that the 10 T memory reads correctly even with 256 bitcells on the bitline down to 320 mV. Finally, the memory shows good Hold SNM performance. The first bits observed to fail to hold their data occur at $V_{DD} < 230$ mV, as seen in the distribution shown in Fig. 14.

Fig. 15 shows the measured leakage power of the test chip at two different temperatures and the expected savings from V_{DD} scaling. At 27 °C, the 10 T memory saves 2.5X and 3.8X in leakage power by scaling from 0.6 V to 0.4 V and 0.3 V, respectively and over 60X when V_{DD} scales from 1.2 V to 0.3 V. V_{DD} scaling also gives the expected savings in active energy per read

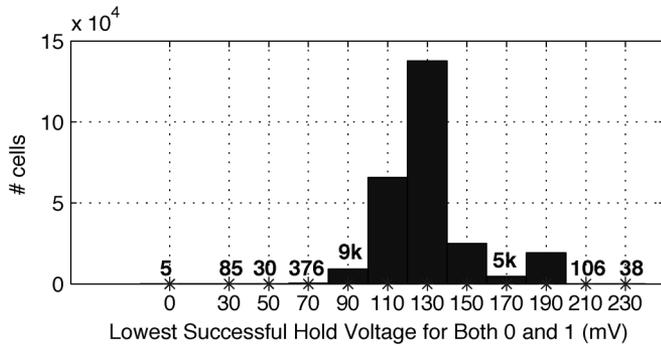


Fig. 14. Measured distribution of minimum voltage at which bitcells hold both “0” and “1”.

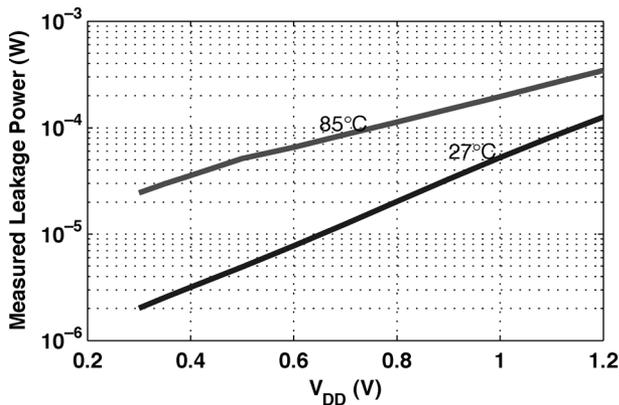


Fig. 15. Measured leakage power from the memory test chip.

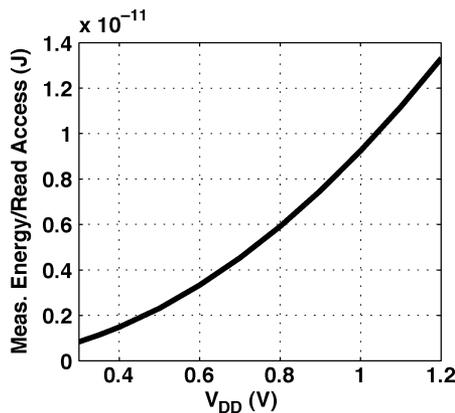


Fig. 16. Measured active energy per read access.

access, as shown in Fig. 16. Fig. 17 shows the measured frequency of operation versus V_{DD} (the 1.2 V speed of 200 MHz is a simulation result, because the testing board did not support high-speed testing). The maximum measured operating speed at 400 mV is 475 kHz.

Pushing V_{DD} even lower exposes the limitations for both read and write operations as a small fraction of bits begins to fail. These failures occur for the same set of bits in a repeatable fashion. For certain bits at low voltage, a read access shows that the bitcell holds a “0” when in fact it holds a “1”. This error is non-destructive, which we confirm by raising the supply

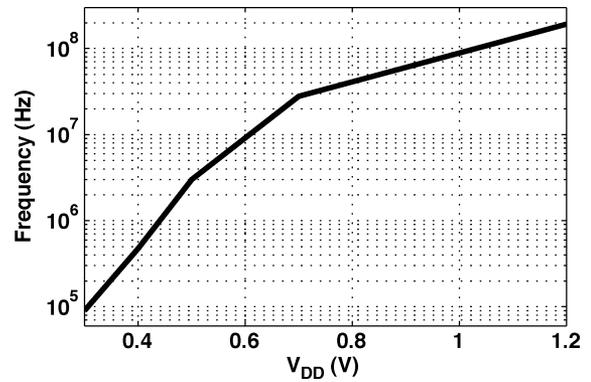


Fig. 17. Measured frequency of operation versus V_{DD} .

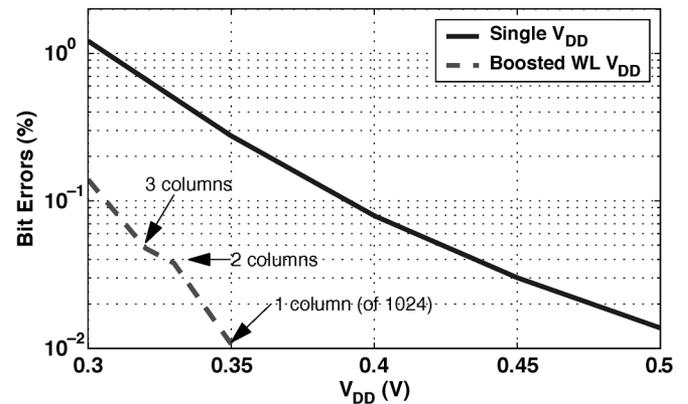


Fig. 18. Measured percentage of bit errors for read versus V_{DD} . Boosting the WL voltage dramatically reduces these errors.

voltage and re-reading the cell and invariably reading the correct value. Also, these bit errors tend to gather along a small number of specific columns. The fact that this error exists in a small fraction of cases indicates that it results from local device variation, and we can isolate the problematic transistors. The fact that the bits exhibiting problems cluster along specific columns indicates that variation in the sensing inverter, I_{RD} in Fig. 12, has shifted its switching threshold, V_M , towards V_{DD} . Now, specific bitcells along this column that have read access transistors weakened by local variation cannot hold the read bitline above V_M of inverter I_{RD} . Several experiments confirm that this is the mechanism for read failures. First, we can independently lower the supply voltage of the sense inverters, $I_{RD}(c)$. This lowers the V_M for the inverters, and the measured bit error rate decreases. Secondly, we can increase temperature, which provides the expected improvement in discerning a “1” (c.f. Fig. 8). Finally, we can increase the voltage of the wordline drivers, which pulls the transistors that weakened by local variation back toward the mean and rapidly decreases the read bit errors. Fig. 18 shows the measured percentage of bit errors during read access versus V_{DD} . The error rate without the wordline boosted by 100 mV also is shown. In sub-threshold, the extra gate voltage on the read access transistors provides over 10X (due to the sub-threshold slope) of extra current drive. As with above-threshold memories, the extra current speeds operation. It also makes the design more robust to mismatch.

By aggressively choosing a block having 256 rows on a single bitline, we pushed the limits of read operation and exposed the limits to scaling read accesses that result from local variation. Also, using a simple inverter for sensing makes it harder to read a “1” correctly. As the bitline separation plots have shown, V_M of the sensing inverter lies too close to the logical “1” value at some corners and temperatures. Boosting the wordline voltage offers one simple change that dramatically reduces the error rate and allows this memory to read without error at 320 mV. A better solution to improve the read reliability and robustness to local device variation is to replace the inverter with a new sensing scheme, for which many relevant sense amps are available in the literature.

The limit to V_{DD} scaling for write manifests when write accesses fail for specific bitcells. Write functionality was tested using a high voltage write, a low-voltage write of the opposite value, and finally a high voltage read. This test isolates the bits for which sub-threshold write fails. These errors aggregate in bits along specific rows. As with the read errors, local device variation is the culprit, and the predominance of row-wise errors suggests that the failure mechanism involves the row peripherals. Referring back to Fig. 12, write limitations first appear along specific rows whose pull-up device, M_P , is strengthened by local variation. Thus, when M_P turns off during a write access, its larger leakage pulls V_{DD} up closer to V_{DD} . Some of the bitcells can still switch under these conditions, but V_{DD} reaches a steady-state voltage that is high enough to prevent some bitcells from overpowering the pMOS to write a ‘0’ into the memory. In these bitcells, local mismatch has made the internal pMOS relatively stronger than the access transistor to the point that the write driver cannot flip the cell at the steady-state V_{DD} voltage. Measurements confirm that this is the case. First, the lowest functional supply voltage decreases at higher temperature. Since the leakage through M_P gets relatively weaker compared with the nMOS access transistors, this confirms the mechanism for failure. More importantly, the write errors decrease when the supply voltage to the wordline increases. The higher wordline voltage increases V_{GS} for the write access transistors and makes them more capable of producing voltage droop on V_{DD} . Fig. 19 shows the percentage of bit errors measured during write both with and without 100 mV of wordline boosting. With boosting, the memory can write without error at 380 mV at 27 °C and 350 mV at 85 °C.

As with the limitations on read, simple changes to the peripheral circuits can push the lowest operational V_{DD} even lower. Specifically, the leakage through M_P can be reduced using one of several well-known methods (e.g., stacking, RBB, etc.) A better solution to the write issue that maintains the same basic architecture and approach is to induce a specific voltage drop on V_{DD} intentionally. In the extreme, replacing M_P with an inverter will drive V_{DD} all the way to 0 V. Then, as long as the write wordline remains asserted, the bitcells will develop the correct internal data when V_{DD} goes back high regardless of local variations. A disadvantage of this extreme case is the energy penalty associated with discharging and re-powering the V_{DD} rail and all of the bitcells in the row. An alternative is to use a circuit (e.g., diode connected FET) to force V_{DD} to

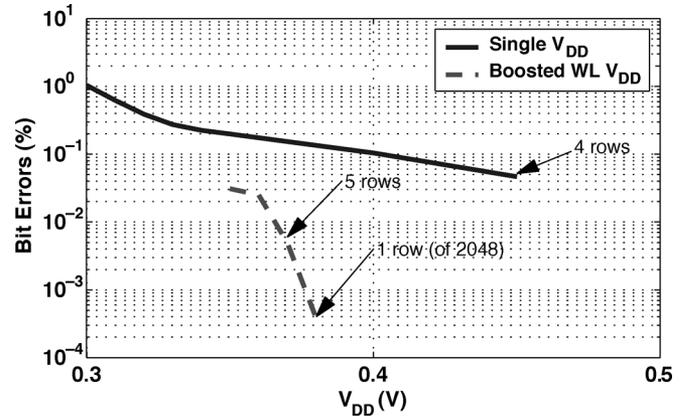


Fig. 19. Measured percentage of bit errors for write versus V_{DD} . Again, boosting the WL voltage dramatically reduces these errors.

some intermediate value that is low enough to ensure write but that uses less energy.

V. SUMMARY AND CONCLUSIONS

Sub-threshold SRAM provides the dual advantages of minimizing total memory energy consumption and of providing compatibility with minimum-energy sub-threshold logic. Traditional 6 T SRAM cannot function in sub-threshold because it fails to write and because the Read SNM degrades badly. Furthermore, bitline leakage in 6 T SRAMs limits the number of bitcells on a bitline to 16. Measurements of a 256-kb 65-nm bulk CMOS test chip show that our 10 T bitcell fundamentally solves the Read SNM problem, overcomes the write problem, and relaxes the bitline integration limitation to allow sub-threshold operation. With one redundant row and column per block and a boosted wordline, the memory functions without error to below 380 mV. At 400 mV, it consumes 3.28 μW and works up to 475 kHz. Although aggressive design exposes the limitations of the architecture in terms of its robustness to local device variation, the bit errors result primarily from problems in the peripheral circuits. Simple proposed changes to the periphery promise to push the limits of SRAM operation to even lower V_{DD} .

REFERENCES

- [1] A. Wang and A. Chandrakasan, “A 180 mV FFT processor using sub-threshold circuit techniques,” in *IEEE ISSCC Dig. Tech. Papers*, 2004, pp. 292–293.
- [2] N. Kim, K. Flautner, D. Blaauw, and T. Mudge, “Circuit and microarchitectural techniques for reducing cache leakage power,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 2, pp. 167–184, Feb. 2004.
- [3] K. Osada, J. L. Shin, M. Khan, Y. Liou, K. Wang, K. Shoji, K. Kuroda, S. Ikeda, and K. Ishibashi, “Universal-Vdd 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, pp. 1738–1744, Nov. 2001.
- [4] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, “SRAM leakage suppression by minimizing standby supply voltage,” in *Proc. ISQED*, 2004, pp. 55–60.
- [5] K. Kanda, T. Miyazaki, M. K. Sik, H. Kawaguchi, and T. Sakurai, “Two orders of magnitude leakage power reduction of low-voltage SRAM’s by row-by-row dynamic V_{DD} control (RRDV) scheme,” in *Proc. IEEE Int. ASIC/SOC Conf.*, Sep. 2002, pp. 381–385.
- [6] A. Bhavnagarwala, S. Kosonocky, S. Kowalczyk, R. Joshi, Y. Chan, U. Srinivasan, and J. Wadhwa, “A transregional CMOS SRAM with single, logic V_{DD} and dynamic power rails,” in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 292–293.

- [7] H. Yamauchi, T. Iwata, H. Akamatsu, and A. Matsuzawa, "A 0.8 V/100 MHz/sub-5 mW-operated mega-bit SRAM cell architecture with charge-recycle offset-source driving (OSD) scheme," in *Symp. VLSI Circuits Dig. Tech. Papers*, 1996, pp. 126–127.
- [8] K. Osada, Y. Saitoh, E. Ibe, and K. Ishibashi, "16.7-fA/cell tunnel-leakage-suppressed 16-Mb SRAM for handling cosmic-ray-induced multierrors," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1952–1957, Nov. 2003.
- [9] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Yang, B. Zheng, and M. Bohr, "A SRAM design on 65 nm CMOS technology with integrated leakage scheme," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 294–295.
- [10] A. Agarwal, H. Li, and K. Roy, "A single- V_t low-leakage gated-ground cache for deep submicron," *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 319–328, Feb. 2003.
- [11] T. Enomoto, Y. Oka, and H. Shikano, "A self-controllable voltage level (SVL) circuit and its low-power high-speed CMOS circuit applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 7, pp. 1220–1226, Jul. 2003.
- [12] K. Osada, K. Yamaguchi, Y. Saitoh, and T. Kawahara, "SRAM immunity to cosmic-ray-induced multierrors based on analysis of an induced parasitic bipolar effect," *IEEE J. Solid-State Circuits*, vol. 39, no. 5, pp. 827–833, May 2004.
- [13] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, "Low-power embedded SRAM modules with expanded margins for writing," in *IEEE ISSCC Dig. Tech. Papers*, 2005, pp. 480–481.
- [14] M. Yamaoka, K. Osada, R. Tsuchiya, M. Horiuchi, S. Kimura, and T. Kawahara, "Low power SRAM menu for SOC application using Yin-Yang-feedback memory cell technology," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2004, pp. 288–291.
- [15] E. Seevinck, F. List, and J. Lohstroh, "Static noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 5, pp. 748–754, Oct. 1987.
- [16] B. Calhoun and A. Chandrakasan, "Analyzing static noise margin for sub-threshold SRAM in 65 nm CMOS," in *Proc. ESSCIRC*, 2005, pp. 363–366.
- [17] B. Calhoun and A. Chandrakasan, "A 256 kb sub-threshold SRAM in 65 nm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, 2006, pp. 628–629.
- [18] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, "A read-static-noise-margin-free SRAM cell for low- V_{dd} and high-speed applications," in *IEEE ISSCC Dig. Tech. Papers*, 2005, pp. 478–479.
- [19] A. Alvandpour, D. Somasekhar, R. Krishnamurthy, V. De, S. Borkar, and C. Svensson, "Bitline leakage equalization for sub-100 nm caches," in *Proc. ESSCIRC*, 2003, pp. 401–404.
- [20] K. Itoh, A. Fridi, A. Bellaouar, and M. Elmasry, "A deep sub- V_t single power-supply SRAM cell with multi- V_T , boosted storage node and dynamic load," in *Symp. VLSI Circuits Dig. Tech. Papers*, 1996, pp. 132–133.



Benton Highsmith Calhoun (S'05–M'06) received the B.S. degree in electrical engineering with a concentration in computer science from the University of Virginia, Charlottesville, VA, in 2000, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, in 2002 and 2006, respectively.

In January 2006, he joined the faculty at the University of Virginia as an Assistant Professor in the Electrical and Computer Engineering Department. His research interests include low-power digital circuit design, sub-threshold digital circuits, SRAM design for end-of-the-roadmap silicon, variation tolerant circuit design methodologies, and low-energy electronics for medical applications. He is a coauthor of *Sub-threshold Design for Ultra Low-Power Systems* (Springer, 2006).

Dr. Calhoun serves on the Technical Program Committee for the International Symposium on Low Power Electronics and Design (ISLPED).



Anantha P. Chandrakasan (M'95–SM'01–F'04) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer sciences from the University of California, Berkeley, in 1989, 1990, and 1994, respectively.

Since September 1994, he has been with the Massachusetts Institute of Technology, Cambridge, where he is currently the Joseph F. and Nancy P. Keithley Professor of Electrical Engineering. His research interests include low-power digital integrated circuit design, wireless microsensors, ultra-wideband radios, and emerging technologies. He is a coauthor of *Low Power Digital CMOS Design* (Kluwer, 1995) and *Digital Integrated Circuits* (Pearson Prentice-Hall, 2003, 2nd edition). He is also a co-editor of *Low Power CMOS Design* (IEEE Press, 1998), *Design of High-Performance Microprocessor Circuits* (IEEE Press, 2000), *Leakage in Nanometer CMOS Technologies* (Springer, 2005), and *Sub-threshold Design for Ultra Low-Power Systems* (Springer, 2006).

Dr. Chandrakasan has received several awards including the 1993 IEEE Communications Society's Best Tutorial Paper Award, the IEEE Electron Devices Society's 1997 Paul Rappaport Award for the Best Paper in an EDS publication during 1997, the 1999 Design Automation Conference Design Contest Award, and the 2004 DAC/ISSCC Student Design Contest Award. He has served as a technical program co-chair for the 1997 International Symposium on Low Power Electronics and Design (ISLPED), VLSI Design'98, and the 1998 IEEE Workshop on Signal Processing Systems. He was the Signal Processing Subcommittee Chair for ISSCC 1999–2001, the Program Vice-Chair for ISSCC 2002, the Program Chair for ISSCC 2003, and the Technology Directions Subcommittee Chair for ISSCC 2004–2006. He was an Associate Editor for the IEEE JOURNAL OF SOLID-STATE CIRCUITS from 1998 to 2001. He serves on the SSCS AdCom and is the meetings committee chair. He is the Technology Directions Chair for ISSCC 2007.