A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy

Naveen Verma, *Student Member, IEEE*, and Anantha P. Chandrakasan, *Fellow, IEEE*

***Abstract*—Aggressively scaling the supply voltage of SRAMs greatly minimizes their active and leakage power, a dominating portion of the total power in modern ICs. Hence, energy constrained applications, where performance requirements are secondary, benefit significantly from an SRAM that offers read and write functionality at the lowest possible voltage. However, bit-cells and architectures achieving very high density conventionally fail to operate at low voltages. This paper describes a high density SRAM in 65 nm CMOS that uses an 8T bit-cell to achieve a minimum operating voltage of 350 mV. Buffered read is used to ensure read stability, and peripheral control of both the bit-cell supply voltage and the read-buffer's foot voltage enable sub-V_t write and read without degrading the bit-cell's density. The plaguing area-offset tradeoff in modern sense-amplifiers is alleviated using redundancy, which reduces read errors by a factor of five compared to device up-sizing. At its lowest operating voltage, the entire 256 kb SRAM consumes 2.2 μW in leakage power.**

***Index Terms*—Cache memories, CMOS memory circuits, leakage currents, low-power electronics, redundancy, SRAM chips.**

I. Introduction

VOLTAGE scaling affords significant advantages in digital circuits by virtue of the CV_{DD}^2 active energy savings. Of course, the ensuing reduction in the operating speed also means that a particular circuit block takes longer to complete a required operation. As a result, the leakage power, which comes about as a result of the idle sub-V_t currents, integrates over a longer time, and the leakage energy goes up. This opposing trend give rise to a minimum energy supply voltage, which, for most practical digital circuits, occurs below the threshold voltage of the devices [1]. Importantly, however, this argument assumes the circuit can be operated at exactly the optimal speed, and then be shut off after the operation is complete, to eliminate the leakage power. Unfortunately, however, SRAMs often need to retain or buffer their data for some length of time unrelated to their own access period and cannot be shut off accordingly. In this case, minimizing leakage power is critical, and voltage scaling is even more powerful, since it significantly reduces the leakage current by alleviating drain-induced barrier lowering (DIBL). For instance, in a 65 nm process, the leakage current reduction from a V_{DD} of 1 V to 0.3 V due to DIBL is over 4x, and the leakage power savings is over 10x. Low-voltage standby modes help in this manner, but are limited in their power reduction due to active-switching and leakage during operational modes. Accordingly, this paper presents a sub-V_t SRAM that provides full read and write operation down to 350 mV [2]. That voltage makes it compatible with minimum-energy logic, but, more importantly, also minimizes the active and leakage power of the array. Previous instances of sub-V_t SRAMs have achieved ultra-low-voltage operation by adding devices within the cell, or employing hierarchy to approach standard logic topologies. For instance, a 10T cell operates at 400 mV [3], and register-file structures use multiplexed reads to operate at 310 mV [4] and 180 mV [5]. However, this design maximizes the cell density and relies on peripheral circuit assists to resolve sub-V_t design challenges. Finally, the difficulties with sense-amplifier scaling in the presence of variation induced offsets are becoming more pronounced in advanced technologies, and are stressed in this design. Hence, an alternative to device up-sizing is proposed.

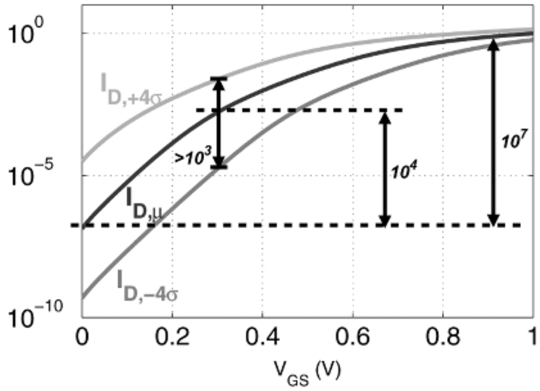


Fig. 1. I_D versus V_{GS} behavior of a MOSFET in a 65 nm technology.

The following sections start by describing the challenges of sub-V_t design specific to SRAMs, and then present the specific circuits employed in this design, including the 8T bit-cell and its peripheral circuit assists. Finally, the technique of sense-amp redundancy is discussed, and prototype results are presented.

II. Subthreshold SRAM Design Challenges

Fig. 1 shows the I_D versus V_{GS} behavior of a MOSFET, where the drain current increases exponentially in sub-V_t and far more slowly in strong inversion. Two effects are shown that are of critical importance to SRAMs in the sub-V_t regime; the first is threshold voltage variation, and the second is the degradation in the on-to-off ratio of the current.

Manuscript received March 14, 2007; revised June 13, 2007. This work was supported by the Defense Advanced Research Projects Agency (DARPA). IC fabrication was provided by Texas Instruments Incorporated. The work of N. Verma was supported by the Intel Foundation Ph.D. Fellowship Program and NSERC.

The authors are with the Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: nverma@mtl.mit.edu; anantha@mtl.mit.edu).

Digital Object Identifier 10.1109/JSSC.2007.908005

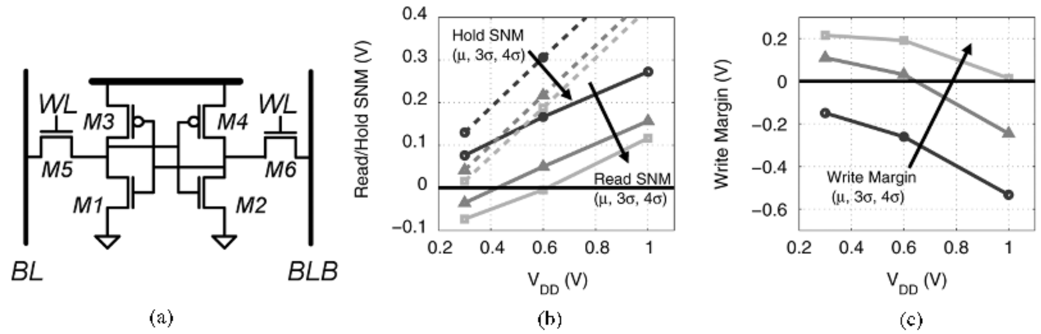


Fig. 2. Functionality of (a) conventional 6T bit-cell is lost at low voltages due to (b) read/hold SNM failures and (c) write margin failures.

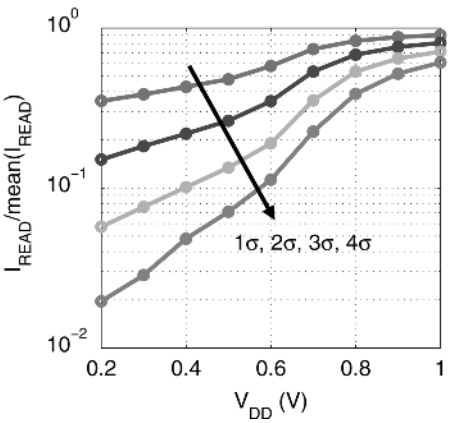


Fig. 3. In sub-V_t, mean I_{READ} is greatly reduced, but normalized I_{READ} distribution shows that further degradation due to variation is also severe.

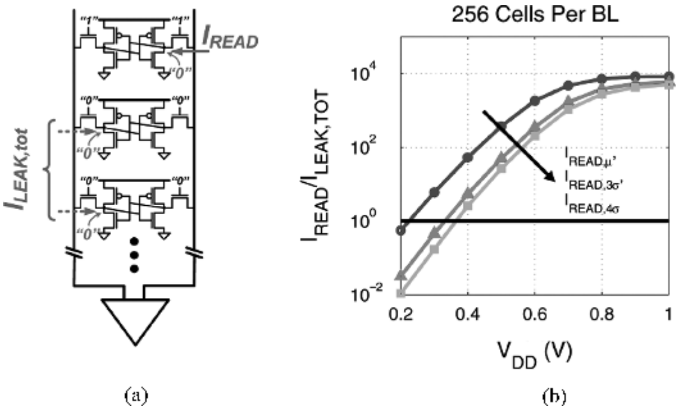


Fig. 4. Bit-line leakage depends on the data stored in the unaccessed cells and is worst for the case (a) where the leaking devices have a large voltage drop; (b) the worst-case leakage exceeds the weak-cell I_{READ} at low voltages.

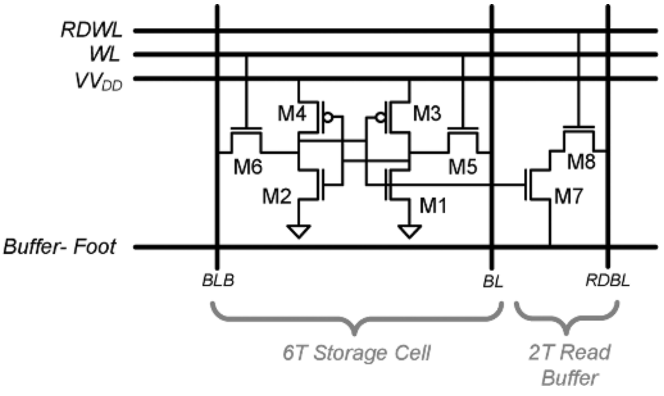


Fig. 5. 8T bit-cell uses two-port topology to eliminate read SNM and peripheral assists, controlling $\mathrm{Buffer–Foot}$ and VV_{DD}, to manage bit-line leakage and write errors.

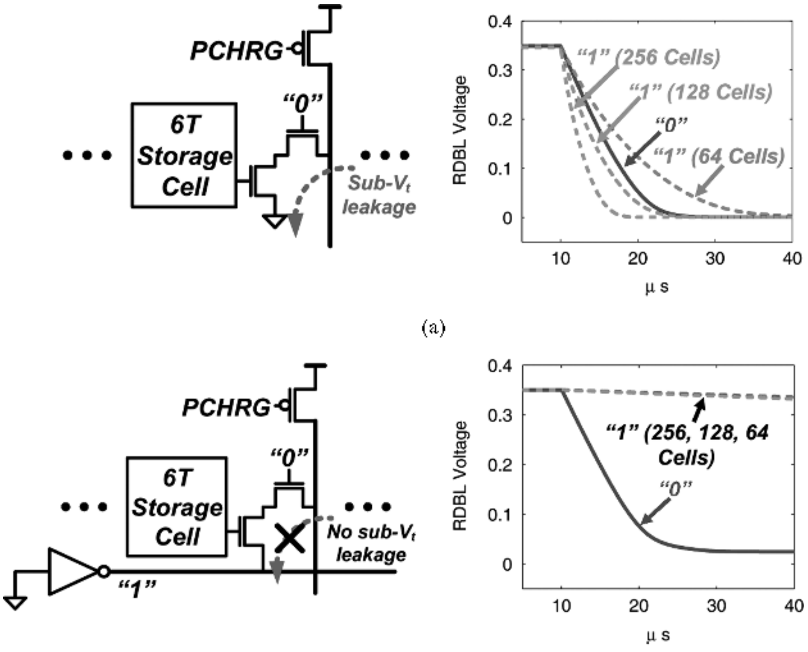
Threshold voltage shifts, a prominent result of processing variation and random dopant fluctuation [6], [7], effectively cause the sideways offsets shown in Fig. 1. For the case of $\pm 4\sigma$ variation, which occurs commonly for devices in a large SRAM array, the resulting change in the sub-V_t drain current (e.g., at 0.3 V), is over three orders of magnitude. Accordingly, relative device strengths cannot reliably be set using conventional techniques like W/L sizing.

The degradation in $I_{\mathrm{ON}}/I_{\mathrm{OFF}}$, from approximately 10^7 to 10^4, implies that, in sub-V_t, there is a strong interaction between the "on" and the "off" devices when it comes to setting the voltage level of critical signals. Of course, this introduces a relevant failure mechanism where SRAM density requirements call for the integration of many devices on shared nodes. The following subsections more specifically relate these fundamental effects in sub-V_t MOSFETs to the challenges of SRAM design.

A. 6T Bit-Cell Failures

The 6T bit-cell, shown in Fig. 2(a), fails to operate in sub-V_t because of reduced signal levels and increased variation [8]. The ratioed nature of this circuit implies that proper read and write operation depends on the relative strengths of the devices. For instance, the read static noise margin (SNM) [9] requires that the driver devices, $M1/M2$, be stronger than the access devices, $M5/M6$, and, as shown in the Monte Carlo simulations of Fig. 2(b), at low voltages it vanishes and becomes negative. Similarly, the write margin characterizes the ability of the access devices to over-power the load devices, $M3/M4$, and, once again, in Fig. 2(c) it vanishes at low voltages, where, in this case, it is positive.

The hold SNM, however, depends on the basic storage structure composed of the cross-coupled inverters $(M1–M4)$. Fig. 2(c) shows that at the target voltage of 350 mV, hold stability is preserved. Accordingly, in this design, peripheral assists and the bit-cell topology eliminate the read and write limitations so that V_{MIN} can approach the limit set by the hold SNM.



(a)

(b)

Fig. 6. Read-buffer bit-line leakage in (a) conventional case where unaccessed read-buffer foot is statically connected to ground and (b) this design where unaccessed read-buffer foot is pulled up to V_{DD}.

B. Read-Current Distribution

In sub-V_t, we expect a significantly lower read-current, $I_{\rm READ}$, because of the low gate drive. However, the exponential effect of V_t variation severely degrades the weak-cell $I_{\rm READ}$ even further. Fig. 3 normalizes the $I_{\rm READ}$ distribution by the mean $I_{\rm READ}$ to show just the degradation due to variation. In sub-V_t, where the mean $I_{\rm READ}$ is already greatly reduced, the effect is particularly pronounced, and the weak-cell $I_{\rm READ}$ can easily be further degraded by a couple of orders of magnitude.

C. Bit-Line Leakage

A related consequence of the reduced $I_{\rm READ}$ is that the aggregate leakage currents from the unaccessed cells sharing the bit-lines can make conventional data sensing impractical. Typically, we differentially detect a droop on either BL or BLB and expect the alternate bit-line to dynamically remain high. However, as shown in Fig. 4(a), the aggregate leakage currents on the alternate bit-line can exceed $I_{\rm READ}$, depending on the data stored in the unaccessed bit-cells. The problematic leakage currents are maximized for the case shown, were a large voltage drop appears across the leaking devices, and Fig. 4(b) plots the weak-cell $I_{\rm READ}$ normalized to that total leakage current, $I_{\rm LEAK,tot}$ (assuming 256 cells per bit-line). At low voltages, $I_{\rm LEAK,tot}$ exceeds $I_{\rm READ}$ making the droop on the two bit-lines indistinguishable.

III. 8T Subthreshold Bit-Cell

To address the challenges of sub-V_t SRAM, the bit-cell shown in Fig. 5 is used. This two-port cell topology has a 6T storage cell and a 2T read-buffer which isolates the data-retention structure during read-accesses. Consequently, the read SNM limitation from Section II-A is eliminated [10]. The other two prominent limitations, namely bit-line leakage and writeability in the presence of variation, are dealt with using the peripheral assists associated with the Buffer–Foot and VV_{DD} controls.

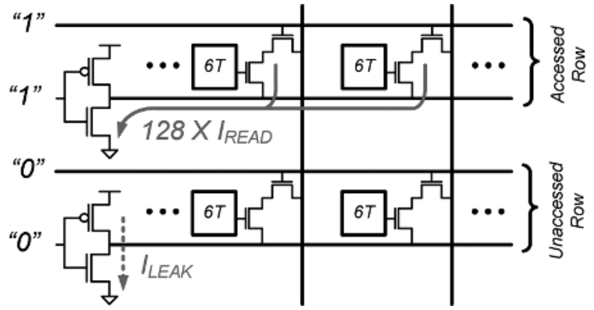


Fig. 7. Read-buffer must sink $I_{\rm READ}$ from all bit-cells in accessed row, and it draws leakage current in all unaccessed rows.

A. "Zero" Leakage Read-Buffer

The bit-line leakage problem in the single-ended 8T cell is analogous to the problem in the 6T case, except that the leakage currents from the unaccessed cells and $I_{\rm READ}$ from the accessed cell affect the same node, *RDBL*. So, the leakage currents can pull down *RDBL* regardless of the accessed cell's state. Fig. 6(a) shows transient simulations where *RDBL* is correctly pulled low by the accessed cell in the solid curve, but it is also erroneously pulled low, by the leakage currents of the unaccessed cells, in the dotted curves. Here, only the case with 64 cells on the *RDBL* results in a minimal sampling window, severely limiting the achievable integration.

In this design, however, the feet of all the unaccessed read-buffers are pulled up to V_{DD}, as shown in Fig. 6(b). Consequently, after *RDBL* is precharged, the read-buffer devices have

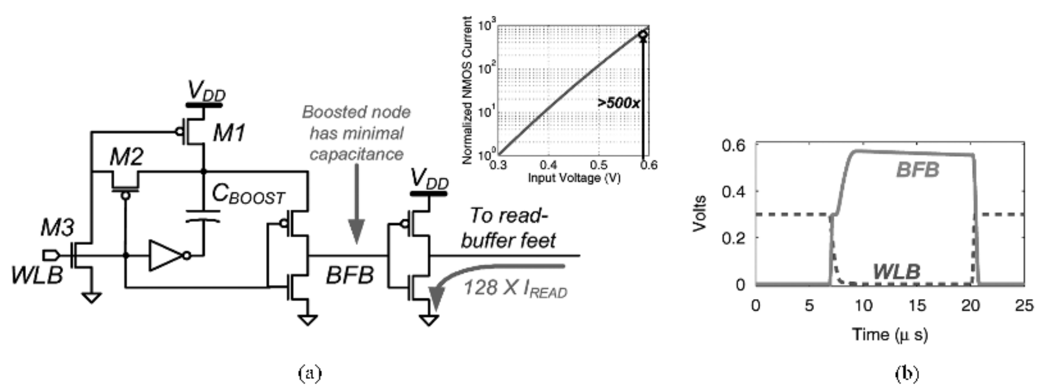


Fig. 8. To resolve read-buffer footer limitation, (a) charge-pump circuit is used and (b) BFB node gets bootstrapped to approximately $2V_{DD}$, increasing current drive of footer by over 500x.

no voltage drop across them, and they sink no sub-V_t leakage current. The transient simulations in Fig. 6(b) now show that *RDBL* correctly remains high in the dotted curves even when 256 cells are integrated. Some residual droop is still visible; however, this comes about as a result of gate leakage from the read-buffers' access devices and junction leakage from their drains.

An important concern with this approach is that the peripheral nMOS footer device needs to sink I_{READ} from all cells in the accessed row. As shown in Fig. 7, this design has 128 cells per row, making the current requirement of the footer device impractically large. Unfortunately, this device faces a two-sided constraint, and cannot simply be up-sized to that drive strength, since it would impose too much leakage current in the unaccessed rows; additionally, the resulting area increase would offset the density advantage of using a peripheral assist.

Instead, in this design, the nMOS footer is driven with the charge-pump circuit shown in Fig. 8(a). This ensures that its gate drive is at least 600 mV instead of 350 mV, and since the footer is in sub-V_T, its current increases exponentially by over a factor of 500, as shown. As a result, the nMOS footers can be nearly minimum sized, and they consume negligible leakage power in the unaccessed rows. Additionally, because their gate nodes have minimal capacitance, the charge pumps and boost capacitors can be physically small, occupying just slightly more area than a couple of bit-cells. The charge-pump circuit itself is suitable for this ultra-low-voltage application since it uses a pMOS, $M1$, to precharge the boost capacitor and is free from threshold voltage drops. The transient simulation in Fig. 8(b), shows that when a row gets accessed, its *BFB* node gets bootstrapped to nearly $2V_{DD}$, and the following nMOS can easily pull down the feet of the accessed read-buffers.

B. Internal Cell Feedback Control

Write failures occur because, in the presence of variation, we cannot guarantee that the strength of the access devices is more than the strength of the load devices. However, it is possible to enforce the desired relative strengths using circuit assists. For instance, the appropriate bit-line voltage can be pulled below ground or the word-line voltage can be boosted above V_{DD} in order to increase the gate-drive of the nMOS access devices. Unfortunately, both of these approaches require boosting a large capacitance, either the bit-line or word-line, beyond one of the rails. An alternate strategy, that avoids generation of an explicit bias voltage, involves weakening the pMOS load devices by reducing the cell supply voltage. Fig. 9 shows that, as the supply voltage is reduced, the strength required of the access devices is eased, which is reflected by decrease in the minimum word-line voltage that results in a successful write. So, in this design, writeability at 350 mV is ensured by boosting the word-line slightly, by 50 mV, but more importantly by reducing the cell supply voltage to weaken the pMOS load devices.

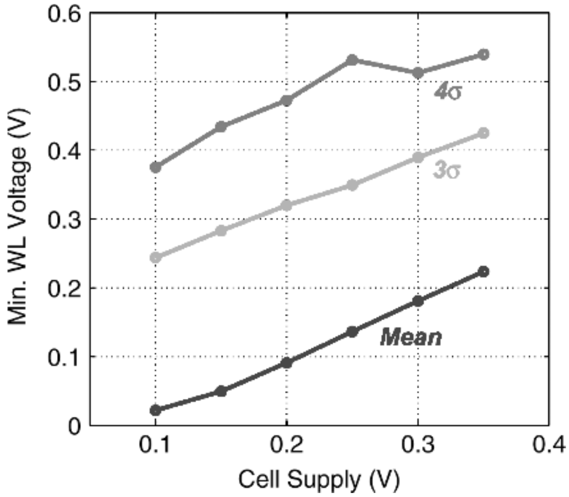


Fig. 9. Minimum word-line voltage resulting in a successful write with respect to the bit-cell supply voltage.

As shown in Fig. 10(a), all cells in each row share a virtual supply node, labeled as VV_{DD}. During the first half of the write cycle, VV_{DD} gets pulled low by the peripheral supply driver. However, as shown in Fig. 10(b), VV_{DD} does not go all the way to ground because all of the accessed cells contribute to pulling it back up. Specifically, one of the bit-lines gets pulled low, causing the corresponding storage node, *QB*, to go low. Accordingly, the alternate pMOS load device tends to turn on, introducing a current path from the opposite storage node to VV_{DD}; in this manner half the bit-cell contributes to pulling VV_{DD} back up through one of its pMOS load devices and one of its nMOS access devices. Fortunately, this interaction is quite accurately controllable, since the pull-down devices of the supply

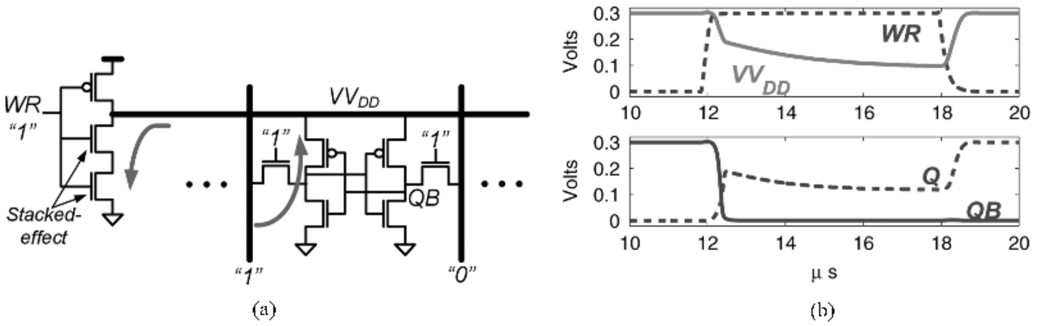


Fig. 10. Virtual V_{DD} scheme. (a) Supporting circuits. (b) Simulation waveforms.

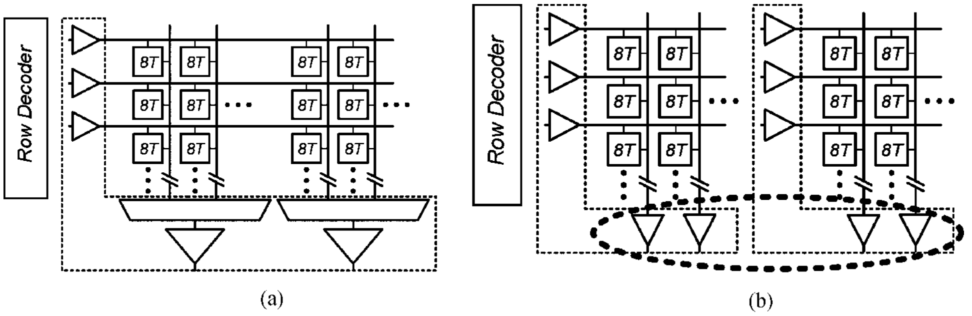


Fig. 11. The number of sense-amps in (a) an interleaved layout is less than the number required in (b) a noninterleaved layout where the accessed cells must be next to each other so that they can share VV_{DD}.

driver are large enough that they experience minimal local variation, and, similarly, the pull-up path through all of the accessed bit-cells tends to average. It is important to note, however, that the supply driver does introduce an additional leakage path in all of the unaccessed rows. To minimize that leakage current, series nMOS pull-down devices are used, taking advantage of the stacked effect [11].

IV. Sense-Amplifier Redundancy

Of course, reducing VV_{DD} as described in Section III-B affects the SNM of all cells in the accessed row, even the ones to which no new data is intended to be written. However, in this design, the supply voltage is already scaled to the hold limit, so further reduction to VV_{DD} will cause loss of data. Consequently, the interleaved layout, shown in Fig. 11(a), where only specific cells in the accessed row get written to through a column MUX, is not suitable, since all cells in that row are susceptible to losing their data. Instead, the layout shown in Fig. 11(b) is used, where all of the cells in each of the sub-rows within the selected row get accessed at the same time, and they can share a VV_{DD} supply. Notice, here, the row decoders and other row periphery can be shared, and only supply drivers and word-line drivers, gated with a sub-row select signal, need to be repeated. Generally, the size of the supply-drivers and word-line drivers scale with their load, reducing the total resulting area overhead.

An important consequence of separating the layout as shown in Fig. 11(b), however, is that adjacent columns can no longer share a sense-amp. As a result, each column has its own sense-amp, making the area of each sense-amp more constrained and increasing the total number in the entire SRAM. This result stresses a general problem observed in deeply scaled technologies. Specifically, the size of the sense-amps is not scaling due to the tradeoff between their statistical offset and their physical size [12]. In this design, that tradeoff is managed in part by using a "full-swing" sensing scheme, where *RDBL* is allowed to discharge completely. Considering the significant speed-up conventionally obtained by small-signal sensing, this might seem like a drastic approach. However, as mentioned in Section III-A the unaccessed read-buffers do impose some minimal droop on *RDBL* due to gate and junction leakage. Oppositely, as the *RDBL* voltage level falls, the unaccessed read-buffers start to drive reverse sub-V_t leakage current from their foot nodes, which are at V_{DD}, to the *RDBL* node. The resulting droop ultimately settles to approximately 120 mV. Unfortunately, as mentioned in Section II-B, I_{READ} variation can cause the read-access time to extend almost arbitrarily, and it can approach the settling time of the transient droop. Consequently, in this design, a static discipline is adopted that guarantees that the correct data value can be sensed on *RDBL* even after the read and droop transients have settled. Specifically, this implies that the offset of all of the sense-amps must be bound by the 120 mV logic "1" level of *RDBL*. To achieve that offset under the imposed area constraints, sense-amp redundancy is employed, as described in the following sections.

A. Sense-Amplifier Offset Sources

Offsets in sense-amps come about as a result of global and local variation in their devices. Global variation refers to die-to-die variation in devices, and local variation refers to mismatch between devices within the same die placed close to each other. Global variation can effect all of the nMOS devices on the chip differently than the pMOS devices, thereby,

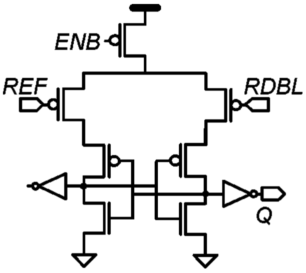


Fig. 12. Differential sense-amp structure cancels effects of global variation.

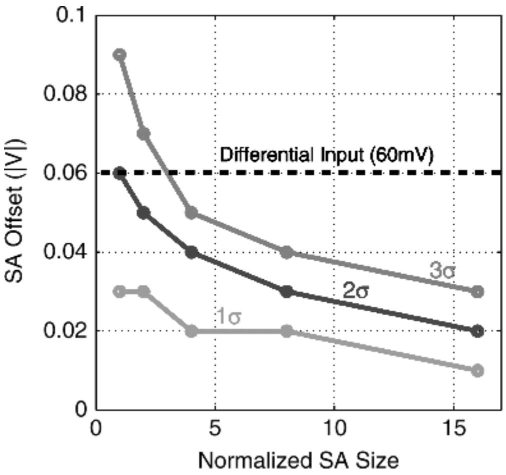


Fig. 13. Monte Carlo simulations of sense-amp statistical offset; at expected input swing (i.e., 60 mV), errors from offset are prominent.

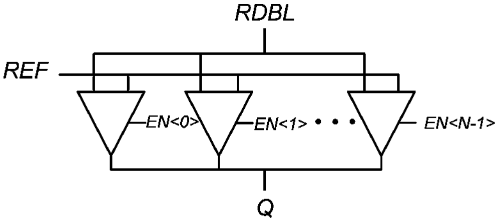


Fig. 14. With sense-amp redundancy, each *RDBL* is connected to N different sense-amps.

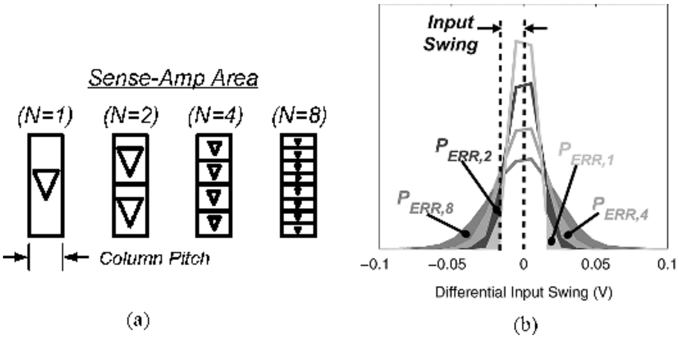


Fig. 15. With sense-amp redundancy, (a) the size of each individual sense-amp must decrease, and (b) individual sense-amp error probabilities, defined as the area under the offset distribution exceeding the magnitude of the input swing, increase.

for instance, skewing the switching threshold of all inverters. Alternatively, local variation can effect the switching threshold differently for each inverter.

Importantly, however, the effect of global variation can be cancelled by using a differential sense-amp, as shown in Fig. 12. The symmetry in this structure ensures that the devices in its two branches will not be subject to systematic differences in processing variation. Of course, the 8T bit-cell of this design uses a single-ended read-buffer and is incompatible with differential sensing. Accordingly, pseudo-differential sensing is used, where *RDBL* drives one of the inputs in Fig. 12, and an off-chip reference drives the other high-impedance input. So, differentially, a 60 mV signal on *RDBL* must be resolved.

The remaining source of offset is local variation, which is modeled as a random effect, whose standard-deviation is inversely related to the square root of the device areas [13], [14]. This gives rise to the area–offset tradeoff that is also shown in the Monte Carlo curves of Fig. 13. In this design, where there are a total of 1024 sense-amps, considerable up-sizing would be required to keep the number of failures from offset to an acceptable limit.

B. Sense-Amplifier Redundancy Concept

As shown in Fig. 14, sense-amplifier redundancy requires that *RDBL* from each column be connected to N different sense-amps. Each of these has the differential structure shown in Fig. 12, so their offsets are from local variation and therefore nondeterministic and uncorrelated. Now, one among them is selected whose offset is bound by the high and low logic levels of *RDBL*. So, if the sense-amp can be selected correctly, only one of the N must have sufficiently low offset. A similar approach has been applied to flash ADCs to achieve minimal offset in the thermometer coded comparators [15].

Importantly, though, the total area for all of the sense-amps is constrained. So, increasing the amount of redundancy means each of them must be smaller. For example, as shown in Fig. 15(a), if N equaled 2, each would need to fit into half the allocated area, and, if N equaled 4, each would need to fit into a quarter of the allocated area. Unfortunately, reducing the size of the individual sense-amps in this manner increases the standard deviation of their offset distribution, and correspondingly increases their probability of error. Specifically, the offset distributions in Fig. 15(b) are derived from Monte Carlo simulations, and the error probability for an individual sense-amp, $P_{\mathrm{ERR},N}$, is defined as the area under its distribution where the magnitude of the offset exceeds the input voltage swing expected on *RDBL*. Here, it is clear that, due to the required reduction in its size, the error probability for an individual sense-amp increases as we increase the amount of redundancy, N. However, the ability to select one structure with sufficiently small offset means that the error probability for the entire sensing network is the joint probability that all of the individual sense-amps yield an error. The total error probability, $P_{\mathrm{ERR,tot}}$, is given by the following:

$$P_{\mathrm{ERR,tot}} = (P_{\mathrm{ERR},N})^N. \tag{1}$$

The resulting error probabilities for the overall sensing networks are plotted in Fig. 16 normalized to error probability of a single, full-sized sense-amp. As shown, increased levels of redundancy

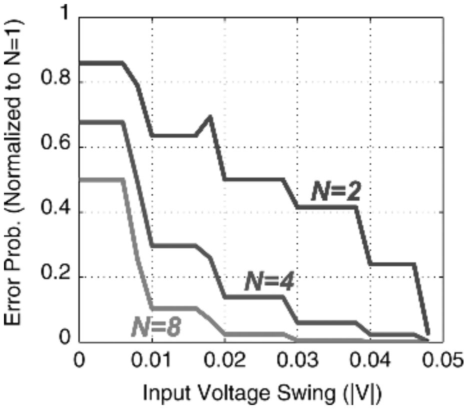


Fig. 16. Increased levels of redundancy significantly reduce the error probability in the overall sensing network.

result in significantly reduced overall error probabilities, and at the input swings expected in this design (i.e., > 50 mV), the resulting improvement is well over an order of magnitude.

C. Sense-Amplifier Redundancy Implementation

The actual implementation of redundancy used in this design incorporates two sense-amps (i.e., $N = 2$). The analysis in Section IV-B considers a general case of up to 8, but at those levels, the total area must be large enough to accommodate at least eight minimum sized structures, and the overhead of the selection logic, which is not considered, becomes significant. With $N = 2$, the selection logic is just two flip-flops and a few logic gates.

The rest of the selection circuitry is shown in Fig. 17. Here, a dummy bit-cell is used with both “0” and “1” data states hard-wired. This cell gets accessed once on power-up, and it enforces the case where *RDBL* is first pulled low, and then where it remains high. Fortunately, the logic “1” and “0” levels of *RDBL* are fairly independent of variation between the accessed bit-cells; as mentioned, logic “1” is set by the aggregate gate, junction, and reverse sub-V_t leakage from all of the read-buffers, and logic “0” is consistently very near ground. Consequently, under a static discipline, the wide distribution in I_{READ} does not limit the integrity with which the dummy cell emulates each data value. Then, the simple state machine in Fig. 17 determines which of the sense-amps can correctly resolve those data values, and only the corresponding structure gets enabled. If both sense-amps work, the first one is selected, and if neither work, the entire SRAM fails.

Fig. 18 shows the normalized overall error probability for the sensing network with the sense-amp sizes actually used in this design. As shown, at the input swings of interest (i.e., ~ 60 mV), the error probability improves by approximately a factor of five compared to a single full-sized sense-amp.

V. Prototype SRAM

A prototype of the SRAM, incorporating the 8T bit-cell, peripheral assists, and sense-amp redundancy, is implemented in a 65 nm CMOS technology. The test-chip consists of eight blocks, each with 256 rows and 128 columns, and has a total capacity of 256 kb. A die photograph of the prototype is shown in Fig. 19. The implemented SRAM achieves full read and write functionality down to 350 mV and retains data down to 300 mV, indicating that the bit-cell and peripheral assists are successful at enabling a V_{MIN} that is close to the retention limit. The following sections describe the characterization results of the prototype with regards to its leakage power, active performance, and active power.

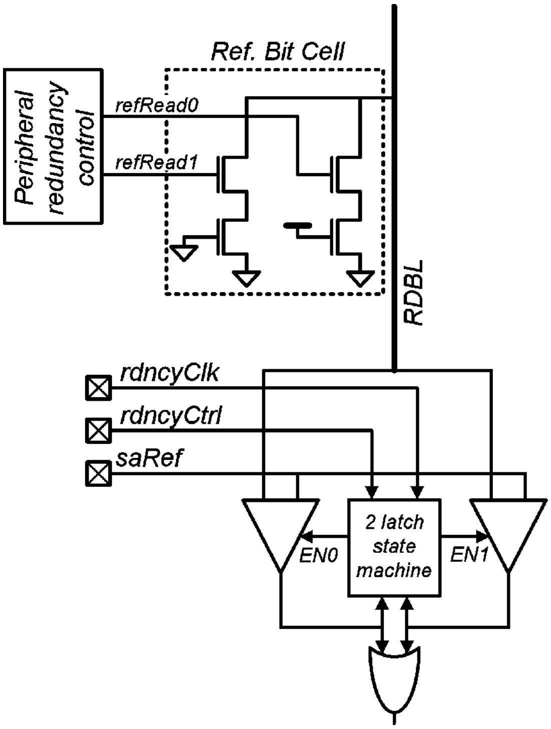


Fig. 17. Redundancy selection circuitry consists of a dummy bit-cell and selection state machine.

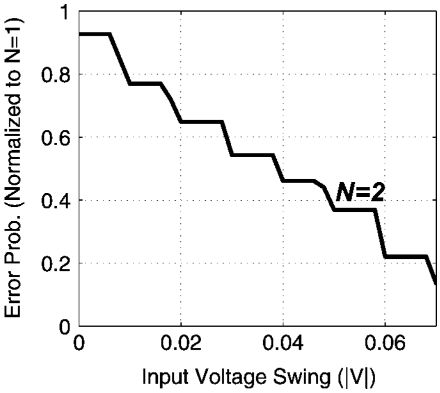


Fig. 18. Overall error probability for implemented sense-amp redundancy scheme improves by a factor of 5 compared to a single sense-amp scheme.

A. Leakage Power

Fig. 20 shows the leakage power of the SRAM with respect to supply voltage for 0 °C, 27 °C, and 75 °C. At the V_{MIN} of 350 mV, the total leakage power is 2.2 μW, representing over a factor of 20 in leakage power savings compared to a supply

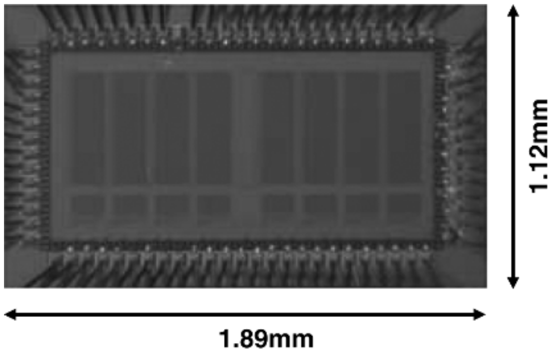


Fig. 19. Die photo of prototype SRAM.

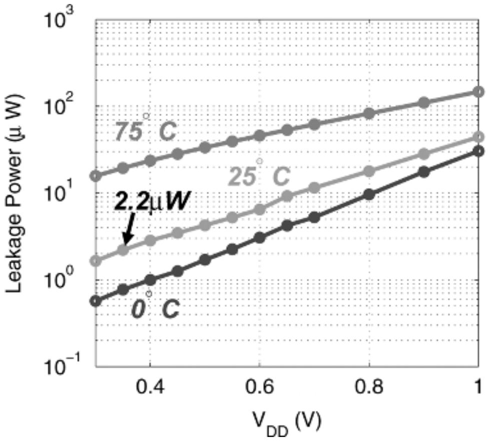


Fig. 20. Prototype SRAM leakage power; at the $V_{\rm MIN}$ of 350 mV, the entire SRAM draws 2.2 μW of leakage power.

voltage of 1 V. As mentioned, the SRAM also retains data down to 300 mV where the total leakage power is 1.65 μW.

The area and leakage power of this SRAM can be compared to a conventional 6T design, and the 10T sub-V_t design in [3]. From the actual cell layouts, this design represents an area overhead of approximately 30% compared to a 6T design and an area savings of approximately 30% compared to the 10T design. Additionally, the leakage power savings of this design, compared to a conventional 6T design, with a projected $V_{\rm MIN}$ of approximately 700 mV [16], [17], is over 5x.

B. Active Performance

Fig. 21 shows the active read and write performance of the prototype SRAM. with respect to the supply voltage. As expected, the speed is significantly reduced in sub-V_t, and at 350 mV, the SRAM operates at 25 kHz.

C. Active Power

Fig. 22 shows the total (i.e., active plus leakage) power, in the solid curves, and just the leakage power, in the dotted curves, with respect to the operating frequency. The leakage power remains a dominant portion of the total power for a very wide range of frequency, so leakage minimization efforts are well justified.

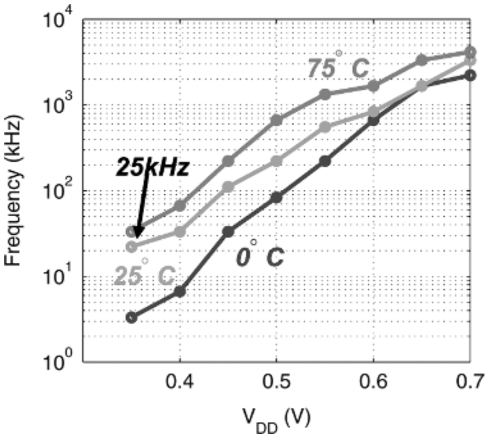


Fig. 21. SRAM speed with respect to V_{DD}.

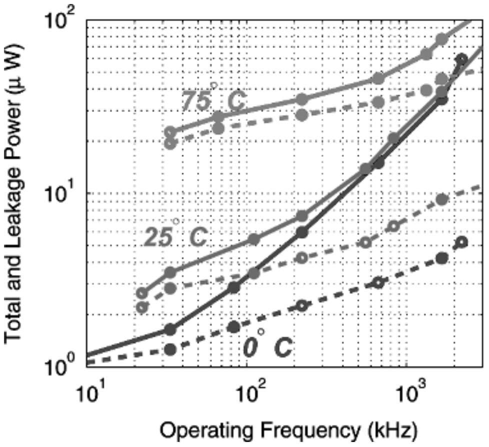


Fig. 22. Total power (solid curves) and leakage power (dotted curves) with respect to operating frequency.

VI. Conclusions and Summary

Voltage scaling is an effective strategy for minimizing the power consumption of SRAMs. Further, as SRAMs continue to occupy a dominating portion of the total area and power in modern ICs, the resulting total power savings are significant. Unfortunately, however, conventional SRAMs, based on the 6T bit-cell, fail to operate at voltages below approximately 700 mV both because of reduced signal levels and because of increased variation. In sub-V_t, in particular, threshold voltage variation has an exponential effect on the drive current, resulting in increased cell instability and a severely degraded read-current. To address these limitations, an 8T bit-cell is incorporated into a 65 nm 256 kb SRAM, and it achieves full read and write functionality deep into the sub-V_t regime at 350 mV. At this voltage, the total leakage power is 2.2 μW, and the operating speed is 25 kHz. The significantly reduced speed is expected in sub-V_t and is acceptable for low throughput, energy-constrained applications. At 350 mV, the leakage power represents almost 85% of the total power consumption, so, leakage reduction is a critical consideration. Additionally, the tradeoff between the size of a sense-amp and its statistical offset is emerging as a primary limitation to SRAM scaling in advanced technologies. In this design, enabling sub-V_t write requires the use of circuit assists that result in a layout where sense-amp multiplexing between adjacent columns is impractical. Accordingly, the sense-amp scaling limitation is stressed, necessitating a different approach to managing the offset–area tradeoff. The concept of sense-amp redundancy is introduced, and it is demonstrated that, for a given area constraint, errors in the sensing network due to offsets can be reduced by over an order of magnitude. In this design, a factor of five improvement is expected with the implemented scheme, which incorporates a simple start-up control loop.

ACKNOWLEDGMENT

The authors thank Prof. B. H. Calhoun and M. Bhardwaj for their support and feedback.

REFERENCES

[1] A. Wang, A. Chandrakasan, and S. Kosonocky, “Optimal suppply and threshold scaling for subthreshold CMOS circuits,” in *Proc. IEEE Comput. Soc. Annu. Int. Symp. VLSI*, Apr. 2002, pp. 5–9.

[2] N. Verma and A. Chandrakasan, “A 65 nm 8T sub-V_t SRAM employing sense-amplifier redundancy,” in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 328–329.

[3] B. Calhoun and A. Chandrakasan, “A 256 kb subthreshold SRAM in 65 nm CMOS,” in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2006, pp. 480–481.

[4] J. Chen, L. Clark, and T.-H. Chen, “An ultra-low-power memory with a subthreshold power supply voltage,” *IEEE J. Solid-State Circuits*, vol. 41, no. 10, pp. 2344–2353, Oct. 2006.

[5] A. Wang and A. Chandrakasan, “A 180 mV FFT processor using subthreshold circuit techniques,” in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2004, pp. 292–293.

[6] T. Mizumo, J.-I. Okamura, and A. Toriumi, “Experimental study of threshold voltage fluctuations using an 8 k MOSFET’s array,” in *Proc. IEEE Symp. VLSI Technology*, May 1993, pp. 41–42.

[7] S.-W. Sun and P. G. Y. Tsui, “Limitations of CMOS supply-voltage scaling by MOSFET threshold-voltage variation,” *IEEE J. Solid-State Circuits*, vol. 30, no. 8, pp. 947–949, Aug. 1995.

[8] A. Bhavnagarwala, X. Tang, and J. Meindl, “The impact of intrinsic device fluctuations on CMOS SRAM cell stability,” *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, Apr. 2001.

[9] E. Seevinck, F. J. 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.

[10] L. Chang, D. M. Fried, J. Hergenrother, J. W. Sleight, R. H. Dennard, R. Montoye, L. Sekaric, S. J. McNab, A. W. Topol, C. D. Adams, K. W. Guarini, and W. Haensch, “Stable SRAM cell design for the 32 nm node and beyond,” in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2005, pp. 128–129.

[11] Y. Ye, S. Borkar, and V. De, “A new technique for standby leakage reduction in high performance circuits,” in *Proc. IEEE Symp. VLSI Circuits*, Jun. 1998, pp. 40–41.

[12] K. Zhang, K. Hose, V. De, and B. Senyk, “The scaling of data sensing schemes for high speed cache design in sub-0.18 μm technologies,” in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2000, pp. 226–227.

[13] M. Pelgrom, H. Tuinhout, and M. Vertregt, “Transistor matching in analog CMOS applications,” in *IEDM Dig. Tech. Papers*, Dec. 1998, pp. 915–918.

[14] P. Kinget, “Device mismatch and tradeoffs in the design of analog circuits,” *IEEE J. Solid-State Circuits*, vol. 40, no. 6, pp. 1212–1224, Jun. 2005.

[15] M. P. Flynn, C. Donovan, and L. Sattler, “Digital calibration incorporating redundancy of flash ADCs,” *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 50, no. 3, pp. 205–213, May 2003.

[16] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, “A SRAM design on 65 nm CMOS technology with integrated leakage reduction scheme,” in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2004, pp. 294–295.

[17] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. S. K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, “Low power embedded SRAM modules with expanded margins for writing,” in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2005, pp. 480–481.

Naveen Verma (S’04) received the B.A.Sc. degree in electrical and computer engineering from the University of British Columbia, Vancouver, BC, Canada, in 2003, and the M.S. degree from the Massachusetts Institute of Technology, Cambridge, MA, in 2005. He is currently pursuing the Ph.D. degree at the Massachusetts Institute of Technology.

His research interests include low-power mixed signal circuits in the areas of analog-to-digital converters, SRAMs, and implantable biological systems.

Mr. Verma was the recipient of the Intel Foundation Ph.D. fellowship and the NSERC Postgraduate fellowship.

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 Academic, 1995), *Digital Integrated Circuits* (Pearson Prentice-Hall, 2003, 2nd edition), and *Subthreshold Design for Ultra-Low Power Systems* (Springer 2006). He is also a co-editor of *Low Power CMOS Design* (IEEE Press, 1998), *Design of High-Performance Microprocessor Circuits* (IEEE Press, 2000), and *Leakage in Nanometer CMOS Technologies* (Springer, 2005).

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 2004–2007. 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.