Analysis Towards Minimization of Total SRAM Energy Over Active and Idle Operating Modes

Naveen Verma, *Member, IEEE*

Abstract—**Computational requirements in highly energy constrained applications are driving the need for ultra-low-power processors. In such devices SRAMs pose a primary energy limitation. This paper analyzes SRAM energy in practical applications using state-of-the-art power-management techniques. The design targets and array biasing for energy minimization are developed. Compared with generic logic, these are characterized by the important difference that SRAMs generally need to retain data. This restricts the use of power-gating for leakage elimination, and thus this paper considers the application of low-leakage data-retention biasing during the idle-mode. The resulting energy tradeoffs have important distinctions, and these are analyzed in the presence of practical variation levels.**

Index Terms—**CMOS memory circuits, data-retention voltage, energy minimization, power-aware computing, SRAM.**

I. Introduction

Energy-efficiency is a paramount concern in modern processors. Emerging architectures favor the integration of increasing amounts of memory on-chip due to performance benefits that come at relatively modest energy cost. SRAMs, in particular, play a critical role due to their achievable access speeds, their compatibility with standard processes, and the exponential density-scaling trend that their designers have been able to successfully enforce. Compared with generic logic, however, SRAMs are severely constrained by the simultaneous need for very high density, low leakage, high performance, and long-term data retention. Although a diverse range of power-management approaches have been developed to address these, the prominent role played by SRAMs makes their power-consumption a key concern in severely energy constrained applications such as wireless sensor networks, mobile multimedia, and implantable/wearable biomedical devices. Fig. 1 illustrates three devices with increasingly severe energy constraints, and these culminate in the case of a 0.3 V ULP DSP [1] where the embedded SRAM consumes 69% of the total processor power.

Due to the importance of leakage-energy in SRAMs, one of their notable features is the general need for long-term data retention. This restricts the application of power-gating, which provides effective leakage mitigation in generic logic. As a result, SRAM energy characteristics differ from those of generic logic. This paper analyzes SRAM energy incorporating the alternate power-management techniques that may be applied. For this, a practical array is modeled. The model applies generically to arrays of any size and any bit-cell topology. For analysis, however, a specific array is considered, corresponding to 256 × 256 0.25 μm^2 6 T bit-cells in a commercial 45 nm LP CMOS process. This is used to establish practical targets for the most critical optimization parameters (namely V_{DD} and V_t) with respect to application-dependant performance constraints. Following this, the effect of device variation is incorporated so that its impact on the original analysis and tradeoffs can be understood.

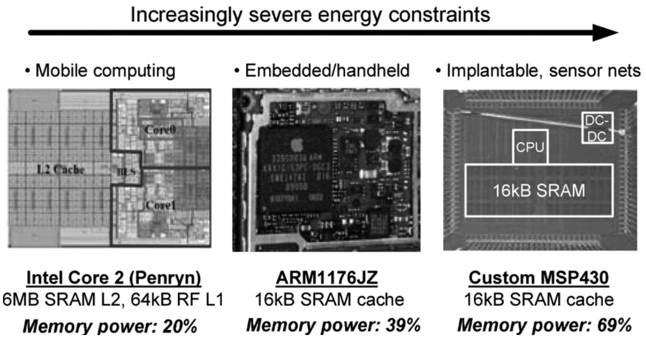


Fig. 1. Three low-power applications (45 nm Intel Core 2 [2], 90 nm ARM1176JZ, and 65 nm 0.3 V MSP430 [1]) culminating in SRAM consuming 69% of the chip power.

II. SRAM Operation and Power-Management

In order to integrate a substantial amount of SRAM on-chip, bit-cell are typically grouped into sub-arrays (usually up to 256×256 bits), as shown in Fig. 2. The need for separate peripheral control circuitry for each sub-array degrades the area-efficiency somewhat, but, the performance and energy are improved thanks to the reduced bit-line and word-line capacitances *and* the ability to apply power-management assists at the individual sub-array level. As the number of sub-arrays becomes very large, the active energy of selecting and multiplexing data to the output of the bank can be significant. However, the associated circuitry does not face the same density, stability, and leakage power constraints as bit-cells within the sub-array. Thus, they can employ more aggressive energy reduction techniques and more relaxed device sizing choices.

Energy minimization has been studied for generic logic circuits [3], [4]. SRAMs, however, require some distinct considerations. First, the structure of the array elevates the importance of leakage energy versus active energy. Active-switching is mitigated since only one word-line must be asserted and since the bit-lines can exploit low-swing reads and column-multiplexed

Manuscript received February 02, 2010; revised April 18, 2010; accepted June 08, 2010.

The author is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: nverma@princeton.edu).

Digital Object Identifier 10.1109/TVLSI.2010.2055906

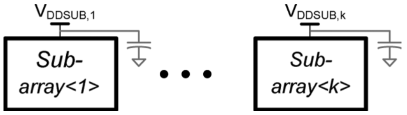


Fig. 2. SRAM structure composed of multiple sub-arrays.

writes. Leakage, on the other hand, is aggravated since it integrates over a read delay that is limited by one tiny bit-cell. This path faces high variability and does not benefit from the delay averaging experienced in multistage timing paths [5].

Second, the need for long-term data retention limits the use of high-V_t power switches. These can virtually eliminate leakage-currents in combinational logic when performance reduction beyond the minimum energy point is tolerable [4]. In SRAMs, however, data must generally be retained in the array. Thus, while power-gating may be applied at the end of some system-level operating phase, the period of this phase is generally not correlated with just the circuit's access-speed.

Fortunately, SRAMs can exploit a low-leakage idle-mode analogous to the (nearly) zero-leakage idle-mode of generic logic. Instead of gating the supply-voltage, the SRAM voltage may be reduced to the data-retention voltage (V_{DRV}). This is the minimum voltage where the bit-cell's hold-margin is preserved, but its leakage current is *reduced* as much as possible [6]. Accordingly, after completing a set of array-accesses (N_{ACC}) associated with a target operation, a sub-array can switch into idle-mode, where the noise-margins are relaxed since no accesses are performed and only data-retention is required. The functionality constraints and energies are summarized in Fig. 3. During the active-mode, the supply-voltage must be large enough to meet the SRAM read/write margins. Both active-switching and static-leakage energy are consumed during this period, T_{ACC}. T_{ACC} may be optimized freely as long as it meets the application's throughput constraint specified by the retention-cycle time, $T_{\mathrm{CYC,RTN}}$. During $T_{\mathrm{CYC,RTN}}$ the set of SRAM accesses required must be completed (i.e., $T_{\mathrm{ACC}} < T_{\mathrm{CYC,RTN}}$), and the corresponding data must be retained in the array. If T_{ACC} is less than $T_{\mathrm{CYC,RTN}}$, the SRAM may enter the idle-mode for the remainder of the retention-cycle where static-leakage energy is consumed, but at a much lower rate than the active-mode. Additionally, however, some overhead energy must be incurred in order to switch the total capacitance that is coupled to the supply-voltage. This can be achieved with simple (passive) circuits [7], but the ensuing transition time affects the idle mode energy (recovery back to the active mode can be achieved quickly (i.e., within a clock cycle) by way of an array power supply switch [8]). It is worth noting that the idle mode is useful not only for managing the leakage-power of the sub-array under consideration, but also the leakage-power of all inactive sub-arrays shown in Fig. 2, since as few as one sub-array may enter the active-mode during a retention-cycle.

A. Energy Components

In this section, the energy components from Fig. 3 will be modeled. The attempt is to use generic parameters so that the methodology can apply to a broad range of bit-cells and array configurations. Ultimately, the energy components will be combined to analyze the total sub-array energy during a retention-cycle. The parameters related to the sub-array that are used in the energy model are summarized in Fig. 4.

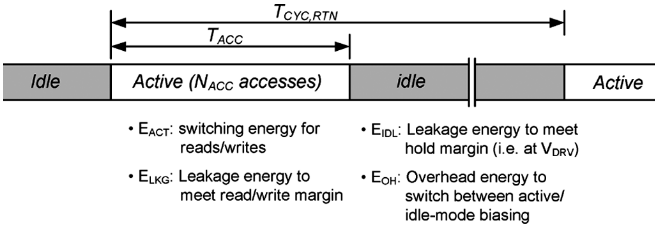

T_CYC,RTN
T_ACC
Idle
Active (N_ACC accesses)
idle
Active
• E_ACT: switching energy for reads/writes
• E_LKG: Leakage energy to meet read/write margin
• E_IDL: Leakage energy to meet hold margin (i.e. at V_DRV)
• E_OH: Overhead energy to switch between active/idle-mode biasing

Fig. 3. Summary of energy components contributing to total SRAM energy.

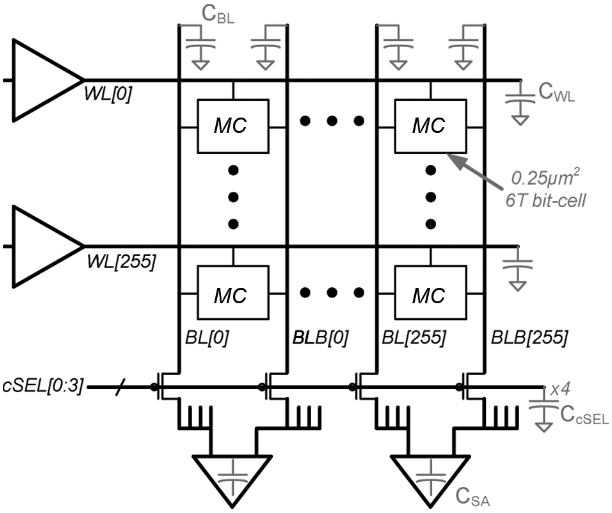


Fig. 4. Sub-array highlighting parameters critical for determining energy.

1) Active Energy (E_{ACT})*:* The active energy associated with capacitive switching during sub-array reads is given by

$$E_{\mathrm{ACT,RD}} = N_{\mathrm{ACC}} \Big[C_{WL} V_{DD}^2 + C_{\mathrm{cSEL}} V_{DD}^2 + \frac{i}{m} C_{SA} V_{DD}^2 + i C_{BL} V_{DD} V_{\mathrm{SNS}} \Big]. \quad (1)$$

Here, N_{ACC} is the total number of array accesses. During each of these, one word-line capacitance (C_{WL}) and one column-select-line capacitance (C_{cSEL}) is assumed to switch. Additionally, there is capacitive switching associated with the sense-amplifiers (C_{SA}), of which there are i/m (where i is the number of columns and m is the column-multiplexing ratio). Finally, the i bit-line capacitances (C_{BL}) switch, but only by an amount required by the small-signal sense-amplifier (V_{SNS}).

Similarly, the active energy during sub-array writes is given by (2)

$$E_{\mathrm{ACC},WR} = N_{\mathrm{ACC}} \Big[C_{WL} V_{DD}^2 + C_{\mathrm{cSEL}} V_{DD}^2 + \frac{i}{m} C_{BL} V_{DD}^2 + i \frac{m-1}{m} C_{BL} V_{DD} V_{\mathrm{SNS}} \Big]. \quad (2)$$

Here, the selected bit-lines are switched completely while the half-selected bit-lines droop (to V_{SNS}) according the state of the bit-cell.

2) Leakage Energy (E_{LKG})*:* The leakage energy of the sub-array during the active mode is given by

$$\begin{aligned} E_{\mathrm{LKG}} &= ij \int\limits_{T_{\mathrm{ACC}}} I_{\mathrm{LKG},BC} V_{DD} dt \\ &= ij I_{\mathrm{LKG},BC} V_{DD} T_{\mathrm{ACC}}. \end{aligned} \tag{3}$$

Here, the leakage current of a bit-cell is represented by $I_{\mathrm{LKG},BC}$, and it depends strongly on V_{DD} (through DIBL) and on V_t. The bit-cell leakage power is multiplied by the size of the array (i-columns $\times j$-rows) and the length of the access-period (T_{ACC}) to represent the energy. Since both $I_{\mathrm{LKG},BC}$ and T_{ACC} also depend on temperature, characterization for absolute energy requires consideration over the relevant temperature range.

3) Idle-Mode Energy (E_{IDL})*:* The idle-mode energy of the sub-array is given by (4), shown at the bottom of the page. Here, T_{Droop} corresponds to the time it takes the supply voltage to droop to the data-retention voltage. If the supply does not reach the data-retention voltage, no idle-mode energy is consumed (only overhead energy is consumed to recover the partial supply voltage droop). If the supply does reach the data-retention voltage, it is assumed to remain there, and the corresponding leakage current of a bit-cell is represented by $I_{\mathrm{DRV},BC}$. Although either the array supply-voltage or ground may be manipulated to reduce the leakage in the idle-mode, for this analysis, supply-voltage reduction will be considered. As with $I_{\mathrm{LKG},BC}$, $I_{\mathrm{DRV},BC}$ depends on the temperature, and hence characterization for absolute energy requires consideration over the relevant temperature range.

4) Overhead Energy (E_{OH})*:* The overhead energy that must be incurred in order to recover the sub-array supply-voltage is given by (5), shown at the bottom of the page, (V_{Droop} is the voltage to which the supply droops, but it cannot be less that the data-retention voltage, V_{DRV}). Here, the total capacitance that is coupled to the supply voltage is represented by C_{VDD}, and the overhead of the control circuitry required to enforce the idle-mode biasing is represented by E_{CNTRL}. Assuming, passive voltage regulation, E_{CNTRL}, is dominated by the wire capacitance to drive the supply switches: $C_{\mathrm{CNTRL}} V_{DD}^2$ (approaches to minimize this overhead have been proposed [9]). It should be noted that E_{OH} implies that the idle-mode is only viable if the resulting energy savings exceed the overhead. Specifically

$$E_{OH} < ij I_{\mathrm{LKG},BC} V_{DD} (T_{\mathrm{CYC,RTN}} - T_{\mathrm{ACC}}) - E_{\mathrm{IDL}}. \tag{6}$$

B. Importance of V_{DD} and V_t

The supply voltage and threshold voltage are critical parameters in determining the total energy of the sub-array. Supply-voltage directly affects the active energy in (1) and (2), and it also affects the leakage energy in (3), both directly and through an impact on performance (T_{ACC}).

The threshold voltage is critical in determining all of the leakage currents (i.e., $I_{\mathrm{LKG},BC}$, $I_{\mathrm{DRV},BC}$) as well the array performance (T_{ACC}). Additionally, as discussed in Section IV it has an important impact on the level of variation expected in the bit-cell devices. The importance of threshold voltage comes about due to the prominence of leakage energy in SRAMs. Aside from the fact that SRAM sub-arrays consist of a *high number of leakage paths*, the use of intentionally *tiny* devices implies extreme V_t variation.

III. SRAM Energy Analysis

For analysis, the sub-array parameters summarized in Table I are used. A practical SRAM based on a 45 nm LP process and 0.25 μm^2 6 T bit-cells is simulated with parasitic capacitances (for C_{WL}, C_{BL}, $C_{c\mathrm{SEL}}$, C_{SA}, C_{VDD}, and C_{CNTRL}) included from layout extraction. Currently, the effect of variation is omitted (and will be included in Section IV). The minimum achievable V_{DRV} strongly depends on how variability affects the bit-cell hold-margin, but for the initial analysis a V_{DRV} of 0.4 V is supposed.

The parameters N_{ACC} and $T_{\mathrm{CYC,RTN}}$ are determined by the application. $T_{\mathrm{CYC,RTN}}$ has important implications for determining the *relative* importance of each energy component, while N_{ACC} causes most components (except E_{OH}) to scale similarly. For the analysis, $N_{\mathrm{ACC}} = 1024$ is assumed, corresponding to one access of every bit-cell in the 256×256 sub-array (since a column multiplexing ratio of 4 implies that 64 bits are accessed every cycle). To elucidate the energy

$$\begin{aligned} E_{\mathrm{IDL}} &= \begin{cases} ij \int_{T_{\mathrm{CYC,RTN}} - T_{\mathrm{ACC}} - T_{\mathrm{Droop}}} I_{\mathrm{DRV},BC} V_{\mathrm{DRV}} dt, & (\text{for } T_{\mathrm{CYC,RTN}} > T_{\mathrm{ACC}} + T_{\mathrm{Droop}}) \\ 0, & (\text{otherwise}) \end{cases} \\ &= \begin{cases} ij I_{\mathrm{DRV},BC} V_{\mathrm{DRV}} (T_{\mathrm{CYC,RTN}} - T_{\mathrm{ACC}} - T_{\mathrm{Droop}}), & (\text{for } T_{\mathrm{CYC,RTN}} > T_{\mathrm{ACC}} + T_{\mathrm{Droop}}) \\ 0, & (\text{otherwise}) \end{cases} \end{aligned} \tag{4}$$

$$E_{OH} = \begin{cases} C_{\mathrm{VDD}} V_{DD} (V_{DD} - V_{\mathrm{DRV}}) + E_{\mathrm{CNTRL}}, & (\text{for } T_{\mathrm{CYC,RTN}} > T_{\mathrm{ACC}} + T_{\mathrm{Droop}}) \\ C_{\mathrm{VDD}} V_{DD} (V_{DD} - V_{\mathrm{Droop}}) + E_{\mathrm{CNTRL}}, & (\text{otherwise}) \end{cases} \tag{5}$$

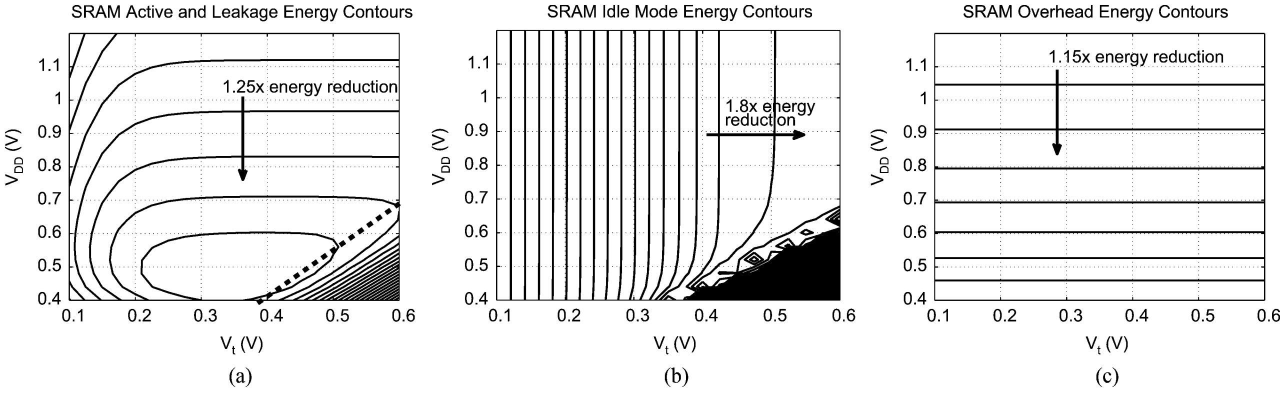


Fig. 5. Log-magnitude of sub-array energy components versus V_{DD} and V_t (for $T_{\rm CYC,RTN} = 1$ ms; with $N_{\rm ACC} = 1024$, this corresponds to an access period of over 1 MHz). (a) Active-mode energy components $(E_{\rm ACT} + E_{\rm LKG})$. (b) Idle-mode leakage energy component $(E_{\rm IDL})$. (c) Overhead energy component (E_{OH}).

TABLE I
SUB-ARRAY PARAMETER SUMMARY FOR ENERGY ANALYSIS

Sub-array Parameters	
Technology	45nm LP CMOS
Bit-cell	0.25 μm^2, 6T
Array configuration	256 cells x 256 cells
Column-muxing	4:1 (i.e., 64 sense-amplifiers)
BL swing (V_{SNS})	0.2V
N_{ACC}	1024
$T_{CYC,RTN}$	10ms, 1ms, 100μs, 10μs
Access throughput (N_{ACC}/ $T_{CYC,RTN}$)	100kHz, 1MHz, 10MHz, 100MHz

trends, various values of $T_{\rm CYC,RTN}$ are considered that are relevant for the energy constrained applications mentioned in Section I. It should be noted that the choice of array configuration has some impact on the relative magnitudes of the energy components. Namely, the word-line and column-select-line components of the active energy do not scale with the number of rows (j), and, with fewer rows, the access delay $(T_{\rm ACC})$ can also be reduced, altering the leakage and idle-mode energies. The analysis methodology may be applied to various sub-array configurations. As mentioned in Section II-A, some of the energy components depend on temperature. Simulations at various temperatures show that the absolute energy values vary accordingly. However, the key trends presented below are persistent (they have been derived for a temperature of 25 °C).

A. SRAM Energy Component Scaling

In this section the energy components from Section II-A are plotted to isolate their characteristics. Fig. 5(a) shows the active mode energy (i.e., $E_{\rm ACT} + E_{\rm LKG}$) as log-magnitude contours. This plot removes the effect of idle-mode current, and therefore behaves similar to that of generic digital logic [4]. Namely, a minimum energy point is formed based on the active-mode switching and leakage energies. At high V_{DD}, energy is dominated by switching and scales at an approximate rate of CV_{DD}^2 until leakage-energy becomes significant due to rapid performance $(T_{\rm ACC})$ degradation. V_t scaling effects leakage-current and performance; as operation approaches the sub-V_t regime, the two effects begin to negate each other, diminishing the impact of V_t scaling. In deep sub-V_t (i.e., lower-right region), energy begins rising due to increased static-current from degraded logic-levels.

Fig. 5(b) shows the idle-mode energy $(E_{\rm IDL})$ plotted as log-magnitude contours. Since an array supply voltage of $V_{\rm DRV}$ is assumed, the dependence on V_{DD} is minimal (until $T_{\rm ACC}$ begins to approach $T_{\rm CYC,RTN}$- this case is discussed below). At low V_t, the equally spaced vertical contours indicate exponential leakage-current reduction (by a constant factor of 1.8). At high V_t and high V_{DD}, the separation between contours begins increasing since leakage sources not strongly dependant on V_t (i.e., gate and junction leakage) begin dominating. At high V_t and low V_{DD}, the contours begin tapering together, indicating rapid reduction in $E_{\rm IDL}$; this is brought-on by rapid performance degradation during the active-mode. As a result, $T_{\rm ACC}$ approaches $T_{\rm CYC,RTN}$, leaving reduced time for the idle-mode. The blocked region (lower-right) corresponds to an invalid V_{DD} and V_t regime where the performance degrades beyond the application constraint $(T_{\rm CYC,RTN})$. Finally, it is important to note that as system-constraints call for increased $T_{\rm CYC,RTN}$, the importance of idle-mode energy increases while that of other components remains relatively constant; as a result, $E_{\rm IDL}$ starts to have a dominating effect.

Fig. 5(c) shows the overhead energy (E_{OH}) plotted as log-magnitude contours. The overhead energy scales predictably as the supply voltage switches between V_{DD} and $V_{\rm DRV}$. In the current analysis, E_{OH} is independent of V_t. Later, in the presence of variation, the minimum achievable $V_{\rm DRV}$ will depend on V_t, affecting E_{OH} accordingly.

B. SRAM Energy Scaling

Combining all of the energy components, the log-magnitude contours for the total energy are shown in Fig. 6. Four different system performance constraints $(T_{\rm CYC,RTN})$ are considered: 1) 10 ms; 2) 1 ms; 3) 100 μs; and 4) 10 μs (with $N_{\rm ACC} = 1024$, these correspond to access periods of slightly over 100 kHz, 1, 10, and 100 MHz, respectively, which are relevant for many energy constrained applications). The blocked region once again indicates the V_{DD} and V_t regime where the system performance constraint cannot be met. It should be noted that, although a wide V_{DD} and V_t range is plotted, functionality constraints (described in Section IV-B) limit the actual range achievable. For the current analysis a data-retention voltage of 0.4 V is assumed; thus, this is plotted as the limit for V_{DD} scaling.

Comparing the plots in Fig. 6, it can be seen that increasing $T_{\rm CYC,RTN}$ (i.e., reducing the system performance requirement) increases the impact of the idle-mode energy, causing the contours from Fig. 5(b) to emerge more prominently. For instance, in Fig. 6(a), threshold voltage increase rapidly improves the total energy. On the other hand, with a high system performance requirement [e.g., Fig. 6(d)], the active energy scaling trend from Fig. 5(a) has increased prominence.

Across all performance cases, an important observation is that *both* the idle-mode and active-mode play significant roles in determining the energy trends [in contrast to the energy trends for generic logic, which are similar to Fig. 5(a)]. For instance, while the active-mode has an energy minimum [i.e., at $V_{DD} = 0.5$ V and $V_t = 0.35$ V in Fig. 5(a)], the total energies continue to decrease with V_{DD} until the performance requirements cannot be met. Intuitively, the increased emphasis on voltage scaling is driven by the fact that the sub-array must retain data (and incur leakage energy) even after the $N_{\rm ACC} = 1024$ accesses are complete. Hence, maintaining performance *at the cost of V_{DD} scaling* has diminished appeal.

The contours of Fig. 6 advocate V_{DD} reduction and V_t elevation until the performance requirement can no longer be met. This implies that the entire retention period $(T_{\rm CYC,RTN})$ should be spent in the active-mode. Of course, the idle-mode retains its importance for mitigating the leakage of inactive sub-arrays in the SRAM configuration of Fig. 2. In practice, the idle-mode helps towards minimizing leakage in the active sub-array as well. In order to ensure robust operation, the active-mode V_{DD} must be set to accommodate the most stringent system performance requirement. In the absence of fine-grained operation-by-operation dynamic voltage scaling (DVS), the required sub-array performance may exceed the instantaneous system requirement, making the idle-mode valuable during the remaining retention period. In fact, even if the number of accesses and the system performance requirements remain constant, some V_{DD} margin is necessary to accommodate process and operating condition variations. Under conditions where the entire margin is not utilized, its overhead can be reduced considerably by evoking the idle-mode. Fig. 7 shows the energy cost of such V_{DD} margin if provisions for the idle-mode are not included. Specifically, the energy components for the $T_{\rm CYC,RTN} = 10$ ms case are normalized to $E_{\rm TOT}$ and shown along a vertical slice corresponding to $V_t = 0.45$ V. Along this slice, a minimum V_{DD} of 0.45 V just meets the performance requirement; this reduces $E_{\rm IDL}$ to zero since the entire $T_{\rm CYC,RTN}$ is spent in the active-mode. An additional energy component, $E_{\rm LKG,CYC,RTN}$, is also plotted. This represents the leakage-energy that would be incurred if V_{DD} is increased and no idle-mode is available. By normalizing $E_{\rm LKG,CYC,RTN}$ to the total original energy $(E_{\rm TOT})$, it can be seen that the leakage-overhead of introducing V_{DD} margin results in a dominating rise in leakage-energy.

Overall, the results of Fig. 6 convey two important messages: 1) SRAM energy reduction benefits from more aggressive V_{DD} reduction and V_t elevation than non-state-retaining logic and 2) SRAM performance enhancement can significantly help reduce energy by enabling further V_{DD} and V_t scaling. The former message can be seen by comparing the contours for $E_{\rm TOT}$ (see Fig. 6) with those for $E_{\rm ACT}$ [see Fig. 5(a)], which follow the behavior of generic logic. The latter message can be seen by observing that the contours in Fig. 6 indicate significantly decreasing energy even into the blocked regime (where the system performance requirement is not met). Both of these design objectives, however, face oppositions in the presence of variation. The next section, analyzes this and revises the energy analysis.

IV. Impact of Variation on SRAM Energy

The need for large on-chip SRAMs implies that variation at the 4-5σ level must be considered, and the use of intentionally tiny devices (in order to maximize cell density) aggravates the standard-deviations exhibited by critical device parameters. This section starts by investigating the device-level implications with regards to variation as the V_{DD} and V_t targets from the previous section are pursued. Following this, the impact on bit-cell metrics is considered, and then, finally, the impact on sub-array energy is derived.

A. Device-Level Impact of Variation

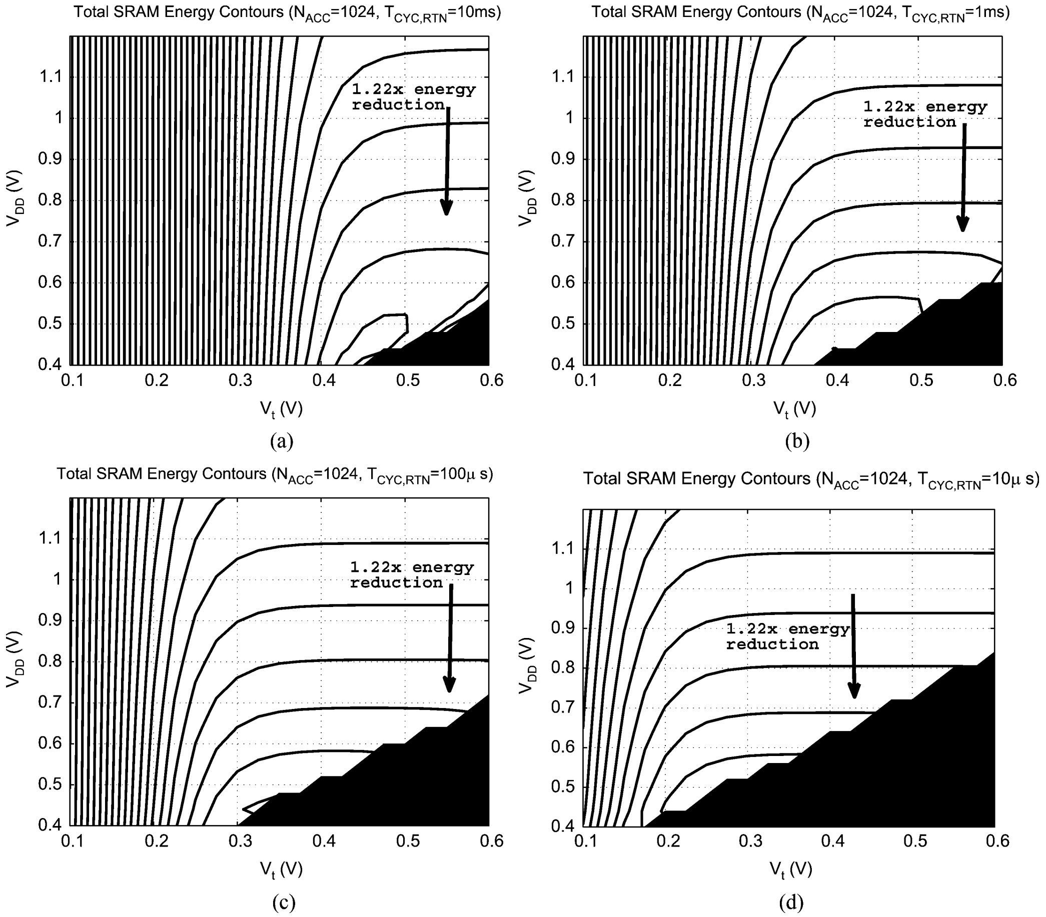
In low-V_{DD} and high-V_t designs, threshold voltage variation is a dominating limitation. Its effect with V_{DD} scaling is considered in Fig. 8(a), which shows the normalized on-current of an nMOS (whose V_t is constant at 0.3 V). As V_{DD} is reduced, the mean current degrades gradually at first and then rapidly as the subthreshold regime is approached. Importantly, however, the deviation between the mean current and the 4σ current widens. As V_{DD} is reduced, threshold voltage variation accounts for a larger proportion of the gate-overdrive, and it eventually leads to exponential degradation in subthreshold, as expected. The resulting impact on bit-cells (which is detailed in Section IV-B) is two-fold: as V_{DD} is reduced: 1) threshold-voltage variation accounts for a larger proportion of the gate-overdrive and 2) the voltage-noise margins required for functional operation are degraded.

Like V_{DD}, V_t scaling according to the targets established in Section III-B is strongly opposed by device variation. As devices are engineered for higher V_t, the standard deviation of their V_t increases. Analytically, this can be seen through the random-dopant fluctuation relationship [10]

$$\begin{aligned}
\sigma V_t &\propto \sqrt{q^2 N_{\rm SUB} W_{\rm DEP}} \\
\sigma V_t &\propto \sqrt{\frac{q(V_t - V_{FB} - 2\phi_F)}{C_{OX}}} \\
\sigma V_t &\propto \sqrt{(V_t - V_{FB} - 2\phi_F)} \\
\sigma V_t &\propto \sqrt{V_t + 0.1}. \qquad (7)
\end{aligned}$$

As a result, threshold voltage variation increases roughly with square-root relationship to V_t (here, $-V_{FB} - 2\phi_F \approx 0.1$, which has been estimated from several data-points corresponding to 65 nm fabs [10]). The corresponding impact on on-current is shown in Fig. 8(b), where a constant V_{DD} of 1.0 V is assumed. As shown, the deviation between mean current and 4σ current widens drastically as V_t is scaled. Once again, the resulting impact on bit-cells is two-fold: as V_t is increased: 1) threshold-voltage variation accounts for a larger proportion of the gate-

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a) (b) (c) (d)

Fig. 6. Sub-array total energy (at room temperature) for various performance requirements (specified by $T_{\mathrm{CYC,RTN}}$). The arrow indicates the direction of energy reduction by a factor of 1.22 x between adjacent contours. (a) Total energy for $T_{\mathrm{CYC,RTN}} = 10$ ms. (b) Total energy for $T_{\mathrm{CYC,RTN}} = 1$ ms. (c) Total energy for $T_{\mathrm{CYC,RTN}} = 100$ μs. (d) Total energy for $T_{\mathrm{CYC,RTN}} = 10$ μs.

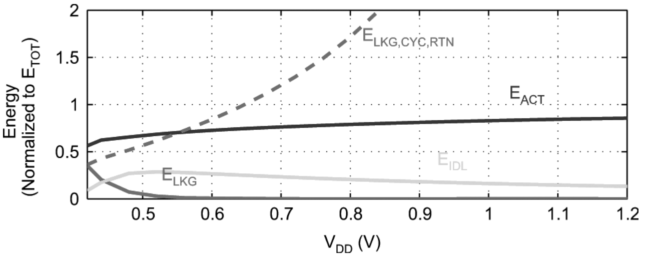
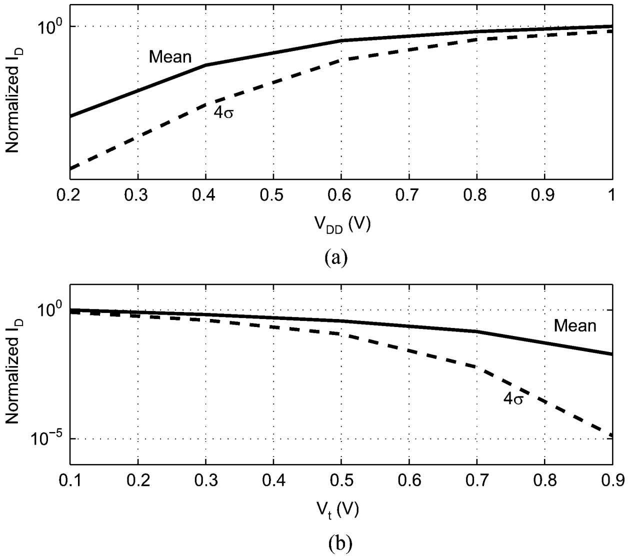


Fig. 7. Normalized energy components at $T_{\mathrm{CYC,RTN}} = 10$ ms versus V_{DD}; the excess energy consumed by leakage is severe if no provision is made for a low-leakage idle-mode.

overdrive and 2) the standard-deviation of threshold-voltage increases, leading to worse absolute device variation.

B. Bit-Cell Level Impact of Variation

The device degradations exhibited in Fig. 8 oppose the low V_{DD} and high V_t targets for energy minimization established in Section III-B. Fig. 9(a)–(c) shows contour plots for the read, hold, and write margins. The contour labels indicate the respective voltage margin at the corresponding V_{DD} and V_t for the 4σ level (with global-variation) of the 0.25 μm^2 bit-cell. The read margin is represented by the read static-noise margin (SNM) [11]. The write-margin measures the ability to force a desired data-state in the bit-cell by enabling its access devices and selectively clamping one bit-line to V_{DD} and the other to ground (here, the negative of the SNM lobe-measurement is



(a)

(b)

Fig. 8. Drain-current versus V_{DD} and V_t showing deviation between nominal and 4σ cases in the presence of variation. (a) Drain-current versus V_{DD}. (b) Drain-current versus V_t.

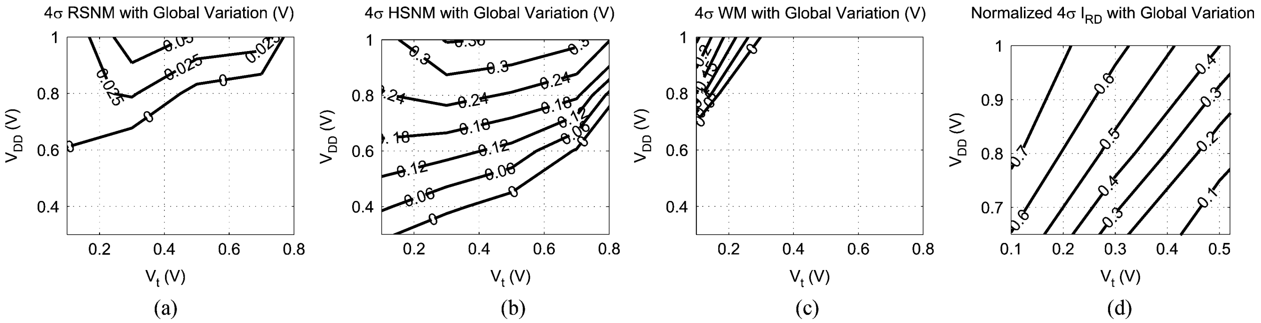


Fig. 9. Bit-cell operational margins and read-current versus V_{DD} and V_t in the presence of variation. (a) Read static-noise-margin. (b) Hold static-noise-margin. (c) Write-margin. (d) Read-current.

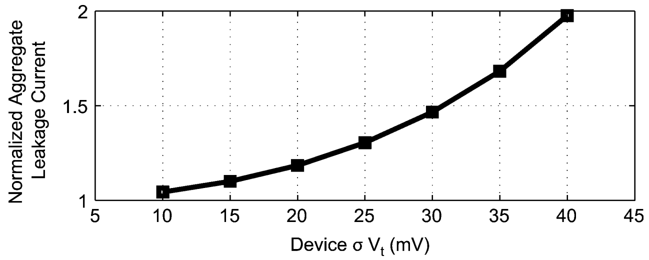


Fig. 10. Total aggregate leakage-current in the presence of variation normalized to nominal aggregate leakage-current assuming no variation.

used [12]). It should be noted, however, that one method of improving write-margin without effecting density is through independent V_t adjustment to strengthen the nMOS devices (access and driver) and weaken the pMOS devices (load). For the purposes of establishing simple trends, however, the current analysis assumes equivalent device threshold voltages.

The hold SNM has important implications on the sub-array's idle-mode energy; although Section III assumed a $V_{\rm DRV}$ of 0.4 V, in reality, $V_{\rm DRV}$ is set by the minimum supply-voltage at which the hold-margin can be guaranteed. Namely, this corresponds to V_{DD} along the contour in Fig. 9(b) where the hold SNM is equal to zero. Since V_t affects the variation exhibited by the bit-cell devices, this minimum V_{DD} increases with V_t in the manner shown, diminishing the idle-mode leakage-energy reduction possible through V_t scaling.

Another important bit-cell metric is read-current. Read-current affects the sub-array performance, which ultimately restricts V_{DD} and V_t scaling for energy savings. The 4σ read-current normalized to the nominal read-current is plotted in Fig. 9(d), showing the increasing severity of performance degradation due to variation as V_{DD} and V_t are scaled.

Another impact of V_t variation is that it increases the aggregate leakage-currents due to their exponential relationship with V_t. As a result, positive perturbations have larger magnitude than negative perturbations, and Fig. 10 shows that the net effect is that the leakage current summed over the entire array is significantly elevated even for modest standard deviations in V_t. Since higher V_t results in higher σV_t, the leakage-energy savings are somewhat degraded.

It is worth noting that several other metrics for read and write margin also exist, especially to represent the impact of dynamic error sources such as power-supply noise, bit-line/word-line transients, and capacitive coupling [13]–[15]. These metrics may be analyzed in a manner similar to that shown in Fig. 9.

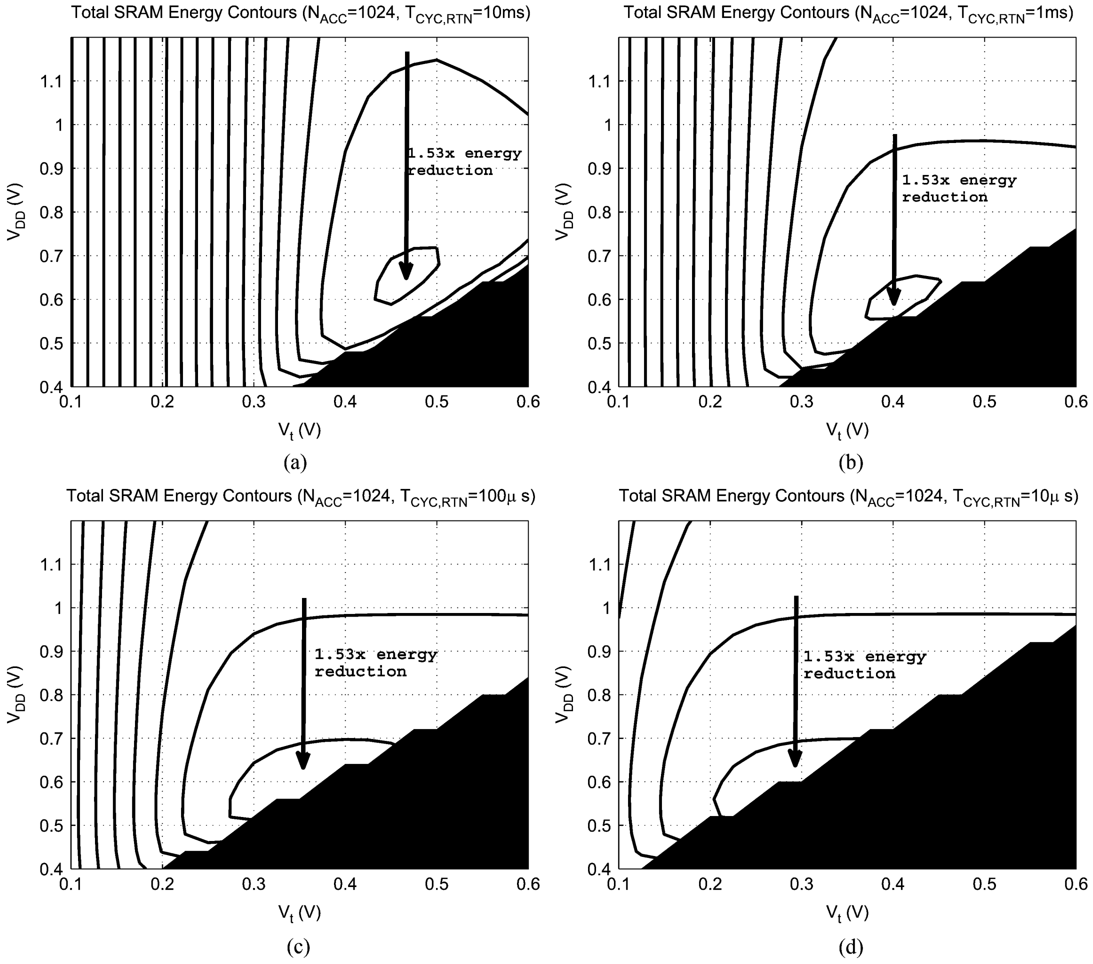
C. Energy Contours With Variation

In addition to degrading operating margins, variation affects the sub-array parameters that determine the energy. For instance, these include the aggregate leakage-current, the data-retention voltage, and the access-time (through read-current). The affect on these parameters with respect to V_{DD} and V_t necessitates revision of the total energy analysis. The revised total energy contours are plotted in Fig. 11. Although the previous targets remain relevant, several notable differences emerge.

For the cases with low performance requirement (i.e., $T_{\rm CYC,RTN} = 10\ \mu{\rm s}, 1\ \mu{\rm s}$), the optimal energy point occurs outside the blocked region. This implies that it is preferable to spend a portion of the retention-cycle in the idle-mode rather than scaling V_{DD} or V_t to extent allowed by $T_{\rm CYC,RTN}$. The underlying reason is that long retention-cycle periods primarily require high V_t in order to minimize leakage-currents. An accompanying effect of V_t elevation, however, is that severe variation degrades performance. Since active-energy is less critical than total leakage-energy in these cases, V_{DD} increase has low relative cost but allows the access-times to be reduced considerably, mitigating the active-mode leakage-energy. As a result, higher V_{DD} is required (while maintaining high V_t).

In all cases, the blocked region is considerably expanded owing to the performance degradation brought on by variation. The effect of this is particularly detrimental in the high throughput cases (i.e., $T_{\rm CYC,RTN} = 100\ \mu{\rm s}, 10\ \mu{\rm s}$), where the energy contours continue decreasing into the blocked region. As a result, the energy-savings achievable by V_{DD} and V_t scaling are restricted.

In addition to affecting the energy trends, variation strongly impacts the absolute total energy. For instance regardless of the V_{DD} and V_t point, the leakage-current is elevated in the presence of variation due to the exponential dependence on V_t. This combines with the $V_{\rm DRV}$ dependence on V_t to increase the absolute leakage-power during the idle mode. Accordingly, with a V_t of 0.5 V, the idle-mode power is increased by over 2×. The minimum total energy achievable in the four cases ($T_{\rm CYC,RTN} = 10$ ms, 1 ms, $100\ \mu{\rm s}$, $10\ \mu{\rm s}$) increases by 1.65×, 1.56×, 1.28×, and 1.33×, respectively, in the presence of variation.



(a)

(b)

(c)

(d)

Fig. 11. Sub-array total energy in the presence of variation (at room temperature) for various performance requirements (specified by $T_{\mathrm{CYC,RTN}}$). (a) Total energy for $T_{\mathrm{CYC,RTN}} = 10$ ms. (b) Total energy for $T_{\mathrm{CYC,RTN}} = 1$ ms. (c) Total energy for $T_{\mathrm{CYC,RTN}} = 100\ \mu$s. (d) Total energy for $T_{\mathrm{CYC,RTN}} = 10\ \mu$s.

V. Conclusion and Design Directions

This paper presents the total energy analysis for SRAMs over active and idle operating modes. For the design perspective, this motivates two important directions for SRAM energy minimization. The first is a need for low-voltage SRAMs that can reliably ensure operating margins in high-V_t (low leakage) technologies. The second is the need to improve performance as much as possible to make the low-V_{DD} and high-V_t operation viable while meeting system throughput constraints.

With regards to V_{DD} and V_t scaling, it is worth noting that practical SRAMs require some engineering margin (i.e., 100–200 mV on V_{DD}). Thus, operation in the range of 0.3 V must be ensured even if a V_{DD} of 0.5 V is targeted by the minimum energy analysis. A promising approach for such voltage levels is the 8 T cell [16]. Its essential benefit is the introduction of a read buffer, which overcomes the read margin limitation of 6 T cells. Additionally, since separate devices are used for reads and writes, the two operations can be optimized individually through selective device design and cell biasing. For instance, the read buffer can be optimized for high read-current (e.g., by lowering V_t's) while the storage cell can be optimized for low-leakage. Thanks to these opportunities, several ultra-low-voltage buffered-read designs have been reported [17]–[19]. Further while 6 T cells require fewer devices, 8 T cells can be more area efficient at low-voltages by allowing the use of *smaller* devices thanks to the robustness against variation that is afforded by the wider operating margins [20]. It is worth noting, however, that the aspect ratio of the two layouts will result in different word-line and bit-line lengths, impacting the active-energies in different manners.

Sense amplifier design is an important area of focus for improving performance without impacting the bit-cells. The essential challenge is maximizing small-signal sensitivity while ensuring robustness to coupling noise, variation, and bit-line signal degradation, which all raise the need for increased sensing margin. For instance, in [19] a replica column is used to estimate bit-line leakage and bias the sense-amplifiers appropriately in the face of signal degradation. In [18] a bank of small sense-amplifiers is incorporated to combat variation by selecting only one that exhibits small enough offset. Finally, in [21] a single-ended sense-amplifier is used for compatibility with 8 T cells, while offset compensation and regeneration are used to achieve small-signal sensing.

Acknowledgment

The author would like to thank Prof. A. Chandrakasan (MIT) and the production SRAM team from TI (D. Buss, M. Clinton,

X. Deng, T. Houston, W.-K. Loh, and H. Mair) for their input and support.

REFERENCES

[1] J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann, and A. Chandrakasan, "A 65 nm sub-v_t microcontroller with integrated SRAM and switch-capacitor DC-DC converter," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2008, pp. 318–319.

[2] V. George, S. Jahagirdar, C. Tong, K. Smits, S. Damaraju, S. Siers, V. Naydenov, T. Khondker, S. Sarkar, and P. Singh, "Penryn: 45-nm next generation intel core 2 processor," in *Proc. IEEE Asian Solid-State Circuits Conf.*, Nov. 2007, pp. 14–17.

[3] R. W. Brodersen, M. A. Horowitz, D. Markovic, B. Nikolic, and V. Stojanovic, "Methods for true power optimization," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, Nov. 2002, pp. 35–42.

[4] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal suppply and threshold scaling for sub-threshold CMOS circuits," in *Proc. IEEE Comp. Soc. Annu. Int. Symp. VLSI*, Apr. 2002, pp. 5–9.

[5] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proc. Int. Symp. Low Power Electron. Des.*, Aug. 2005, pp. 20–25.

[6] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "SRAM leakage suppression by minimizing standby supply voltage," in *Proc. IEEE Int. Symp. Quality Electron. Des.*, Mar. 2004, pp. 55–60.

[7] Y. Wang, H. Ahn, U. Bhattacharya, T. Coan, F. Hamzaoglu, W. Hafez, C.-H. Jan, R. Kolar, S. Kulkarni, J. Lin, Y. Ng, I. Post, L. Wel, Y. Zhang, K. Zhang, and M. Bohr, "A 1.1 GHz 12 μA/mb-leakage SRAM design in 65 nm ultra-low-power CMOS with integrated leakage reduction for mobile applications," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 324–325.

[8] K. Zhang, U. Bhattachalya, 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.

[9] J. B. Kuang, H. C. Ngo, K. J. Nowka, J. C. Law, and R. V. Josi, "A low-overhead virtual rail technique for SRAM leakage power reduction," in *Proc. IEEE Int. Conf. Comput. Des.*, Oct. 2005, pp. 574–579.

[10] K. Takeuchi, T. Fukai, T. Tsunomura, A. T. Putra, A. Nishida, S. Kamohara, and T. Hiramoto, "Understanding random threshold voltage fluctuation by comparing multiple fabs and technologies," in *IEDM Dig. Tech. Papers*, Dec. 2007, pp. 467–470.

[11] 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. 10, pp. 748–754, Oct. 1987.

[12] B. Calhoun, "Low energy digital circuit design using sub-threshold operation," M.S. thesis, Massachusetts Inst. Technol., Boston, 2005.

[13] M. Khellah, D. Khalil, D. Somasekhar, Y. Ismail, T. Karnik, and V. De, "Effect of power supply noise on SRAM dynamic stability," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2007, pp. 76–77.

[14] M. Khellah, Y. Ye, N. S. Kim, D. Somasekar, G. Pandya, A. Farhang, K. Zhang, and V. De, "Wordline and bitline pulsing schemes for improving SRAM cell stability in low-Vcc 65 nm CMOS designs," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2006, pp. 12–13.

[15] B. Zhang, A. Arapostathis, S. Nassif, and M. Orshansky, "Analytical modeling of SRAM dynamic stability," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, Nov. 2006, pp. 315–322.

[16] 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.

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

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

[19] T.-H. Kim, J. Liu, J. Kean, and C. H. Kim, "A high-density subthreshold SRAM with data-independant bitline leakage and virtual ground replica scheme," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 330–331.

[20] N. Verma, "Ultra-low-power SRAM design in high variability advanced CMOS," M.S. thesis, Massachusetts Inst. Technol., Boston, 2009.

[21] N. Verma and A. Chandrakasan, "A high-density 45 nm SRAM using small-signal non-strobed regenerative sensing," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2008, pp. 380–381.

Naveen Verma 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. and Ph.D. degrees in electrical engineering from Massachusetts Institute of Technology, Cambridge, in 2005 and 2009, respectively.

Since July 2009, he has been an Assistant Professor with the Department of Electrical Engineering, Princeton University, Princeton, NJ. His research focuses on ultra-low-power integrated circuits including low-voltage digital logic and SRAMs, low-noise analog instrumentation and data-conversion, and energy-efficient processing algorithms especially for biomedical applications.

Prof. Verma was a corecipient of 2008 ISSCC Jack Kilby Award for Outstanding Student Paper and 2006 DAC/ISSCC Student Design Contest Award.