Approximate SRAMs With Dynamic Energy-Quality Management

Fabio Frustaci, *Member, IEEE*, David Blaauw, *Fellow, IEEE*, Dennis Sylvester, *Fellow, IEEE*, and Massimo Alioto, *Fellow, IEEE*

***Abstract*— In this paper, approximate SRAMs are explored in the context of error-tolerant applications, in which energy is saved at the cost of the occurrence of read/write errors (i.e., signal quality degradation). This analysis investigates variation-resilient techniques that enable dynamic management of the energy-quality tradeoff down to the bit level. In these techniques, the different impacts of errors on quality at different bit positions are explicitly considered as key enabler of energy savings that are far larger than a simple voltage scaling. The analysis is based on the experimental results in an energy-quality scalable 28-nm SRAM and the extrapolation to a wide range of conditions through the models that combine the individual energy contributions. Results show that the joint adoption of multiple bit-level techniques provides substantially larger energy gains than individual techniques. Compared with the simple voltage scaling at isoquality, the joint adoption of these techniques can provide more than 2× energy reduction at negligible area penalty. Energy savings turn out to be highly sensitive to the choice of joint techniques, thus showing the crucial importance of dynamic energy-quality management in approximate SRAMs.**

***Index Terms*— Approximate computing, energy-quality tradeoff, error tolerant, near threshold, SRAM, ultralow-power processing, voltage overscaling.**

I. Introduction

IN THE last years, the approximate computing design paradigm has been investigated in the context of error-tolerant applications [1]–[9]. Such applications can tolerate a certain bit error rate (BER) without severely compromising the correctness of the overall computation or the user experience. The related applications have become predominant with the advent of cloud/mobile computing, e.g., multimedia, big data, Web search, computer vision, machine learning, sensor fusion, and augmented reality [1], [10]. Approximations are inherent in most image, audio, and video lossy compression algorithms [11], and introducing approximation at the hardware level can extend the energy/performance benefits that are achieved when the quality requirements are relaxed.

Due to their nature, error-tolerant applications allow a more aggressive supply voltage (V_{DD}) scaling, compared with the error-free applications where errors are strictly prohibited. However, when V_{DD} scales down, the impact of process variations becomes heavier, and the SRAM BER increases ungracefully (exponentially) at voltages below the minimum operating voltage V_{min} [1]–[10]. Hence, very limited voltage and energy reduction are actually possible under practical quality targets [12], [13]. At the system level, the ungraceful quality degradation is an even more crucial limit, as the SRAM typically limits the overall minimum voltage [14]–[20].

In general, the impact of errors on quality is different for different bit positions. For example, the quality of the user experience in multimedia applications is mainly defined by the most significant bits (MSBs) [21]–[26]. This is true also for a very wide range of applications, such as big data, multimedia, machine learning, and several others [27]. In SRAMs, this observation has been exploited by: 1) storing MSBs in more robust bitcells (i.e., larger transistor size/count and supply voltage), while saving area and/or energy in LSBs by using bitcells with a smaller footprint [21]–[23]; 2) suppressing part of their error correcting code (ECC) bits [24]; and 3) lowering their supply voltage [26]. Unfortunately, the first and second classes of approximate SRAMs set a fixed energy-quality tradeoff at design time [21]–[24], and are, hence, unable to dynamically track the time-varying quality requirement and correspondingly minimize the energy [1]–[10]. These techniques require costly bitcell redesign and manual array reorganization. In addition, none of these three classes addresses the fundamental issue of the ungraceful quality degradation at low voltages and the resulting limit to true energy savings. The first SRAM with dynamic error-quality management and graceful quality degradation was proposed in [12]. Selective techniques were introduced to enable dynamic management of the energy-quality tradeoff down to the bit level. The ability to improve the resiliency of a dynamically adjustable number of bits permits to: 1) achieve graceful quality degradation at low V_{DD}; 2) limit the energy cost of improved resiliency for a given quality; and 3) enable more aggressive voltage/energy scaling for quadratic energy reduction [12].

This paper presents a wide exploration of several selective bit-level techniques to manage the energy-quality tradeoff, through energy-quality models that are solidly based on

Manuscript received July 1, 2015; revised October 7, 2015; accepted November 13, 2015. This work was supported in part by the Singaporean Ministry of Education under Grant MOE2014-T2-1-161 and Grant MOE2014-T2-2-158, in part by the Australian Cancer Research Foundation through the Project entitled Sub-Cycle Error Correction for Resilient Ultralow Voltage VLSI Processing under Grant RG00003061, in part by the National Science Foundation Variability Expedition, and in part by STMicroelectronics for chip fabrication.

F. Frustaci is with the Department of Computer Engineering, Modeling, Electronics and Systems, University of Calabria, Rende 87036, Italy (e-mail: ffrustaci@deis.unical.it).

D. Blaauw and D. Sylvester are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: blaauw@umich.edu; dennis@eecs.umich.edu).

M. Alioto is with the Department of Electronics and Communication Engineering, National University of Singapore, Singapore 117583 (e-mail: malioto@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2015.2503733

measurements of a 28-nm SRAM testchip [12]. As a main contribution, this paper provides: 1) an insight into the energy gains of each technique under a wide range of conditions (e.g., voltage, bank size, and word size), providing justification to trends and technology-independent results whenever possible; 2) a wider comparison that includes five bit-level techniques; 3) the investigation of the impact of ECC code on the energy-quality tradeoff; and 4) the first investigation of the joint adoption of multiple bit-level techniques and their interaction.

This paper is organized as follows. Section II describes the BER-quality relationship in error-tolerant SRAMs. Sections III–V deal with the individual selective techniques to manage the energy-quality tradeoff. The joint adoption of these techniques is explored in Section VI, and its dynamic adjustment is discussed in Section VII. Conclusions are drawn in Section VIII. The Appendix provides details on the analysis methods.

II. Error-Tolerant SRAMs: Errors and Quality

SRAM bitcell read/write errors are mainly determined by the inadequate bitcell read margin (RM) and write margin (WM) [14]. Although these contributions appear to be random across different dice, they repeatedly have the same effect in a given die.[1] Since variations affect RM and WM in an opposite way, the process corner defines the critical margin between the two: the slow–fast (SF) corner makes the bitcell write critical (i.e., inadequate WM is responsible for nearly all bitcell failures), whereas the fast–slow corner makes the read critical. Adequately robust operation is required at both corners to keep write and read failures under control.

Under traditional SRAM designs, no differentiation is made across bitcells, and failures occur uniformly within the array. When V_{DD} is scaled down, the process variations degrade both WM and RM, thus determining the well-known exponential increase in the read and write BERs at lower voltages, which ultimately results in a very ungraceful quality degradation at low voltages [21]–[25]. Although the quality is qualitatively related to the BER, their relationship strongly depends on the application and the data representation. In image and video processing, a widely used metric is the peak signal-to-noise ratio (PSNR), which is defined as the ratio of the largest pixel value and the rms error [29]. This metric retains the meaning of SNR even when used in other applications.[2] Accordingly, PSNR will be used as a quality metric for the following examples in the image/video processing domain, although it actually quantifies the impact of failures in a much broader range of applications. Fig. 1 shows the measured dependence of PSNR on BER for the realized 28-nm SRAM testchip [12]. Practical PSNR targets are in the order of 25–30 dB or higher [25], [29]. Our analysis shows that the measured PSNR is highly consistent across all image benchmarks in [30], with a maximum deviation of only 0.6 dB (and 0.3 dB on average), thus confirming the suitability of PSNR as a representative and general metric.

[1]This paper ignores other sources of bit failures that are occasional (e.g., soft errors) or random (e.g., erratic bits) in a specific die, as their error rates are negligible compared with those encountered in error-tolerant applications [28].

[2]In the definition of PSNR, the failures are treated as noise. In applications different from image/video processing, the only difference is that the pixel is replaced by a string of bits that quantify the information of interest (e.g., a sample in signal processing). By definition, the higher values of PSNR correspond to higher quality.

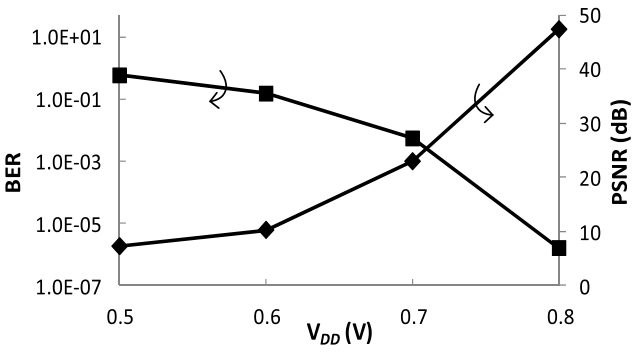

Fig. 1. Measured BER and resulting PSNR versus V_{DD} (SF corner, 22 °C) [12].

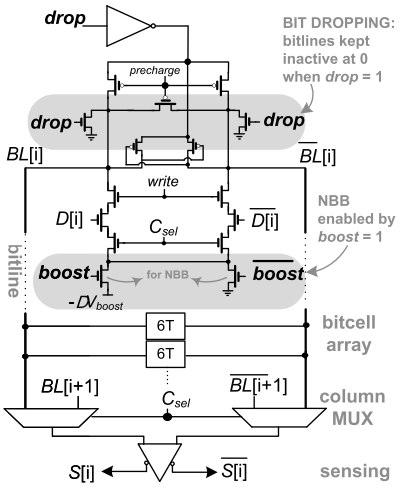

Fig. 2. Schematic of the selective bitline precharge enabling the bit dropping technique [circuit details on SNBB are also shown (see Section IV for details)].

III. LSB Dropping and Dual-V_{DD} Techniques Under Voltage Scaling

In this section, LSB dropping and dual-V_{DD} techniques are explored to reduce energy when lower bit precision is acceptable. Bit dropping consists in disabling the bitlines corresponding to a given number of LSBs, to linearly reduce the energy at reduced quality. This is different from the dual-V_{DD} scheme in [26], where the LSBs are instead powered at lower supply voltage, rather than being dropped. At the circuit level, bit dropping is implemented within the bitline precharge circuit, as shown in Fig. 2. The drop signal disables the precharge circuit during read and write operations, thus saving dynamic energy. Fig. 3 shows the PSNR versus the number of dropped LSBs, based on measurements on the testchip for video processing in [12] and the analysis methodology in the Appendix, assuming that the 32-bit SRAM word comprises four 8-bit pixels. From Fig. 3, the quality of the

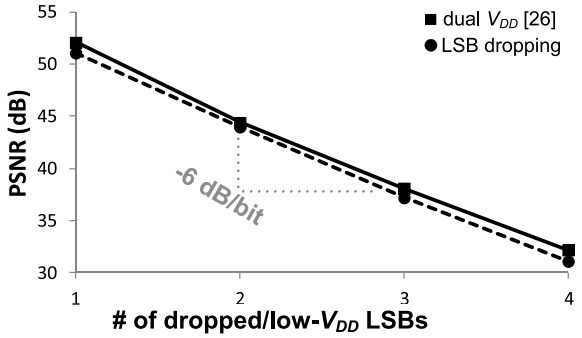

Fig. 3. Measured PSNR for bit dropping and dual-V_{DD} scheme [26] versus number of dropped/low-V_{DD} LSBs at $V_{DD} = 0.8$ V (write-critical corner, 22 °C, $V_{DD} = 0.5$ V for low-V_{DD} LSBs in dual-V_{DD} scheme).

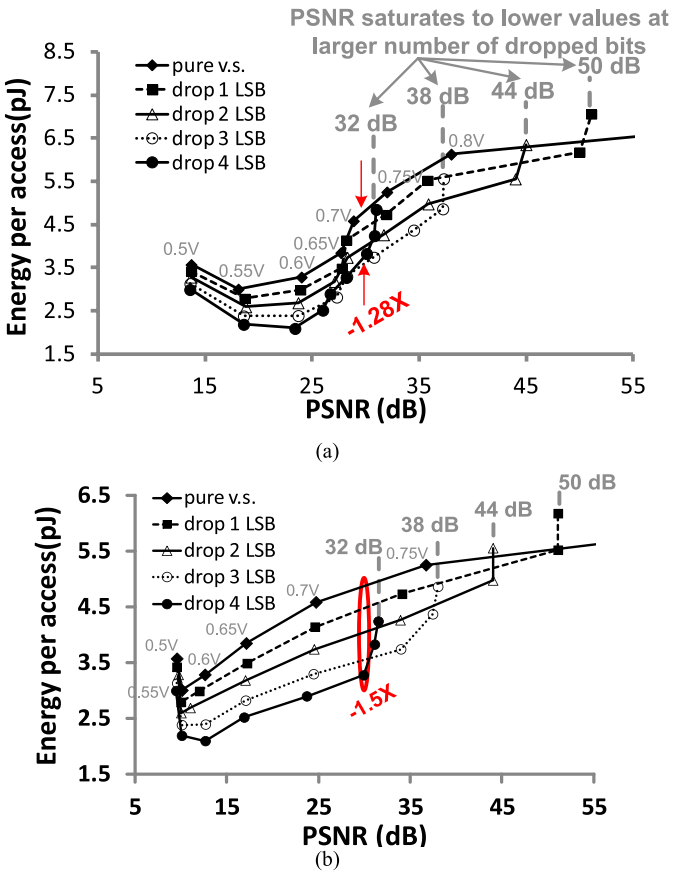

Fig. 4. Measured energy-quality tradeoff for pure voltage scaling and bit dropping technique (22 °C). (a) Read-critical corner. (b) Write-critical corner.

dual-V_{DD} scheme in [26] is approximately the same as bit dropping, as expected from the ungraceful BER increase in LSBs at lower voltages, which makes LSBs mostly incorrect, which is indeed equivalent to the case of their complete suppression. From Fig. 3, the quality decays by 6 dB for each additional dropped bit, as expected from the quantization noise theory[3] [32].

The above results clearly show that the LSB dropping is a more energywise approach, since it completely eliminates the energy associated with dropped bitlines at very similar quality as in [26]. The circuit implementation of LSB dropping is very simple and has negligible area cost, as it simply needs to precharge both bitlines to zero, when the corresponding bit needs to be dropped. Instead, the scheme in [26] requires the insertion of buffers in the intermediate sections of wordlines, which increases the memory area, makes it harder to maintain regularity, and gives rise to reliability issues. From the point of view of the energy-quality tradeoff, LSB dropping is equally effective at any corner, as it simply reduces the activity and, hence, the energy associated with the corresponding bitlines, regardless of voltage and process variations.

Fig. 4(a) and (b) shows the energy-quality tradeoff under joint bit dropping and voltage scaling for read- and write-critical corners, respectively. In this case, the quality is degraded due to the loss of information on the dropped LSBs and the bitcell failures at low voltage. The former contribution dominates at higher voltages [0.75 V or more in Fig. 4(a) and (b)], and the PSNR, hence, saturates to a value that decreases when a larger number of bits is dropped. On the other hand, at lower voltages, the bitcell failures dominate and the PSNR is further reduced. By the definition of PSNR, the absolute value at which it saturates at higher voltages depends on the specific data set considered (e.g., image and frame). However, the distance between the value at which PSNR saturates at different numbers of dropped bits is independent of the specific data set. More specifically, an increase in the number of dropped bits by one reduces the effective resolution by one bit, thus degrading the SNR due to quantization (hence, PSNR) by 6 dB, as expected from the quantization noise theory [32]. All PSNR-energy values have been measured at the maximum operating frequency, thus the failures are due to degraded cell margins, and not to timing failures. It is worth noting that the PSNR is calculated with respect to the original reference image, that is quantized with 8-bit per pixel, hence, when no error occurs, the PSNR assumes an infinite value by definition.

From a design standpoint, the number of dropped bits needs to be set to minimize the energy while achieving the targeted quality. On this respect, let us compare the energy-quality curves of two arbitrary configurations that drop i and $(i-1)$ LSBs in Fig. 4(a) and (b), under the same V_{DD}. Here, i is assumed to be small enough to make the targeted PSNR achievable, i.e., the quality under i dropped bits saturates at a PSNR that is larger than the targeted one [clearly, the same holds for the case of $(i-1)$ dropped bits, as the quality saturates at even larger PSNR]. Due to the above discussed quality saturation (i.e., energy vertical asymptote), the curve with i dropped bits is more energy efficient for PSNR lower than its saturation value, compared with the configuration with $(i-1)$ dropped bits. For example, dropping four bits is more energy efficient than dropping three or less bits for PSNR <32 dB.

Once the number of dropped bits is set, the voltage scaling enables further energy reduction, which in Fig. 4(a) and (b) is, respectively, up to 1.28× and 1.5× for the read- and write-critical corners, for practical quality targets of 30 dB and higher.

[3]LSB dropping is equivalent to degrading resolution by one bit, increasing the quantization noise and degrading the SNR by 6 dB/bit [32].

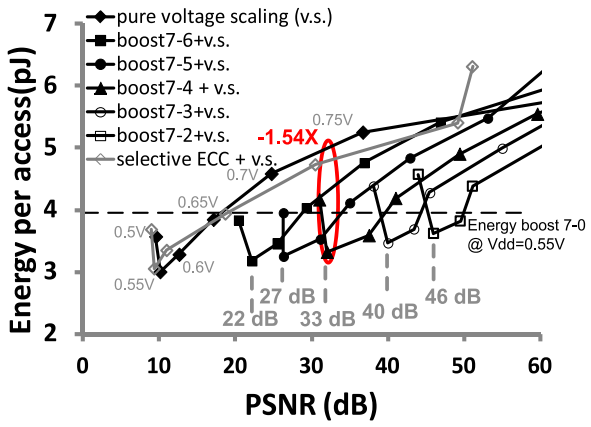

Fig. 5. Measured energy-PSNR tradeoff under SNBB (write-critical corner, 22 °C, −130 mV NBB voltage) and SECC Hamming(15, 11).

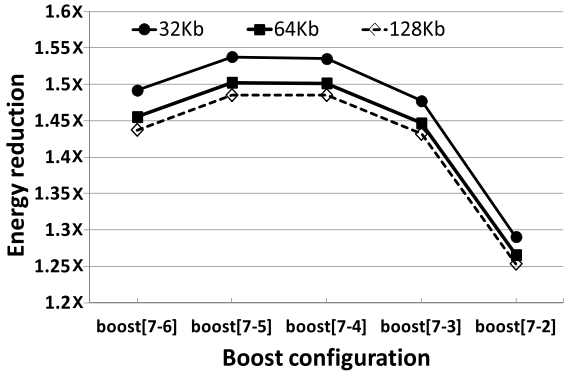

Fig. 6. Net energy saving of SNBB versus simple voltage scaling for different SRAM sizes (simulation, SF corner, $V_{DD} = 0.55$ V, $T = 22$ °C).

From the same figures, the minimum energy point is slightly affected by the bit dropping. Indeed, the minimum energy under pure voltage scaling is obtained at $V_{DD,min} = 0.55$ V, whereas it is obtained at 0.6 V when three or more LSBs are dropped.

Finally, the energy savings associated with bit dropping are essentially proportional to the number of dropped bits, regardless of the array size and the word length. Hence, the latter two parameters do not affect the above presented results.

IV. Selective Assist Under Voltage Scaling

LSB dropping achieves linear energy reduction by reducing the effective resolution of the stored data. More substantial energy savings (e.g., quadratic) require aggressive V_{DD} reduction, which in turn is severely limited by the ungraceful quality degradation at low voltages. To make such degradation graceful, the traditional variation-resilient (assist) techniques can be unconventionally used to improve the robustness of an adjustable number of MSBs (as set by the quality target), leaving LSBs unprotected to minimize the assist energy cost [12]. The ultimate goal of this approach is to make the energy-quality tradeoff more favorable when a chip is skewed toward the write-critical process corner. In the following, we will consider negative bitline boosting (NBB) as a representative example of the write-assist technique. Any other column-based assist technique is suitable for the same purpose.

In NBB, a strong zero is written on the bitcell by setting the corresponding bitline voltage to a negative voltage $-\Delta V_{boost}$ instead of ground [33], thus improving the write-ability of the bitcell. Only two additional transistors per column are needed in the precharge circuit to select either ground or $-\Delta V_{boost}$ voltage, which is provided off-chip for simplicity,[4] as depicted in Fig. 2. In columns, where NBB is activated, the write energy per bitline is increased by a factor $[(V_{DD} + \Delta V_{boost})/V_{DD}]^2$, due to the increase in the bitline voltage swing. In the following, the notation boost$[i - j]$ indicates that selective NBB (SNBB) is enabled for the columns $i \ldots j$. As an example, the boosting voltage that is required to ensure a given BER target at $V_{DD} = 0.5$ V is shown in Table I. The latter shows that the BER improves very rapidly when increasing ΔV_{boost}, and the values of ΔV_{boost} that cover practical BER targets are in the range of 100–150 mV.

From the above considerations, the SNBB permits to reduce the energy cost of NBB, by restricting it to a fraction of the bit positions. This provides significant energy savings compared with a traditional approach, where errors are equally prevented at all bit positions, as discussed.

A. SNBB Under Voltage Scaling

The measured energy-PSNR tradeoff obtained under SNBB is plotted in Fig. 5 [12]. Being a write-assist technique, SNBB is effective at the write-critical corner, and irrelevant at the read-critical corner (the related curves are omitted accordingly). The quality improves when increasing the number of boosted columns for a given V_{DD}, thus permitting to further reduce V_{DD} (i.e., energy) for a given quality target. The net energy saving comes from the difference of the energy reduction due to the decrease in V_{DD} and the additional energy spent for bitline boosting. As an example, in Fig. 5, the boost[7-4] configuration reduces energy (voltage) by up to 1.54× (from 0.75 down to 0.55 V) compared with a pure voltage scaling at isoquality. Other SNBB schemes offer different energy/quality tradeoffs. For low values of PSNR around 25–30 dB, boost[7-6] is the most energy-efficient configuration, with an energy saving over a pure voltage scaling of 1.31×. At higher PSNR (∼45 dB), boost[7-2] turns out to be the most energy efficient, with an energy saving of 1.5×.

Observe that the pure voltage scaling exhibits a minimum energy point that is placed at impractically low-quality targets, and is, hence, unreachable in practical cases. Instead, the minimum energy point under SNBB is always within practical quality targets. From Fig. 5, the energy-quality curves under SNBB are similar, as they essentially differ for a rigid shift to the left, for schemes with a lower number of boosted bitlines. A left-shift by ∼6 dB is observed when the number of boosted

[4] An on-chip implementation would slightly increase the energy entailed by NBB, due to the nonideal efficiency of the boosting circuitry. In this case, the proposed selective techniques would exhibit an even larger advantage, as a more pronounced energy reduction would be allowed by the selective suppression of NBB in LSBs.

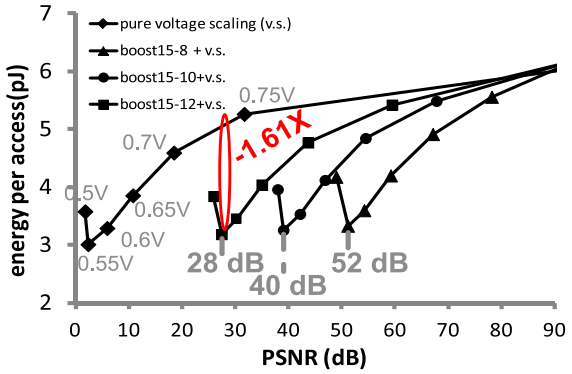

Fig. 7. Energy-PSNR tradeoff under SNBB on the same array with 16-bit subword (measurements, write-critical corner, 22 °C).

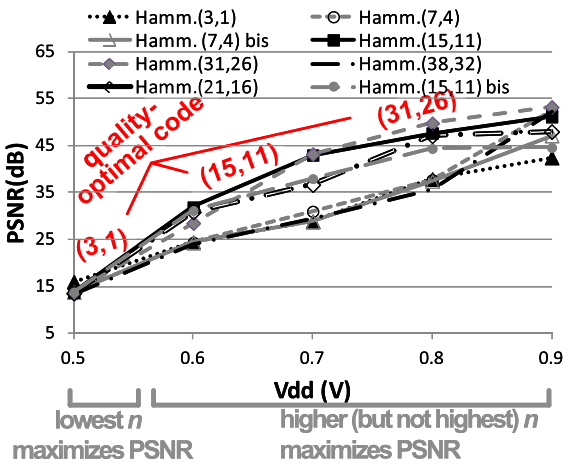

Fig. 8. PSNR versus supply voltage for several ECC schemes (read-critical corner).

bitlines is reduced by one, as expected from the corresponding resolution reduction [32].

B. Impact of Array Size and Subword Length

Larger array sizes increase the total number of errors but keep the same BER and the same random distribution across bit positions. On the other hand, the quality is essentially affected by the probability of having errors in MSBs, which clearly does not change due an increase in the array size due to the same random distribution. Hence, the quality is essentially unaffected by the array size. This was confirmed by extensive MATLAB simulations that assumed random distributions with the same BER as the considered testchip[5] (omitted for the sake of brevity). Results show that the PSNR varies by less than 1 dB when increasing the array size from 32 to 128 kb at $V_{DD} = 0.7$ V or less. Similarly, the energy savings are essentially the same regardless of the array size, as shown in Fig. 6. Thus, the above results on the energy-quality tradeoff are generally valid, independent of the array size.

The results presented in Section VI-A can be generalized to arbitrary array subword length. Indeed, for the same reasons discussed in Section VI-A, the energy-quality curves under SNBB are expected to approximately differ by a 6-dB rigid shift to the left, for each reduction in the number of boosted bitlines by one. This is confirmed in Fig. 7, which plots the energy-quality tradeoff for a 32-kb array with the 32-bit word reorganized into two 16-bit sub- words. For example, the minimum energy point of the boost[15-10] configuration is placed at 12 dB to the left compared with boost[15-8], as the latter has two additional boosted bitlines compared with the former.

To gain a deeper insight into the impact of the subword length, let us estimate the PSNR degradation when $B < N$ columns are boosted and the same information (i.e., with the same peak value) is represented with doubled number of bits (i.e., from N to $2N$). For the N-bit subword, the errors and the corresponding noise are confined in the least $N - B$ positions. In the case of $2N$-bit subword, the additional N unprotected LSBs introduce an overall noise contribution that is comparable[6] with the noise contribution of the single bit that is immediately more significant (i.e., the LSB in the N-bit subword). Hence, the PSNR degradation due to the subword length doubling is equivalent to the PSNR degradation due to errors in the LSB of the original N-bit subword, which in turn amounts to ∼6 dB, as discussed above. This is confirmed in Figs. 5 and 7, as the energy-PSNR curves in Fig. 7 are shifted to the left by ∼6 dB when boosting four bitlines in a 16-bit subword (boost[15-12] in Fig. 7), as compared with the 8-bit subword with the same number of boosted bitlines (boost[7-4]) in Fig. 5.

A more pronounced energy benefit from SNBB is found in arrays with a longer subword under the same number of boosted bitlines. As an example, Fig. 8 shows that the SNBB enables an energy saving by up to 1.61× with respect to pure voltage scaling under 16-bit subword, as compared with 1.54× found in 8-bit subword with the same four boosted bitlines (see Fig. 6). This is because the energy cost of the same number of boosted bitlines is amortized across a larger number of columns in the case of doubled subword length, thus the NBB energy cost becomes a smaller fraction of the total energy.[7]

V. Selective ECC Under Voltage Scaling

As another fundamental class of variation-resilient techniques, ECC corrects errors regardless of their nature (write or read). As opposed to the traditional uniform ECC that equally protects all bits [35], selective ECC (SECC) mitigates failures only in bit positions that have a stronger impact on quality [36]. In [36], extra (redundant) columns were added to

[5]In MATLAB simulations, the measured bitcell failure positions were simply randomly shuffled to mimic the behavior of a large number of dice, as an appropriate for the random variations under consideration.

[6]In the $2N$-bit subword, the total weight of the added faulty N bits is $\sum_{i=-N}^{-1} 2^i \approx 2^{N-1}$, i.e., it is approximately equal to the weight 2^{N-1} of the immediately more significant bit (i.e., the LSB in the N-bit subword). Hence, the noise contribution of the added N bits is equivalent to the noise contribution of errors in the LSB in the original N-bit subword.

[7]Energy (including bitcell, bitline and sensing energy) increases linearly as the number of bitlines switching increases. On the other hand, the energy associated with the wordline (from decoder to wordline buffers and wordline wire) increases sublinearly, since their dominant energy contribution (wordline buffers) typically increases in a logarithmic fashion [34].

TABLE I

δ V_{boost} FOR ERROR-FREE WRITE OPERATION (V_{DD}=500 MV)

SRAM size (Kb)	ΔVboost (mV)	BER
32	130	3.15e-5
64	139	1.52e-5
128	145	0.77e-5

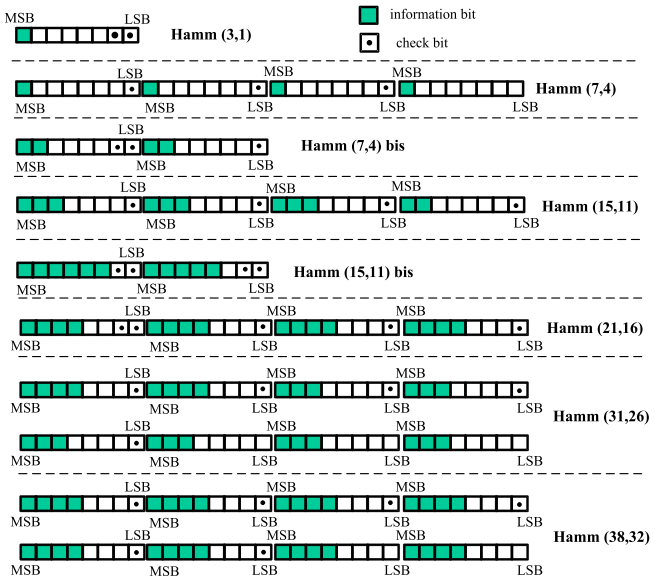

Fig. 9. Description of considered Hamming ECC schemes.

store the ECC check bits at the cost of larger area, and they were selectively activated depending on the targeted quality. On the other hand, in this paper, we explore an SECC scheme that has no redundant columns, as check bits are available from unused (dropped) LSBs, thus saving area and energy. Interestingly, such reuse of LSBs as check bits for MSBs will be shown to be substantially more energy efficient than the mere bit dropping and dual-V_{DD} schemes. Results will show that this approach is rather effective when a chip is skewed toward the read-critical process corner.

In the following, we consider the class of Single Error Correction (SEC) codes of (n, k) Hamming codes, where k indicates the number of the information bits (i.e., protected) and n is the code length (total number of information and check bits). These codes are well known for their low hardware complexity, whereas alternative codes, such as BCH and Reed–Solomon, suffer from substantially larger complexity, which makes them impractical for error-tolerant applications [37]. Table II summarizes the required number of check bits for various practical numbers of information bits in (n, k) Hamming codes. Table I shows that a reasonable energy overhead due to check bits is achieved for k in the order of tens or more. On the other hand, the low values of k and n are desirable from the quality point of view at very low voltages, since error-tolerant arrays operate in relatively high failure rate regime, as opposed to error-free memories. Hence, SEC codes with the high values of n are actually significantly more prone to failures at low voltages, due to the higher probability of multiple errors. Accordingly, the choice of the ECC code has a major impact on the energy-quality tradeoff at ultralow voltages.

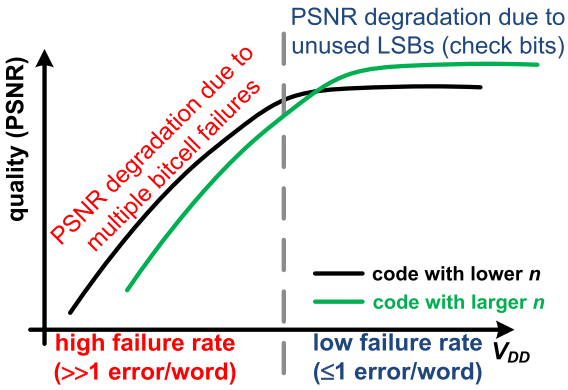

Fig. 10. Qualitative trend of PSNR versus supply voltage for SECC schemes.

TABLE II

REQUIRED NO. OF CHECK BITS VS NO. OF INFORMATION BITS

Number of information bits (k)	Number of checkbits (n-k)	Overhead due to check bits(n-k)/k
1	2	200%
2 to 4	3	75-150%
5 to 11	4	36-80%
12 to 26	5	19-42%
27 to 57	6	11-22%

The impact of the code on the energy-quality tradeoff was investigated via simulations based on the measured error map of the 32-kb memory testchip at different voltages (see the Appendix for details on measurements). The impact of the code on the energy was found to be negligible, as the energy cost of the related SECC encoder/decoder is always lower than 5% across codes and voltages, and typically even lower. Fig. 8 shows the quality versus V_{DD} for the wide range of codes listed in Fig. 9, which summarizes the operation of each code. As an example from Fig. 9, the (3,1) code individually protects each 8-bit subword (i.e., pixel), with its MSB being protected by using the last two LSBs as check bits.

More in general, the check bits can be shared among different subwords, as in the (7, 4) code where the first MSB of each group of four subwords is protected and the last LSB of three subwords acts as check bits. Fig. 8 shows a fundamental difference between very low and higher voltages. At very low voltages, the codes with a smaller n produce a higher PSNR at very low V_{DD}, whereas an opposite trend is observed for higher voltages. For example, at $V_{DD} = 0.5$ V, the (3, 1) code produces the higher PSNR among all codes in Fig. 9. In the same range of very low voltages, all other codes (i.e., with larger n) lead to worse quality due to the very high failure rate and, hence, have a higher probability of experiencing double (or higher order) errors. However, the superiority of the (3, 1) code is observed only at impractically low-quality targets (PSNR $<$18 dB). Hence, the adoption of very low n is not an option from a quality perspective.

Interestingly, the code leading to the best quality actually depends on the quality target range, and it does not necessarily have the largest n. From Fig. 8, for low to medium PSNR targets (25–40 dB), the (15, 11) Hamming code obtains the best performance. For higher PSNR targets (40–45 dB), the (31, 26) code outperforms the others in terms of quality. However, the latter code uses five check bits (the last LSB of each pixel grouped in set of five), whereas the information bits span over a group of eight pixels which implies that the eight pixels should be accessed at a time during a write/read operation (the memory word should be composed of $8 \times 8 = 64$ bits). For even higher PSNR targets (45–53 dB) and V_{DD} between 0.7 and 0.8 V, the highest quality is achieved with the (31, 26) code. Hence, in general, higher quality targets require both larger voltages and the adoption of an SECC code with progressively larger n.

Each PSNR curve in Fig. 8 has the qualitative trend of Fig. 10, which increases as V_{DD} increases, until it reaches a saturation value. At large enough voltages, the BER is exponentially reduced and each subword has at most one error, which can be corrected by the SECC code. In this low failure rate regime, the PSNR is essentially set by the effective length subword, which is reduced by the presence of the LSBs that are used as check bits instead of carrying additional information.

On the other hand, at lower voltages, the failure rate is much higher and the PSNR is dominated by the multiple errors per subwords, as they cannot be corrected by the SECC code. From Fig. 9 and Table II, the fraction of bits utilized as check bits is smaller for SECC codes with larger n, which explains why codes with larger n in Fig. 8 saturate at larger PSNR. However, from Fig. 8, the SECC codes with larger n also lead to a lower PSNR in the high failure rate regime, as shown in Fig. 9. This justifies why the quality-optimal code has a progressively higher n at larger V_{DD}.

As confirmed in Fig. 8, the codes that reserve more than one LSB as check bits have a lower saturation value of PSNR (∼43 dB) compared with the codes using only one LSB as check bit.

A. Design Considerations on Selective ECC

According to the above considerations, quality is substantially affected by both the choice of the SECC code and the operating voltage. Interestingly, the two knobs are interdependent, as the quality-optimal code actually depends on V_{DD} itself. Hence, as already observed for the other selective techniques, SECC and voltage scaling need to be co-optimized to truly minimize energy for a given quality target. However, nearly minimum energy operation can be actually achieved by adopting a single code across the entire voltage range. For example, the adoption of the single (15, 11) code across different voltages is actually a highly reasonable choice in view of the following observations.

1) It has a low number of LSBs used as check bits (only one LSB per 8-bit subword).
2) It exhibits the best PSNR in the range 25–39 dB (i.e., when the quality starts to be acceptable), and has a PSNR that is very close (within 1 dB) to the best even beyond 40 dB.

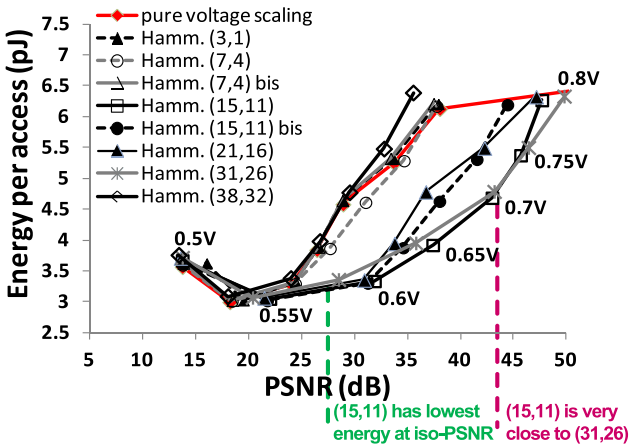

Fig. 11. Energy-PSNR tradeoff for several SECC schemes.

In addition, the (15, 11) code implementation is relatively simple, especially compared with the codes with larger n, due to the rapid complexity increase with larger n. Indeed, such codes are typically implemented as an XOR tree, whose size depends on the values of n and k, as summarized in Table III (two-input XOR is adopted as a building block). Results are obtained through automated synthesis in the adopted 28-nm technology, and the energy has been evaluated through circuit simulations at $V_{DD} = 0.55$ V, which defines the minimum energy point, as will be shown later. From Table II, the (15, 11) code has a significant lower gate count (3.45×), area (3.1×), and energy dissipation (1.31×) compared with the (31, 26) code, while assuring essentially the same PSNR for quality targets above 40 dB. Hence, the (15, 11) code represents a good compromise in terms of circuit complexity (i.e., area/energy overhead) and quality at ultralow voltage. According to the above considerations, the SECC encoder and decoder have a small impact on the total area and energy. The area (energy) cost is only 1% (0.4%) for the (15, 11) code and 4% (1.7%) even for the (38, 32) code.

Regarding the overall energy-quality tradeoff, Fig. 11 shows the array energy-PSNR tradeoff achievable by each SECC. As expected from the above considerations, the (15, 11) code is confirmed to be the most energy-efficient for PSNR ≤43 dB. For larger PSNR values, the (31, 26) code has slightly better energy efficiency, although (15, 11) achieves almost the same PSNR at isoenergy. This is because (15, 11) has essentially the same quality as (31, 26) at given V_{DD}, thus confirming that (15, 11) is the best choice across practical PSNR targets.

Summarizing, the addition of SECC introduces a negligible energy and area overhead, while making quality degradation substantially more graceful than pure voltage scaling. This ultimately justifies the above energy improvements over pure voltage scaling at a given quality.

B. Energy-Quality Tradeoff of Single SECC at Different Values of V_{DD}

In this section, the measured energy-quality tradeoff is analyzed assuming that the (15, 11) code is adopted across the

TABLE III
ENCODER/DECODER CIRCUIT COMPLEXITY VS. ECC SCHEME)

Hamming ECC scheme	Gate count (XOR2) (Enc./Dec.)	Area (equivalent XOR2 gates) (Enc./Dec.)	Energy (fJ) (Enc./Dec.)
(3,1)	0/2	3/12	1.2/11.7
(7,4)	6/9	37/51	11.4/16.4
(15,11)	19/23	114/132	40.1/48
(31,26)	70/75	372/394	126.6/135.7
(38,32)	91/97	489/516	165.8/174.6

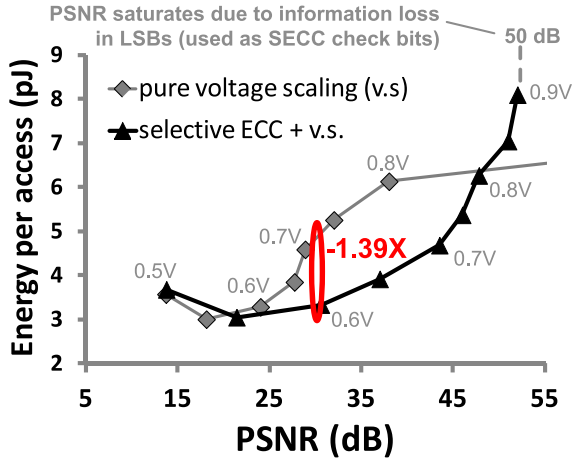

Fig. 12. Energy-PSNR tradeoff for the Hamming (15, 11) SECC technique (measurements, read-critical corner, 22 °C).

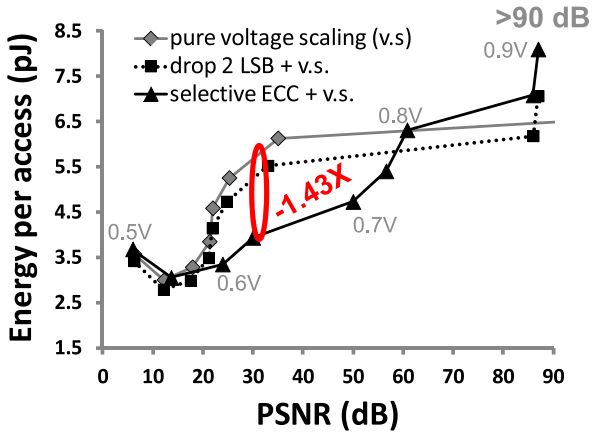

Fig. 13. Energy-PSNR trade-off for the Hamming (15,11) selective ECC technique and 16-bit sub-word (measurements, read-critical corner, 22 °C).

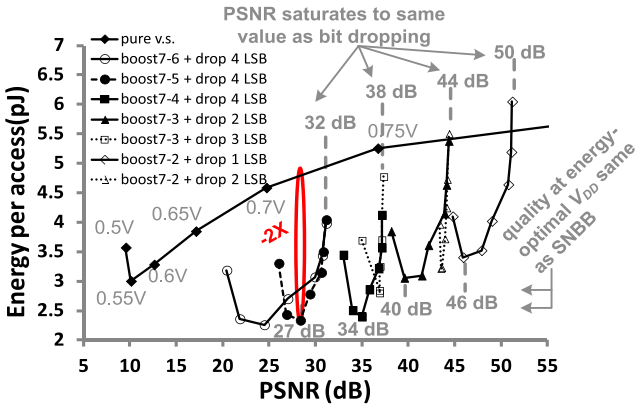

Fig. 14. Energy-PSNR trade-off for the proposed SNBB technique combined with bit dropping (measurements, write-critical corner, 22 °C).

entire considered voltage range, as pointed out in Section V-A. As expected from Fig. 9, the (15, 11) code is able to correct most of the errors occurring in the position 7-5 of each 8-bit subword. The very few remaining errors are due to multiple error events, which cannot be corrected by Hamming(15, 11), and/or to the asymmetry of the code that can protect three MSBs of the first three subwords and only two MSBs of the remaining one (see Fig. 9).

The measured energy-quality tradeoff under the Hamming(15, 11) SECC is plotted in Figs. 5 and 12 for the write- and read-critical corners. As ECC is able to correct single write and read failures, the Hamming(15, 11) SECC scheme typically produces a significant increase in the PSNR, or enables substantial voltage and energy reduction at given PSNR. For example, from Fig. 12, the SECC scheme is able to save 1.39× of total energy at PSNR =30 dB. This net saving includes the energy cost of the SECC encoder/decoder, which is negligible compared with energy dissipated by the array. Measurements indicate that the SECC encoder/decoder accounts for an increase in the energy consumption by less than 3% (2%) during a write (read) access.

At the same time, the additional SECC delay reduces the maximum frequency $f_{\max}$ by up to 4%, Such small performance penalty can be easily recovered through a very small voltage increase (∼10 mV) and, hence, at insignificant energy cost. Overall, the presence of the SECC encoder/decoder enables substantial energy reduction at an energy cost that is only a few percentage points.

As expected, the energy benefit offered by the SECC is smaller than that provided by SNBB in Section IV. Indeed, the SNBB is able to correct all errors occurring in the boosted bitlines, whereas the SECC is able to correct only a single error (see Fig. 9). As expected, the PSNR in Figs. 5 and 12 saturates due to the information loss associated with the usage of one LSB as check bit. In addition, the SECC provides a pronounced energy benefit over a pure voltage scaling for all practical PSNR targets. Instead, at very low voltages, the SECC does not provide any quality (or energy) benefit, because the failure rate becomes so high that double and higher order errors are very likely to occur (which SECC is not able to correct).

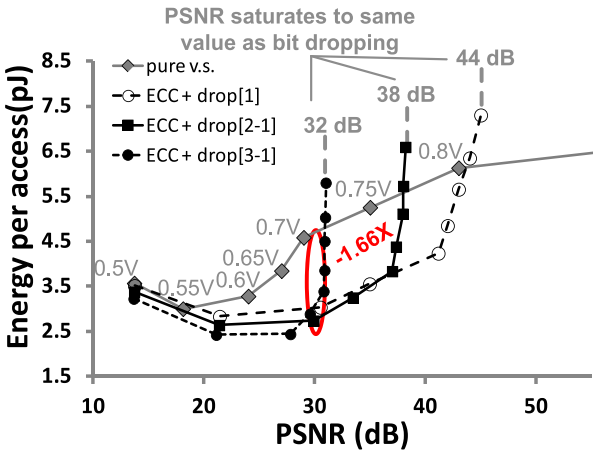

Fig. 15. Energy-PSNR trade-off for the proposed SECC technique combined with bit dropping (measurements, read-critical corner, 22 °C).

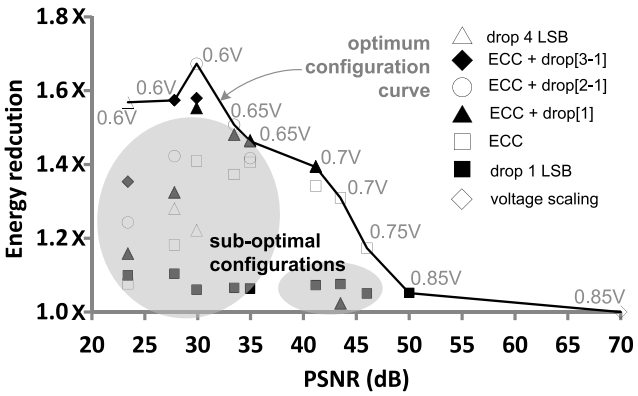

Fig. 16. Energy saving w. r. t. pure voltage scaling vs. PSNR target by dynamically selecting the energy-optimal configuration (measurements, read-critical corner, 22 °C).

C. Impact of Array Size and Subword Length

In general, the effectiveness of SECC may depend on the array size and the subword length. As discussed in Section IV-B, the quality is essentially unaffected by the array size; hence, the above results on the energy-quality tradeoff are valid in general, regardless of the array size. On the other hand, the energy benefit of SECC tends to slightly increase under longer subwords. This can be explained by considering that the quality saturates to a level that is dictated by the subword length (i.e., the corresponding precision), as discussed in Section III. When the quality target is close to the saturation value, the energy penalty tends to increase faster, due to the steeper slope of the energy-quality curve close to the quality asymptote [Fig. 4(a) and (b)]. Since longer subwords saturate at higher quality levels, for intermediate quality targets, the energy has a slower increase than shorter subwords, due to their lower steepness of the energy-quality curve. Hence, for intermediate quality targets, the energy of longer subwords tends to be smaller than shorter subwords, although the resulting energy advantage tends to be rather limited (3% when doubling the subword in the considered testchip).

As an example, Fig. 13 shows the energy-quality tradeoff with subword extended to 16 bits. The only difference compared with the previous numerical examples is that each word comprises two 16-bit subwords, instead of four 8-bit subwords. The bits of each subword have been physically interleaved as in Fig. 16 in order to reuse the same available SECC and boosting auxiliary circuits that have been originally designed for 8-bit subwords. In this way, the same Hamming(15, 11) SECC scheme (and, hence, energy/area overhead) is implemented, thus enabling a fair comparison with the (15, 11) scheme that was previously considered for four 8-bit subwords. As expected, the PSNR under the 16-bit subword saturates at a much larger value of PSNR (>90 dB), as compared with the saturation value of 50 dB under 8-bit subwords (see Figs. 5 and 12). This justifies why the energy benefit of SECC over pure voltage scaling in Fig. 13 is slightly increased to 1.43× at PSNR ∼30 dB, compared with the 1.39× achieved in Fig. 12 for 8-bit subwords. Moreover, Fig. 13 confirms the energy saving opportunity that the SECC offers by reusing the LSBs as check bits compared with the simple LSBs dropping. In the considered case, keeping the LSBs inactive entails a loss of resolution of two LSBs (one LSB for each 8-bit group, thus two LSBs for each 16-bit subword) and the obtained energy saving is only 1.09× at PSNR ∼30 dB.

VI. Synergistic Adoption of Multiple Selective Techniques for Energy-Quality Management

A. Joint Adoption of SNBB, Bit Dropping, and Voltage Scaling

When a chip is skewed toward the write-critical corner, the selective assist SNBB can be combined with bit dropping to further improve the energy-quality tradeoff, compared with individual application of these techniques. Several configurations are obtained for the possible combinations of number of boosted columns and dropped bits. For simplicity, Fig. 14 shows the energy-quality tradeoff for the most promising configurations at the write-critical corner, while omitting the least promising ones.

Regarding quality, from the comparison of Figs. 4(b) and 14, the PSNR of each configuration under joint SNBB, bit dropping, and voltage scaling saturates at the same value that was observed for the individual bit dropping and voltage scaling at the same number of dropped bits. This is explained by observing that the quality in Fig. 14 saturates to its asymptotic value for relatively large voltages ($V_{DD} \geq 0.75$ V), at which the bitcell failure rate is so small that the asymptotic quality is the same as bit dropping in Fig. 4(b). On the other hand, at lower voltages, the quality is mainly limited by bitcell failures rather than the dropped bits. Hence, the quality is essentially defined by the adopted SNBB scheme, and is independent of the number of dropped bits. This is clearly shown in Fig. 14, where the quality at the minimum energy point of each energy-quality curve is placed at the same quality as the SNBB scheme in Fig. 5 with the same number of boosted columns. Hence, once again the reduction in the number of boosted bitlines by one leads to a 6-dB PSNR degradation, similar to the individual SNBB approach (Fig. 7).

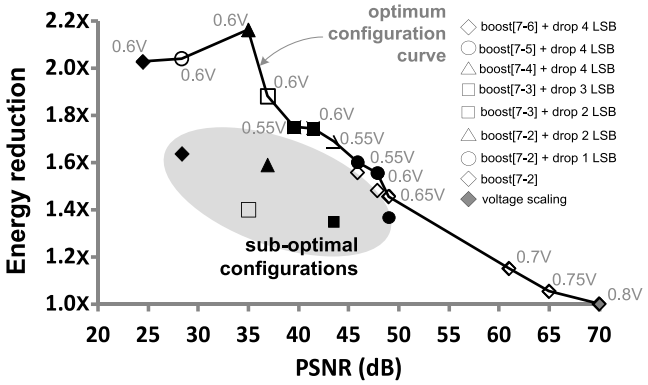

Fig. 17. Maximum energy saving for a given PSNR target tradeoff by dynamically tuning the memory configuration (measurements, write-critical corner, 22 °C).

In addition, the same 6-dB reduction also occurs for the PSNR saturation value when the number of dropped LSBs is increased by one.

In regard to the energy, the comparison of Figs. 5 and 14 shows that the joint SNBB, bit dropping, and voltage scaling provide substantially better quality than the pure voltage scaling. Equivalently, the energy efficiency at a given quality is substantially improved to up to 2× at PSNR ∼30 dB, under boost[7-5] configuration with four dropped bits. This energy saving is larger than that of individual SNBB (1.54× from Fig. 5) and the individual bit dropping with four dropped bits [1.5× from Fig. 4(b)].

The considerations on the impact of array size and subword length in Section IV-B can be repeated for the joint SNBB, bit dropping, and voltage scaling. Hence, the above results and considerations hold regardless of the array size. In regard to the subword length, once again a slightly more pronounced energy benefit is observed under longer subwords (plots are omitted for the sake of brevity).

B. Joint Adoption of SECC, Bit Dropping, and Voltage Scaling

When a chip is skewed toward the read-critical corner, SECC can be combined with bit dropping to enhance the energy efficiency at a given quality. As shown in Fig. 9, the first LSB in any subword of the considered codes is used as check bit, and hence, the joint adoption of SECC requires that only the bits starting from the second least significant position are dropped.

Fig. 15 shows the energy-quality tradeoff of joint SECC, bit dropping, and voltage scaling at the read-critical corner, as measured from the 32-kb testchip array with 8-bit subwords.

For the reasons explained in Section V-B, the adopted SECC code is Hamming(15, 11), and the number of dropped bits ranges from 1 to 3. From Fig. 15, the quality curve saturates to a level that is defined by the number of dropped bits, at relatively large voltages (e.g., 0.7 V). This level matches the saturation value that was observed for the individual bit dropping technique in Fig. 4(a). Again, this is because the failure rate at $V_{DD} \geq 0.75$ V is so small that the SECC actually does not introduce any significant quality improvement, and the quality is actually limited by the information loss due to the dropped LSBs. On the other hand, the quality at low voltages is mainly defined by the failure rate, thus SECC is able to provide significant energy/quality improvement, as shown by the comparison of Figs. 4(a) and 15.

Quantitatively, the energy improvement of the combination of SECC, bit dropping, and voltage scaling at 30-dB PSNR is 1.66×. The latter energy reduction is significantly better than the value of 1.08× obtained under individual 1-bit dropping and 1.28× under 3-bit dropping from Fig. 4(a). The energy reduction enabled by joint SECC, bit dropping, and voltage scaling is also higher than the value of 1.39× obtained with SECC and voltage scaling in Fig. 12.

More interestingly, the overall energy saving (i.e., 1.66×) is even larger than the sum of the savings achieved for each technique (i.e., 1.08× for 1-bit dropping, 1.39× for SECC). In other words, these techniques are synergistic and their appropriate combination can deliver even larger advantages than the sum of their individual improvements.

From Fig. 15, the energy difference across different numbers of dropped bits is much smaller than the difference that was observed for the individual bit dropping with voltage scaling, as shown in Fig. 4(a). This is because the quality is limited by the bitcell failures rather than the information loss in the dropped bits. Hence, dropping a different number of bits does not significantly affect the quality and, hence, the energy-quality tradeoff. For the same reasons clarified in Section V-C, the benefits of the joint SECC and bit dropping are unaffected by the array size, whereas a slight increase in the energy reduction (∼3%) has been observed when doubling the subword length (as in the case of single subword length).

VII. Energy-Optimal Combination of SNBB, SECC, Bit Dropping, and Voltage Scaling

In Section VI, the most promising combinations of SNBB, SECC, bit dropping, and voltage scaling were explored by implicitly assuming that the same single configuration was adopted across all quality targets. However, the different energy efficiencies of each combination at different quality targets can be leveraged by optimally selecting the configuration that minimizes the energy within each target range. To this aim, Figs. 16 and 17 show the maximum energy saving compared with the pure voltage scaling when the configuration is optimally selected among those considered above, under a given quality and, respectively, at read- and write-critical corners.

At the read-critical corner, Fig. 16 shows that the individual bit dropping is the configuration that exhibits the minimum energy at very low-quality targets, with an aggressively high number of dropped bits (e.g., 4). On the other hand, the pure bit dropping with a few dropped bits (e.g., 1) offers a very limited energy reduction over a very wide range of quality targets (PSNR up to 45 dB). At moderate quality targets (PSNR of 25–40 dB), joint SECC, bit dropping, and voltage scaling exhibit the best energy efficiency (1.4×–1.7× energy reduction compared with the pure voltage scaling). Compared with the very low-quality targets, a lower number of dropped bits (1 or 2) needs to be adopted in these

TABLE IV
SUMMARY OF TECHNOLOGY-INDEPENDENT RESULTS (PROCESS-DEPENDENT REPORTED IN PARENTHESIS)

		individual techniques + V_{DD} scaling			**joint techniques + V_{DD} scaling**	
		bit dropping	selective write assist @ write-critical corner	selective ECC @ read-critical corner	bit dropping + selective write assist @ write-critical corner	bit dropping + selective ECC @ read-critical corner
energy-quality-area tradeoff	PSNR (Q)	saturates to value degraded by -6·$N_{dropped}$ dB	errors confined in non-boosted bitlines	- energy-optimal N_{code} increases for larger PSNR target - single code can be used for any PSNR (~0 energy loss)	- saturates to same value as bit dropping - at min. energy point: same PSNR as selective write assist	saturates to same value as bit dropping
	energy (E)	reduced by $N_{dropped}$ times the energy of one column (33%)	E-Q curve rigidly right-shifted by 6·$N_{boosted}$ dB (energy saving: ~35%)	- significant energy reduction at iso-PSNR (28%)	- E-Q curve rigidly right-shifted by 6·$N_{boosted}$ dB - aggressive bit dropping is needed for max energy saving (~54%)	- energy reduction larger than bit dropping and selective ECC (40%) - aggressive bit dropping is needed for max energy saving (~40%)
	area (A)	-	negligible overhead (1%)	negligible overhead (1%)	negligible overhead (1%)	negligible overhead (1%)
impact of array size (*AS*)		- PSNR independent of *AS* - energy saving indep. of *AS*	- PSNR independent of *AS* - energy saving independent of *AS* (only 1.1% worse with doubled *AS*)	- PSNR essentially independent of AS (1dB max deviation, when AS is increased by 4X) - energy saving independent of *AS*	same as bit dropping and selective write assist	same as bit dropping and selective ECC
impact of sub-word size (*SWS*)		- PSNR degradation independent of *SWS* - energy saving $\propto 1/SWS$	- PSNR degraded by -3·*SWS* dB - slightly larger energy saving at larger *SWS* (few %)	slightly better energy saving at larger *SWS* at iso-PSNR (few %)	- PSNR degraded by -3·*SWS* dB - energy saving $\propto 1/SWS$	slightly better energy saving at larger *SWS* at iso-PSNR (few %)

configurations. At very high-quality targets (>40 dB), the energy benefits of joint techniques decrease rapidly. This is because these targets require a voltage that is already large enough such that the BER is relatively small, and hence, the additional techniques to mitigate errors do not provide any significant quality improvement. From Fig. 16, suboptimal configurations offer an energy benefit that is well below the energy-optimal ones, and joint configurations generally exhibit better energy efficiency than the individual techniques. Hence, in practical cases, where the quality target varies over time, the proper selection of the joint configuration (i.e., run-time energy-quality management) is essential to truly minimize the energy.

Similar considerations hold for write-critical corner, as shown in Fig. 17. Again, the energy saving obtained with joint adoption of SNBB, bit dropping, and voltage scaling is maximum at practical quality targets (PSNR ∼25–40 dB), and is in the range of 1.7×–2.24× compared with the pure voltage scaling. For the same above reasons, the energy benefit compared with the pure voltage scaling rapidly decreases for very high-quality targets. Among all configurations, the most energy-efficient at low PSNR has an aggressively high number of dropped bits (e.g., 4) and includes a moderate amount of SNBB (e.g., two columns every eight). For higher quality targets, more aggressive SNBB is needed, and the most energy-efficient configuration includes a progressively lower number of dropped bits. The energy benefit drops rapidly for configurations with a few dropped bits. Observe that the boost configurations in conjunction with the LSBs dropping are able to trade energy with quality only within a small set of PSNR.

As highlighted in Fig. 14, boosting a larger number of MSBs sets the minimum achievable PSNR value, whereas dropping more LSBs affects the maximum achievable PSNR.

From the above considerations, the appropriate choice of the combination of the above techniques can provide substantial energy benefits compared with the simple voltage scaling and even individual selective techniques, under practical PSNR requirements.

In particular, the first choice has to be made between SNBB and SECC at testing or boot time, based on whether the chip is skewed toward write- or read-critical corners.[8] Then, the chosen technique is mixed with bit dropping and voltage scaling as explained in Sections VI and VII. Results showed that the limited bit dropping (e.g., 1 or 2 dropped bits) does not provide significant energy reduction in any practical case, and should be, hence, avoided regardless of the specific corner. In other words, only aggressive bit dropping brings significant energy advantages, and only for low-quality targets.

VIII. CONCLUSION

In this paper, approximate SRAMs for error-tolerant applications have been widely explored through experimental measurements and extrapolation, targeting energy reductions that are beyond what pure voltage scaling traditionally allows. Four highly representative classes of selective (bit level) techniques to manage the energy-quality tradeoff have been considered: 1) bit dropping; 2) dual-V_{DD}; 3) assist; and 4) ECC.

[8]This can be easily done through a complete array scan.

Results showed that the minimum energy point under simple voltage scaling occurs at impractically low-quality targets, and hence, it cannot be really reached. On the other hand, the energy benefits of selective techniques under voltage scaling are well centered around practical quality targets, thus making minimum-energy operation possible. The impact of the design parameters has been discussed and justified, as is summarized in Table IV.

Bit dropping alone is always preferable to dual-V_{DD} approximate arrays, and their energy advantage is observed only at very low-quality targets. At moderate- to high-quality targets, selective assist and ECC provide the largest energy benefits ($\sim$1.5$\times$), compared with the voltage scaling. Results showed that reusing dropped bits as check bits for SECC provides larger energy benefits than keeping dropped bits inactive, as more aggressive voltage scaling is enabled at isoquality. The choice of the ECC code turned out to have a major impact on the energy-quality tradeoff at ultralow voltages. As opposed to error-free memories, SECC in approximate SRAMs with larger number of check bits can actually be more prone to failures, due to the higher probability of multiple errors. Analysis showed that the same ECC code can be used across a wide range of voltages and quality targets, while keeping the energy rather close to the very minimum, thus making ECC dynamic reconfiguration inessential. The array size was shown to have a negligible effect on the above results. On the other hand, the larger word lengths further emphasize the above advantages of selective techniques over pure voltage scaling, with the exception of bit dropping, which is largely unaffected.

The joint adoption of multiple above techniques was shown to provide much higher energy benefits ($>$2$\times$), compared with pure voltage scaling. In combinations that include bit dropping, it was shown that the maximum (asymptotic) quality is set by the number of dropped bits, whereas the quality at the minimum energy point is set by the SECC/assist technique. The energy gain was shown to be very sensitive to the adopted combination, and the qualitative guidelines were provided to identify the energy-optimal combination at different levels of quality. No single combination proved to be the best across different quality targets. This clearly shows that the dynamic management of energy-quality tradeoff that optimally selects the technique combination according to the targeted quality is mandatory in approximate SRAMs.

In summary, the technology-independent results of the above analysis are summarized in Table IV, which can be used as a tool to take preliminary design decisions in approximate SRAMs. In view of the above discussed large energy gains ($>$2$\times$) and negligible area overhead (1%) compared with the traditional voltage scaling, the synergistic adoption of multiple selective techniques and voltage scaling is expected to become a mainstream approach in approximate SRAMs.

APPENDIX

In the following, details are given on how data under different ECC codes and SRAM design parameters have been derived in Sections III–VII.

A. Reference Testchip Array, Error Map, and Energy Model

The energy-quality measurements were performed on a 32-kb SRAM testchip with four 128 $\times$ 64 sub-banks with 32-bit word, 2:1 column multiplexing, and encoder/decoder based on the Hamming(15, 11) code [12]. In this testchip, the corners were emulated by tuning the wordline voltage to match the simulated BER at the corresponding corners [12]. The wordline voltage adjustment also permits to emulate cell failures at the far-end of the tails of the distribution, which has been proven to be useful in the analysis of larger arrays [31].

To explore the quality degradation across different design scenarios, an error map has been derived through comparison of the original and the stored version, after going through a read-after-write access in the available dice. Good agreement was found between the simulated BER and PSNR and the measurements, with an average discrepancy of 3.4 dB.

To explore the energy reduction, a measurement-based SRAM energy model was built by isolating five energy contributions: 1) wordline drivers; 2) encoder/decoder; 3) bitline boosting circuit; 4) bitline drivers (precharge and data drivers); and 5) sense amplifier. The first three were directly measured through explicit supplies, whereas the others were evaluated from the difference. The MATLAB model was built to combine the above contributions and explore design scenarios different from the testchip design (e.g., different array sizes and ECC codes).

B. Analysis of Different SECC Codes

The PSNR-energy and PSNR-V_{DD} tradeoffs for different SECC codes (Figs. 8, 11, and 13) were derived from the 32-kb measured error map. The results of the Hamming(15, 11) code have been directly obtained from measurements, since the correspondent encoder/decoder was actually implemented on the testchip. The other Hamming codes have been investigated via software simulation. The encoder MATLAB models of the studied Hamming codes are fed with the image bitstream. The encoder output is then corrupted according to the measured memory error map. The corrupted data are then inputted to a MATLAB model of a corresponding decoder that fixed the errors accordingly. Finally, the resulting image is reconstructed and compared with the original one. Such a process is repeated for different values of V_{DD} and, hence, for the corresponding measured error map. For each code, the energy of the SRAM was evaluated through the model in the Appendix, whereas the contribution of the encoder/decoder circuits for codes different from Hamming(15, 11) was evaluated via through postlayout simulations (see Table IV).

C. Analysis of Different SRAM Sizes, Subword Length, Dropped LSBs, and Dual-V_{DD}

To extrapolate the energy under array sizes larger than the testchip (see Sections IV-B, V-C), the bitline energy was obtained by linearly scaling the testchip bitline energy according to the number of rows in each bank. The bitline energy increase factor due to NBB was kept fixed, as it does not depend on the bitline capacitance (see Section IV).

Regarding the impact of the subword length (see Sections III-B, VI-B, and V-C), in the example with 16-bit grayscale pixels, the 32-bit data word was split into two 16-bit subwords (instead of the four 8-bit words in the testchip). In this case, since the total word length remains the same, the energy per access is clearly the same as the testchip, assuming the same array capacity.

The impact of 1–2 dropped bits was studied through direct measurements, as the testchip is able to drop up to two LSBs [12]. For a larger number of dropped bits, the energy was evaluated by subtracting the contribution of the bitlines and senseamps that were kept inactive, according to the number of dropped bits and the model in the Appendix. Regarding the quality, the effect of dropping a bit was simply emulated by setting to zero all bits in the corresponding bit position.

The dual-V_{DD} technique in [26] (see Section III and Fig. 3) was studied through extrapolation from testchip measurements, by mixing the error maps measured at the two voltages. In detail, each image was read out at 0.5 and 0.8 V, and the two resulting corrupted images $\mathrm{IMG}_{0.5\ \mathrm{V}}$ and $\mathrm{IMG}_{0.8\ \mathrm{V}}$ were generated in software. Then, the pixel values of image $\mathrm{IMG}_{0.8\ \mathrm{V}}$ were replaced by the pixels in $\mathrm{IMG}_{0.5\ \mathrm{V}}$ in the array locations powered at 0.5 V, as required in the scheme in [26].

ACKNOWLEDGMENT

The authors would like to thank the support of STMicroelectronics for chip fabrication.

REFERENCES

[1] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," in *Proc. 18th IEEE ETS*, May 2013, pp. 1–6.

[2] K. V. Palem, "Energy aware algorithm design via probabilistic computing: From algorithms and models to Moore's law and novel (semiconductor) devices," in *Proc. Int. Conf. CASES*, 2003, pp. 113–116.

[3] K. V. Palem, "Energy aware computing through probabilistic switching: A study of limits," *IEEE Trans. Comput.*, vol. 54, no. 9, pp. 1123–1137, Sep. 2005.

[4] V. Gupta, D. Mohapatra, A. Raghunathan, and K. Roy, "Low-power digital signal processing using approximate adders," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 32, no. 1, pp. 124–137, Jan. 2013.

[5] S. Venkataramani, A. Sabne, V. Kozhikkottu, K. Roy, and A. Raghunathan, "SALSA: Systematic logic synthesis of approximate circuits," in *Proc. 49th ACM/EDAC/IEEE DAC*, Jun. 2012, pp. 796–801.

[6] A. B. Kahng, S. Kang, R. Kumar, and J. Sartori, "Designing a processor from the ground up to allow voltage/reliability tradeoffs," in *Proc. IEEE 16th Int. Symp. HPCA*, Jan. 2010, pp. 1–11.

[7] D. Shin and S. K. Gupta, "Approximate logic synthesis for error tolerant applications," in *Proc. DATE*, Mar. 2010, pp. 957–960.

[8] J. Park, J. Choi, and K. Roy, "Dynamic bit-width adaptation in DCT: An approach to trade off image quality and computation energy," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 5, pp. 787–793, May 2010.

[9] V. K. Chippa, D. Mohapatra, A. Raghunathan, K. Roy, and S. T. Chakradhar, "Scalable effort hardware design: Exploiting algorithmic resilience for energy efficiency," in *Proc. 47th DAC*, 2010, pp. 555–560.

[10] H. Esmaeilzadeh, A. Sampson, M. Ringenburg, L. Ceze, D. Grossman, and D. Burger, "Addressing dark silicon challenges with disciplined approximate computing," in *Proc. ISCA*, 2012, pp. 1–4.

[11] M. Samadi, J. Lee, D. A. Jamshidi, A. Hormati, and S. Mahlke, "SAGE: Self-tuning approximation for graphics engines," in *Proc. 46th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2013, pp. 13–24.

[12] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, and M. Alioto, "A 32 kb SRAM for error-free and error-tolerant applications with dynamic energy-quality management in 28 nm CMOS," in *ISSCC Dig. Tech. Papers*, Feb. 2014, pp. 244–245.

[13] R. Hegde and N. R. Shanbhag, "Soft digital signal processing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 6, pp. 813–823, Jun. 2001.

[14] K. Itoh, "Low-voltage scaling limitations for nano-scale CMOS LSIs," in *Proc. 9th Int. Conf. ULIS*, Mar. 2008, pp. 3–6.

[15] M. Alioto, "Ultra-low voltage VLSI circuits and systems for green computing," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 849–852, Dec. 2012.

[16] M. Yabuuchi *et al.*, "A 45 nm low-standby-power embedded SRAM with improved immunity against process and temperature variations," in *ISSCC Dig. Tech. Papers*, Feb. 2007, pp. 326–606.

[17] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 1, pp. 3–29, Jan. 2012.

[18] F. Frustaci, P. Corsonello, S. Perri, and G. Cocorullo, "Techniques for leakage energy reduction in deep submicrometer cache memories," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 14, no. 11, pp. 1238–1249, Nov. 2006.

[19] P. Corsonello, F. Frustaci, and S. Perri, "Low-leakage SRAM wordline drivers for the 28-nm UTBB FDSOI technology," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 12, pp. 3133–3137, Dec. 2015.

[20] H. Pilo *et al.*, "A 64 Mb SRAM in 22 nm SOI technology featuring fine-granularity power gating and low-energy power-supply-partition techniques for 37% leakage reduction," in *IEEE Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2013, pp. 322–323.

[21] I. J. Chang, D. Mohapatra, and K. Roy, "A priority-based 6T/8T hybrid SRAM architecture for aggressive voltage scaling in video applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 2, pp. 101–112, Feb. 2011.

[22] J. Kwon, I. J. Chang, I. Lee, H. Park, and J. Park, "Heterogeneous SRAM cell sizing for low-power H.264 applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 10, pp. 2275–2284, Oct. 2012.

[23] N. Gong, S. Jiang, A. Challapalli, S. Fernandes, and R. Sridhar, "Ultra-low voltage split-data-aware embedded SRAM for mobile video applications," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 883–887, Dec. 2012.

[24] K. Yi, S.-Y. Cheng, F. Kurdahi, and A. Eltawil, "A partial memory protection scheme for higher effective yield of embedded memory for video data," in *Proc. 13th ACSAC*, Aug. 2008, pp. 1–6.

[25] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, and M. Alioto, "SRAM for error-tolerant applications with dynamic energy-quality management in 28 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 5, pp. 1310–1323, May 2015.

[26] M. Cho, J. Schlessman, W. Wolf, and S. Mukhopadhyay, "Reconfigurable SRAM architecture with spatial voltage scaling for low power mobile multimedia applications," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 1, pp. 161–165, Jan. 2011.

[27] P. Düben *et al.*, "Opportunities for energy efficient computing: A study of inexact general purpose processors for high-performance and big-data applications," in *Proc. DATE*, 2015, pp. 764–769.

[28] A. R. Alameldeen, Z. Chishti, C. Wilkerson, W. Wu, and S.-L. Lu, "Adaptive cache design to enable reliable low-voltage operation," *IEEE Trans. Comput.*, vol. 60, no. 1, pp. 50–63, Jan. 2011.

[29] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[30] *USC-SIPI Image Database*. [Online]. Available: http://sipi.usc.edu/database/?volume=misc, accessed Oct. 5, 2014.

[31] G. Chen, M. Wieckowski, D. Kim, D. Blaauw, and D. Sylvester, "A dense 45 nm half-differential SRAM with lower minimum operating voltage," in *Proc. IEEE ISCAS*, May 2011, pp. 57–60.

[32] B. P. Lathi, *Modern Digital and Analog Communication Systems*, 3rd ed. London, U.K.: Oxford Univ. Press, 1998.

[33] N. Shibata, H. Kiya, S. Kurita, H. Okamoto, M. Tan'no, and T. Douseki, "A 0.5-V 25-MHz 1-mW 256-kb MTCMOS/SOI SRAM for solar-power-operated portable personal digital equipment—Sure write operation by using step-down negatively overdriven bitline scheme," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 728–742, Mar. 2006.

[34] N. H. E. Weste and D. M. Harris, *CMOS VLSI Design*, 4th ed. Reading, MA, USA: Addison-Wesley, 2011.
[35] M. Spica and T. M. Mak, "Do we need anything more than single bit error correction (ECC)?" in *Proc. Rec. Int. Workshop Memory Technol., Design, Test.*, Aug. 2004, pp. 111–116.
[36] I. Lee, J. Kwon, J. Park, and J. Park, "Priority based error correction code (ECC) for the embedded SRAM memories in H.264 system," *J. Signal Process. Syst.*, vol. 73, no. 2, pp. 123–136, 2013.
[37] C. W. Slayman, "Cache and memory error detection, correction, and reduction techniques for terrestrial servers and workstations," *IEEE Trans. Device Mater. Rel.*, vol. 5, no. 3, pp. 397–404, Sep. 2005.

Fabio Frustaci (S'06–M'15) received the M.S. and Ph.D. degrees in electronics engineering from the Mediterranea University of Reggio Calabria, Calabria, Italy, in 2003 and 2007, respectively.

He joined the Department of Computer Science, Modeling, Electronics and Systems Engineering, University of Calabria, Rende, Italy, in 2007, where he is currently an Assistant Professor. In 2006, he was a Visiting Research Associate with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA. From 2011 to 2013, he was a Visiting Researcher with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. He has authored over 40 papers in VLSI design. His current research interests include ultralow-power and high-performance design, variability-tolerant VLSI circuits, design techniques for emerging technologies (Quantum Cellular Automata), reconfigurable architectures (field-programmable gate array), and hardware-oriented stereovision.

Dr. Frustaci was a member of the Technical Program Committee of several conferences (the International Conference on Emerging Trends in Engineering and Technology, the International Conference on Computer Design, the International NEW Circuits and Systems, and the European Workshop on CMOS Variability).

David Blaauw (M'94–SM'07–F'12) received the B.S. degree in physics and computer science from Duke University, Durham, NC, USA, in 1986, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign, Urbana, IL, USA, in 1991.

He was with Motorola, Inc., Austin, TX, USA, where he was the Manager of the High Performance Design Technology Group. Since 2001, he has been a Faculty Member with the University of Michigan, Ann Arbor, MI, USA, where he is currently a Professor. He has authored over 450 papers and holds 40 patents. His work has focused on VLSI design with a particular emphasis on ultralow-power and high-performance design.

Dr. Blaauw was the Technical Program Chair and General Chair of the International Symposium on Low Power Electronic and Design, the Technical Program Co-Chair of the ACM/IEEE Design Automation Conference, and a member of the International Solid-State Circuits Conference Technical Program Committee.

Dennis Sylvester (S'95–M'00–SM'04–F'11) received the Ph.D. degree in electrical engineering from the University of California at Berkeley, Berkeley, CA, USA.

He co-founded Ambiq Micro, Austin, TX, USA, a fabless semiconductor company developing ultralow-power mixed-signal solutions for compact wireless devices. He has held research staff positions with the Advanced Technology Group, Synopsys, Mountain View, CA, USA, and Hewlett-Packard Laboratories, Palo Alto, CA, USA, and visiting professorships with the National University of Singapore, Singapore, and Nanyang Technological University, Singapore. He is currently a Professor of Electrical Engineering and Computer Science with the University of Michigan, Ann Arbor, MI, USA, and the Director of the Michigan Integrated Circuits Laboratory, a group of ten faculty members and over 70 graduate students. He has authored over 400 articles along with one book and several book chapters. He holds 22 U.S. patents. His current research interests include the design of millimeter-scale computing systems and energy efficient near-threshold computing.

Prof. Sylvester also serves as a Consultant and Technical Advisory Board Member for electronic design automation and semiconductor firms in his research areas.

Massimo Alioto (M'01–SM'07–F'15) received the M.Sc. and Ph.D. degrees from the University of Catania, Catania, Italy, in 1997 and 2001, respectively.

He was an Associate Professor with the University of Siena, Siena, Italy, a Visiting Scientist with Intel Labs–Circuits Research Laboratory, Hillsboro, OR, USA, in 2013, and a Visiting Professor with the University of Michigan, Ann Arbor, MI, USA, from 2011 to 2012, the Berkeley Wireless Research Center, University of California at Berkeley, Berkeley, CA, USA, from 2009 to 2011, and the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2007. He is currently an Associate Professor with the National University of Singapore, Singapore, where he is the Director of Integrated Circuits and Embedded Systems. He has authored or co-authored over 200 publications in journals (over 75, mostly in the IEEE TRANSACTIONS) and conference proceedings, and two books. His current research interests include ultralow-power VLSI circuits and self-powered systems.

Prof. Alioto was an IEEE CASS Distinguished Lecturer from 2009 to 2010, and currently serves as the Associate Editor-in-Chief of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He served as a Guest Editor of several journal special issues, and an Associate Editor of a number of IEEE and ACM journals. He was the Technical Program Chair of various conferences [the International Conference on Environmental and Computer Science (ICECS), European Workshop on CMOS Variability, the International NEW Circuits and Systems, and the International Conference on Microelectronics (ICM)], and the Track Chair of several other conferences (the International Conference on Computer Design, the International Symposium on Circuits and Systems, ICECS, VLSI-SoC, the Asia Pacific Conference on Circuits and Systems, and ICM).