SRAM for Error-Tolerant Applications with Dynamic Energy-Quality Management in 28nm CMOS

Fabio Frustaci[1], *member, IEEE*, Mahmood Khayatzadeh[2], *member, IEEE*, David Blaauw[2], *Fellow, IEEE*, Dennis Sylvester[2], *Fellow, IEEE*, Massimo Alioto[3], *IEEE Senior Member*

[1]DIMES - University of Calabria, Rende (Italy)
[2]EECS – University of Michigan, Ann Arbor (MI)
[3] ECE–National University of Singapore (Singapore)

***Abstract*—In this paper, a voltage-scaled SRAM for both error-free and error-tolerant applications is presented that dynamically manages the energy/quality trade-off based on application need. Two variation-resilient techniques, write assist and Error Correcting Code, are selectively applied to bit positions having larger impact on the overall quality, while jointly performing voltage scaling to improve overall energy efficiency. The impact of process variations, voltage and temperature on the energy-quality tradeoff is investigated. A 28nm CMOS 32kb SRAM shows 35% energy savings at iso-quality and operates at a supply 220mV below a baseline voltage-scaled SRAM, at the cost of 1.5% area penalty. The impact of the SRAM quality at the system level is evaluated by adopting a H.264 video decoder as case study.**

***Index Terms*—Error-tolerant, error-free, energy-quality tradeoff, ultra-low power processing, Near-threshold, SRAM, approximate computing, resiliency.**

I. INTRODUCTION

Voltage scaling is widely adopted to improve energy efficiency, thanks to the quadratic dependence of the dynamic energy dissipation [1], [2]. At the system level, the minimum voltage V_{min} that ensures correct operation is typically limited by the SRAM that is embedded in the system. Indeed, as V_{DD} is scaled down, resiliency issues concerning write/read margin degradation of the memory bitcells become severe due to the stronger impact of process variations [3]-[6].

In the last few years, error-tolerant design paradigms have been proposed [7]-[11], in which errors in the data computation or storage due to operation at $V_{DD} < V_{min}$ are actually acceptable, as long as they are within bounds and hence maintain adequate quality of the output signal. Such occasional errors can be tolerated in applications that involve computation that is statistical in nature (i.e., occasional errors are irrelevant), involve human perception (which is imperfect in nature) or physical signals (which are affected by noise inherently). Some examples of such error-tolerant applications are multimedia processing, "big data" processing (e.g., data analytics), web search, computer vision, machine learning, sensor fusion, augmented reality. Most of these applications have already become predominant with the advent of cloud/mobile computing [7]-[8]. For example, in multimedia video processing, the video stream quality can be tolerable even if some pixels are corrupted due to SRAM operation below V_{min} [12]-[15]. In the rest of the paper, image/video processing applications will be targeted for simplicity, although the ideas can be immediately extended to general error-tolerant applications.

In this paper, a highly flexible SRAM with dynamically adjustable energy-quality tradeoff is introduced for use in both error-free and error-tolerant applications [15]. The fundamental concept is to spend additional energy (e.g., assist) to improve the robustness of few MSBs and hence have a graceful quality degradation at low voltages. This requires the insertion of selective techniques that alter the energy-quality tradeoff at the bit level. As a result, for a given quality target, V_{DD} can be scaled more aggressively than traditional voltage scaling to reduce the overall energy, thus enabling larger energy saving. The proposed approach permits to use standard 6T bitcells designed for nominal voltage and distribute the same supply voltage to all bitcells within the array, thereby avoiding the requirement of re-developing the bitcell for lower voltages, as opposed to previous work on error-tolerant SRAMs [12]-[14], [16].

This paper investigates the bit-level optimization to minimize the overall energy for a given quality target. To show the concept, two specific bit-level approaches are considered: 1) MSB bitlines are selectively boosted to mitigate errors due to inadequate write margin, 2) the LSBs are actually used as check-bits in a selective Error Correction Code (ECC) that protects MSBs. Interestingly, the second

technique is proved to offer better energy savings than the traditional bit dropping technique, where low-order bits are simply kept inactive to linearly reduce energy [7]. Measurements on a 32-kb SRAM testchip in 28nm show an energy reduction by up to 35% at iso-quality, which adds to the reduction offered by pure voltage scaling.

This paper is organized as follows. In Section II, quality in SRAMs under voltage scaling is discussed. Section III describes the proposed selective approach to dynamically minimize the energy for a given quality target. Section IV discusses the testchip design and measurements results at nominal temperature, under a specific benchmark. Section V discusses the impact of the benchmark, temperature and reports measurements from multiple dice. As a case study, experimental results are used in Section VI to evaluate the impact of errors on the quality of a H.264 decoder. Conclusions are drawn in Section VII.

II. Quality Degradation in Aggressively Voltage-Scaled SRAMs

At voltages below V_{min}, SRAM bitcells statistically fail due to inadequate bitcell speed for the targeted frequency target (parametric failures) or inadequate read and write margin (functional failures), both due to random process variations. Parametric failures are avoided by operating within the maximum operating frequency of the array.

In functional failures, read and write margins are conflicting: the write margin is mainly set by the cell alpha ratio (ratio of access and pull-up transistor strength), whereas the read margin is set by the cell beta ratio (ratio of pull-down and access transistor strength). This results in degraded write-ability when process variations skew the corner towards SF (slow-NMOS, fast-PMOS), whereas read-ability tends to dominate the failure rate at the FS corner (fast-NMOS, slow-PMOS). Read and write margins rapidly degrade as V_{DD} scales down, and the resulting bitcell error rate (BER) increases approximately exponentially, as shown in Fig. 1 for a 32kb SRAM memory simulated in 28nm for both SF and FS corner (room temperature has been considered as typical in ultra-low power application). The same figure also reports the resulting quality of

an image or frame[1], as measured by the well-known Peak Signal-to-Noise Ratio (PSNR) metric [12]-[13] (higher values indicate better quality). Due to the ungracefully rapid degradation of quality at low voltage, there is no real tradeoff between energy and quality, as very limited reduction in V_{DD} is allowed for realistic quality targets (e.g., PSNR in the order of 30 dB or higher) [4], [17]. Accordingly, pure voltage scaling below V_{min} does not bring significant energy benefits for realistic quality targets.

To mitigate the quality degradation at low voltages, some recent work has exploited the different impact on quality of the errors occurring in different bit positions. As an example, Fig. 2 shows the quality (PSNR) in the above array when errors are selectively injected in a single bit position. From this figure, errors degrade quality much more strongly when they occur in MSBs rather than LSBs, as generally true for video processing algorithm. Based on this observation, the work in [13], [14] lowers V_{DD} only for the LSBs to preserve quality on MSBs. However, the rapid BER degradation at LSBs makes the energy reduction very limited. In other words, to achieve appreciable energy benefits, V_{DD} needs to be scaled so much that most of LSBs are essentially wrong. In that case, better energy reduction would be achieved by simply dropping those bits, instead of retaining them as errors. In addition, pronounced voltage differences across different bit positions pose performance issues, as their access time becomes significantly different.

In [12], MSBs (LSBs) are stored in 8T (6T) bitcells, leveraging the stronger robustness of 8T bitcells at low voltages (i.e., $V_{min,8T}$ is lower than $V_{min,6T}$). In this approach, the energy-quality tradeoff is not adjustable, since it is set at design time by the capacity of the 8T and 6T banks. Also, there is no real energy-quality tradeoff when scaling V_{DD} below $V_{min,6T}$, since the rapid degradation of LSBs make them fail in most cases. Once again, better energy reduction would be achieved by simply dropping those bits. Similarly, in [16] only 6T cells are employed but the cells storing the MSBs are oversized to reduce the BER at low voltages. Once again, the energy-quality tradeoff is set at design time, and no real tradeoff is observed at V_{DD} below the voltage V_{min} of LSBs. Moreover, equal energy per access is consumed at any bit

[1] The memory is supposed to store a grayscale 128x128 image (the image is divided into 4 slices so the memory is written and read four times).

position in [12] and [16], thus missing the opportunity to further reduce the energy by tolerating some limited quality degradation due to occasional failures in LSBs.

In the following section, we introduce a novel approach that permits more favorable and dynamic energy-quality tradeoff, thanks to more graceful quality degradation at low V_{DD}.

III. Enabling True Energy-Quality Tradeoff and Dynamic Adjustment

As discussed in the previous section, MSBs have a stronger impact on quality compared to LSBs. This suggests the idea that the quality degradation at low voltages can be made more graceful by introducing selective bit-level circuit techniques (e.g., read/write assist) that protect only few MSBs. This technique substantially improves the quality while keeping the extra energy cost small, since it is limited to few bit positions. As a result of the more graceful degradation, more aggressive voltage scaling is possible compared to pure voltage scaling, thereby enabling more substantial energy reduction than the latter.

To dynamically track different quality targets, the number of bit positions where extra energy is spent to reduce the bit error rate needs to be flexibly adjusted. Then, for a given quality target, such extra energy needs to be optimally allocated among bit positions to minimize the overall energy. This mechanism permits to cover a wide range of quality targets including error-free applications, in which extra energy is uniformly distributed across all the bit positions.

In this paper, we consider two possible selective techniques that enable selective bit-level enhancement of robustness: the selective Negative Bitline boosting (NBL) and the selective Error Correction Code (ECC). The proposed concept can be generalized to any bit-level selective technique that enhances the bitcell robustness on a column basis.

A. Selective Negative Bitline Boosting

As discussed above, the BER in dice close to the SF corner is mainly limited by write errors. Negative Bitline boosting (NBL) has been used to improve bitcell write-ability through bitline precharge at the slightly negative voltage $-\Delta V_{boost}$ to write a stronger "0" [20], [21]. This robustness enhancement comes

at the cost of larger bitline energy due to the larger voltage swing. Fig. 3 depicts the simulated BER versus the amount of negative boosting voltage ($-\Delta V_{boost}$) in a 28-nm 32kb SRAM array at the SF corner under NBL. From this figure, more negative bitline boosting improves BER (i.e., quality) at quadratic energy cost, due to the increase of bitline voltage swing by ΔV_{boost}. It might be noted that NBL may upset unselected cells within the selected (BL boosted) column and unselected rows, if ΔV_{boost} is large enough to turn their access transistor on. However, practical values of ΔV_{boost} needed to suppress write errors are certainly smaller than the access transistor threshold voltage, hence such issue is never observable in realistic operating conditions and designs. This is clearly shown in Fig. 3, as values of ΔV_{boost} larger than 200mV are never needed in practical cases.

In our approach, NBL was applied non-uniformly by boosting only the columns associated with the MSBs. This bit-level knob permits to limit the extra energy cost of NBL to the most important portion of the word, preserving the BER only where needed. From a design point of view, the voltage ΔV_{boost} is set to achieve a targeted BER at the bit level, whereas the number of columns with boosted bitline defines the overall quality (more MSBs need to be boosted for higher quality targets). For simplicity, we adopted a single voltage ΔV_{boost} for all columns, and the number of columns with NBL was fully adjustable at run time according to the scheme in Fig. 4. From this figure, NBL is enabled in columns whose *boost* signal is high, which sets the low bitline voltage to $-\Delta V_{boost}$ instead of ground through transistor M1. The *boost* signal entails the overhead of only one latch[2] every four columns (assuming that a word contains four pixels) and two additional transistors per column (M1-M2). The boosting configuration (i.e., which positions are boosted) is adjusted on the fly by writing on the *boost* register. In error-free mode, NBL is enabled at all columns, while in the error-tolerant mode it is enabled only in columns associated with an appropriate number of MSBs, as will be discussed in Section IV.

The ability to adjust the number of columns with NBL permits to achieve more graceful quality degradation at low V_{DD}, while enabling the capability to optimally allocate the extra energy for NBL under

[2] Such relatively small area overhead can be further reduced by storing the configuration in an additional SRAM array row with output hardwired to transistors M1-M2.

a given quality target. As shown in Fig. 5 for a 28nm 32kb SRAM test-chip for different NBL configurations ($-\Delta V_{boost} = -130$mV is provided off-chip), the write energy increases by a factor of up to 1.9X when applying NBL to a progressively larger number of columns. Accordingly, our selective NBL approach permits to considerably reduce the additional energy of NBL when lower quality is targeted (i.e., when less columns are boosted). Such tradeoff will be explicitly explored in Section V.

B. Selective Error Correction Code

As write errors were addressed by the selective NBL approach in Subsection A, we introduce another method that can also address read errors (especially for dice close to the FS corner, as discussed above). A recent technique that has been proposed for low-quality targets is to drop the LSBs to save their switching energy, at the cost of reduced arithmetic precision. In this technique a linear energy saving is achieved when the number of used columns is progressively reduced. Instead, we propose to use such dropped bits as check bits of a selective Error-Correction Code (ECC) that protects only MSBs, as opposed to traditional ECC schemes that equally protect all bit positions with extra check bits [22]. Intuitively, strengthening MSBs through the unused LSBs makes the quality degradation more graceful and permits to down-scale voltage more aggressively with quadratic energy benefit. In Section V, the energy benefits of selective ECC will be quantified through measurements.

Fig. 6 depicts the selective ECC scheme that was adopted in this work as a representative example. In each 32-bit word *D[31:0]* (four 8-bit pixels), a single-error-detection error-correction Hamming (15,11) code was adopted with 4 check bits and 11 protected bits. In particular, the MSBs of pixels[3] were protected by the fours LSBs of the four pixels *D[0]*, *D[8]*, *D[16]*, and *D[24]*. The selective ECC scheme is dynamically enabled by signal *ECC_sel*. During a write, the check-bits and the bits to be protected (15 bits in total) of the input word *D[31:0]* are fed to the ECC Encoder. During a read operation, the check-bits and the protected bits of the read word *S[31:0]* are inputted to the ECC Decoder and the final output *Out[31:0]* is reconstructed. Fig. 6 shows an example where a read error occurs at bit *D[23]* under the proposed

[3] As a (15,11) Hamming code was adopted, three MSBs were protected in pixel 0...2, and two in pixel 3.

selective ECC scheme. The proposed technique entails the insertion of the simple logic implementing the Hamming code (24 XOR gates) without requiring any memory array modification, as opposed to traditional ECC that is based on the insertion of redundant columns [22]. The proposed technique can also be jointly adopted with a traditional ECC code, adding further protection against failures. Compared to the herein adopted Hamming(15,11) code, more complex codes may be also used to achieve more effective protection at the expense of higher complexity. Our preliminary analysis revealed that the simple Hamming (15,11) code is a reasonably good compromise between the range of achievable quality PSNR (30dB or more in real applications) and complexity.

The resulting architecture incorporating both selective NBL and ECC is shown in Fig. 7, which includes the precharge scheme in Fig. 4 for each bitline, an ECC Hamming (15,11) Encoder and Decoder for selective ECC. To enable the comparison with traditional bit dropping, this feature was also included in the array. As in Fig. 7, bit dropping is implemented in the bitline precharge circuit by adding a *drop* signal, which disables the precharge circuit and connects the bitline pair to ground, thus saving dynamic energy. Table I summarizes the advantages of the proposed selective techniques over prior art [7], [12], [14], [16], [22].

IV. Test-Chip Design and Measurement Results

The concepts described in the previous sections were implemented in a 28-nm 32-kb SRAM testchip, whose micrograph is shown in Fig. 8 and the main information are reported in Table II. The memory array is divided into four banks (each with 128 rows x 64 columns), and a 2:1 column multiplexing is adopted. The selective ECC Encoder and Decoder consist of a tree of 2-input XOR logic and the related area penalty is only 1.2%. Including the selective NBL scheme, the total overhead associated with the proposed techniques is 1.5%. A digitally tunable pulse generator produces the internal timing signals to enable wordline, precharge and sensing. The on-chip testing harness includes the generation of input address and data patterns, the at-speed acquisition of errors occurring in bitcells, and an interface to upload settings and

download error data. A standard high-density 6T bitcell was adopted, and the negative bitline voltage under NBL was set to $-\Delta V_{boost}$=-130mV to ensure write-ability within approximately 5 standard deviations, as appropriate for the array size (Table III summarizes the yield versus ΔV_{boost}).

To assess the proposed techniques, a 128x128 8-bit grayscale image (*peppers* testbench [24]) was divided into four slices (64 x 64 bit) and stored in the memory. Then, the image was read out from the memory to detect bitcell failures through comparison with the original image. The worst-case corner for write (read) errors is emulated by tuning the wordline underboosting (overboosting) voltage. This permits to study the impact of random variations at different corners [23]. The amount of wordline under/overboosting was set to make the measured failure rate equal to the value expected from Monte Carlo simulations at the targeted corner. For example, a 100-mV (110-mV) wordline voltage decrease (increase) was needed to emulate the SF (FS) process corner, compared to the supply voltage V_{DD}=0.5V. During the SRAM operation, the same supply voltage was clearly applied for both write and read operation. For a given quality target, the lower bound to supply voltage scaling is set by either the write or the read failure rate, depending on which dominates at the considered corner.

The measured maximum frequency vs. V_{DD} is plotted in Fig. 9, which has a relatively linear trend within the 0.5-0.9 V range, as transistors operate above threshold. The measured energy-quality tradeoff when sweeping V_{DD} from 0.5 V to 0.8 V with a 50-mV step is plotted in Fig. 10 for various configurations, under the write-critical SF corner. The configurations in this figure include pure voltage scaling (where no additional NBL energy is spent), selective NBL applied to different groups of bits (7...6, 7...4, ..., 7...0) along with voltage scaling, and selective ECC with voltage scaling. Under pure voltage scaling, the quality degrades very rapidly and becomes unacceptable below 0.73 V (assuming a typical lower bound of 30 dB, as shown in Fig. 11), hence the minimum energy at 0.55 V is not really achievable under realistic quality targets. On the other hand, the uniform insertion of NBL boosting in all bit positions permits to achieve error-free operation at $V_{DD} = 0.55\,V$ (i.e., no bitcell fails, and hence PSNR is infinite), but no real tradeoff is achieved between energy and quality and the energy is 33% higher than pure voltage scaling.

Introducing the proposed selective NBL technique, the extra energy for write assist can be adjusted and traded off for quality. For example, selective NBL on the first 4 MSBs (boost[7-4]) permits to achieve the same quality of as pure voltage scaling at 0.73 V, while reducing energy by up to 35% and enabling operation at the much lower voltage of 0.55V. Conversely, boost[7-4] permits to achieve approximately the same minimum energy as the pure voltage scaling, while achieving a dramatic quality improvement (20 dB). Flexibility in the energy-quality tradeoff is offered by the availability of other selective NBL schemes: boost[7-6] has the minimum advantage over pure voltage scaling (24%) due to its worse quality (PSNR=25dB), while boost[7-2] has a 33% energy advantage due to the larger number of boosted bitlines and better quality (PSNR=46dB). From Fig. 10, such advantage is consistently obtained within the range of practical PSNRs of 30 dB or greater, as qualitatively shown in the sample images in Fig. 11.

In practical cases, a given quality target defines a corresponding energy-optimal configuration, as shown in Fig. 12. This figure shows that the proposed selective NBL scheme enables a further energy saving that is 28% on average and up to 35% for practical values of PSNR (30 dB or greater), when compared to pure voltage scaling. Interestingly, the proposed approach also reduces energy by 18% compared to the error-free case (boost[7-0]), confirming that selective NBL saves energy when compared to a uniform approach. The energy-quality tradeoff is plotted in Fig. 13, showing that the energy-optimal configuration has progressively larger number of boosted bitlines (from [7-6] to [7-2]) for increasing PSNR targets.

Selective ECC is more effective in the read-critical corner, where selective NBL is actually ineffective since write errors are much more infrequent than read errors. Fig. 14 shows the measured energy-quality tradeoff for the same test chip for a wordline tuned to emulate the read critical corner (FS). In this case, the proposed selective ECC technique reduces energy by 28% at iso-quality, compared to pure voltage scaling at 0.7 V. From the same figure, dropping one LSB offers limited energy reduction compared to pure voltage scaling at iso-voltage. Interestingly, reusing the LSB to protect the MSBs as described in Section III.B reduces energy by 19% compared to dropping the LSB under the same quality target. This is because

the quality improvement enabled by the otherwise unused LSBs can be again traded off for lower energy, thereby enabling a quadratic energy saving, as opposed to the linear energy reduction offered by traditional bit dropping. It is worth noting that PSNR saturates to a maximum PSNR of about 50dB, because of the reduced precision due to the unused LSB. The resulting energy saving of the selective ECC over pure voltage scaling is plotted in Fig. 15, which shows an average 23% energy reduction for practical values of PSNR≥30dB.

As expected, the selective ECC is not effective at very low $V_{DD} < 0.55\,V$ because the adopted Hamming (15,11) code cannot correct multiple errors, which are very likely at such low voltages. For V_{DD}≥0.55V, multiple failures in the coded bits are less likely and the ECC becomes effective, as shown by the measured error rate versus bit position in Table IV. As shown in this table, the number of errors on the three MSBs (bits 7-5) is drastically reduced. As an example, at V_{DD}=0.6 V, the error rate in the MSB (bit 7) of each pixel is reduced from 1.1% to 0.2%, i.e. by 84%. Similarly, the number of errors at bits 6 and 5 are respectively reduced by 87% and 66%. The error rate improvement is slightly lower for bit 5 due to the asymmetry of the protected bits (the Hamming code allows to protect only 2 MSBs of pixel 3, while protecting 3 MSBs of all other pixels). This also explains why some errors still occur at bit position 5 at V_{DD}=0.7 V. Fig. 16 shows the energy overhead of the selective ECC encoder (during write) and decoder (during read) versusV_{DD}, which is confirmed to be negligible (lower than 3%).

Interestingly, the proposed selective approach enables significantly more aggressive voltage scaling at given quality compared to pure voltage scaling, as shown in Figs. 17a-b for the write- and read-critical corner. From these figures, the proposed selective NBL (ECC) at write-critical (read-critical) corner reduce the minimum voltage V_{min} that ensures a given quality by 220 mV (100 mV), when targeting a PSNR of 30dB. This enhanced voltage scalability can be leveraged to fill the V_{min} voltage gap between logic (which have lower V_{min}) and SRAM arrays in aggressively voltage scaled systems, and hence adopt the same voltage for both logic and memory.

Finally, it is worth noting that the above energy savings and voltage scalability are further improved in

arrays larger than the considered 32-kb array. For example, larger array capacity requires more aggressive boosting in NBL (i.e., more negative -ΔV_{boost}) to ensure correct operation (see Fig. 3). Hence, the energy benefit of selectively limiting NBL to a few bit positions becomes even more advantageous. Also, the overhead associated with the selective ECC encoder/decoder becomes an even smaller fraction of the overall area/energy.

V. Results across Benchmarks, Multiple Dice and Temperature

The above reported benefits were found to be highly consistent across different benchmark images taken from [24]. In Fig. 18a (b), the measured PSNR is reported under selective NBL (ECC) for the write-critical (read-critical) corner with the voltage being set to 0.55V (0.6V), to achieve a PSNR of approximately 32 dB. From this figure, the image quality is highly consistent across benchmarks, with a maximum PSNR difference being only 0.5dB between the *lena* and *baboon* benchmarks. As a result, the same 220-mV (100-mV) improvement in voltage scalability has been measured across all the considered benchmarks for the write-critical (read-critical) corner.

Measurements were repeated in 19 dice, adopting the same wordline voltage tuning used to emulate write critical and read critical corners, and the same amount of negative boosting (-ΔV_{boost}=-130mV). As shown in Figs. 19a-b, the above benefits are approximately obtained for all tested dice. The average measured energy saving (voltage scalability improvement) for a PSNR=30dB is 27.2% (-198mV) at write-critical corner. Such energy (V_{min}) improvement becomes 26.6% (-109mV) under read-critical corner.

The temperature increase impacts the write and read bitcell failures in opposite ways, as it influences PMOS and NMOS transistors (and hence transistor strength ratio, which defines the margins) in different ways. Simulations show that write (read) margin increases (decreases) as temperature increases. This agrees with measurements, which showed that higher temperatures lead to a smaller (larger) number of write (read) failures. Fig. 20a depicts the energy-quality tradeoff for the write-critical corner at T=80°C, and shows that at V_{DD}=0.7V the PSNR is still acceptable (30dB) without applying any NBL, as opposed to

the lower PSNR=25dB at T=22°C). This makes the selective NBL technique less effective than room temperature, as reported in Table V.

Conversely, as temperature increases, the number of read failures increases and the selective ECC is able to mitigate them starting at higher voltage, as compared to the case at room temperature (see Table IV). This is because read failures are more likely to occur at higher voltage, due to the more significant read margin degradation at higher temperature. Indeed, at V_{DD}=0.7V the selective ECC is able to suppress all errors at T=22°C, whereas some failures occur at 80 ^{o}C. However, the benefit at lower voltages is more limited. Fig. 20b plots the resulting measured energy-quality tradeoff for the read-critical corner at T=80°C, confirming that the selective ECC techniques still provides significant benefits even at high temperatures.

The enhancement in voltage scalability measured at 80°C is shown in Figs. 21a-b. From this figure, the supply voltage can be reduced by 160mV (71mV) at iso-quality, compared to pure voltage scaling for the write- (read-) critical corner. Compared to Fig. 17, higher temperatures clearly reduce the benefits in terms of voltage scalability, due to the increased write margin. At the same time, the energy benefit is reduced to 7.9% (11.1%) for the write-(read-) critical case, since leakage is not affected by the above techniques and is responsible for a larger fraction of the total energy (see the measured energy breakdown in Fig. 22). In typical mobile applications, such significant benefit degradation is not observed, as the operating temperatures are certainly much lower than 80°C.

VI. A Case Study: H.264 Video Decoder

In this section we extend the analysis to a H.264 video decoder, as representative example of a complete system employing the above considered SRAM array as a component. Fig. 23 depicts the architecture, which is partitioned into the following fundamental tasks: entropy decoding, dequantization, inverse DCT and motion compensation (MC). In the MC block, the current frame is reconstructed from the previous frame, which is stored in the on-chip SRAM array (after being transferred from the external SDRAM). The size of the on-chip SRAM required to store a complete QCIF frame is 198 kb (176 × 144

pixels per frame). In state-of-the-art low-power video decoder, the SRAM array is actually much smaller (32 kb or less) to substantially reduce its energy per access while still meeting the required throughput [25]. Accordingly, 32-kb frame macro blocks are downloaded from the off-chip SDRAM to the on-chip SRAM.

The architecture in Fig. 23 was modeled in Matlab [26], and the SRAM bitcell failures were injected according to the measured error map of our test chip under a given condition (voltage, temperature). Through the error-tolerant SRAM, bit errors degrade the quality of the incoming frame, and hence affect the subsequent frame through MC. Fig. 24 shows the resulting output quality of the decoder versus voltage when applying the selective NBL technique on bits 7...4. The values refer to the average PSNR of the first 20 frames of the QCIF video benchmark *football* in the YUV format [27] (both inter- and intra- frame errors are considered). From Figs. 17a and 24a, the trend of quality is similar at both the output of the SRAM and the decoder. Analysis for other benchmarks showed that results are largely independent of the specific video stream.

Under selective ECC and read-critical corner, the quality trend versus voltage is reported in Fig. 25. As observed in Fig. 25a, the PSNR achievable with selective ECC saturates (PSNR~30dB for V_{DD} ≥0.8V) due to the preliminary image compression[4]. It is worth noting that values of PSNR beyond 50-60 dB are not of practical interest, since the frames go through a lossy compression in H.264 encoding, hence their quality is already degraded compared to the original frame. In detail, the PSNR at the decoder output saturates for large values of V_{DD} that prevent errors from occurring. Indeed atV_{DD}=0.85V, the test chip is read-failure free but the PSNR is about 34dB.

VII. Conclusions

In this paper, we presented an SRAM array for error-tolerant applications whose energy-quality tradeoff can be adjusted dynamically over a wide range of quality targets (including error-free operation),

[4] MPEG compression degrades the PSNR to 34 dB compared to the original frame, even ignoring the memory failures at low voltage. Without compression and still ignoring the memory failures, the unused LSB leads to a PSNR saturation at 45dB (see Fig. 17b). Hence, for V_{DD} ≥0.8 V the PSNR is mainly limited by the inaccuracies introduced by the compression, rather than bit drop dropping.

thanks to the graceful quality degradation that was obtained at low voltages through selective (bit-level) techniques. Indeed, the impact of errors at different bit positions is explicitly considered, and extra energy is spent to protect MSBs to enable more aggressive scaling throughout the array, thus enabling further reducing voltage/energy reduction compared to pure voltage scaling. Among other possible bit-level selective techniques, we introduced two techniques to demonstrate the concept: the selective Negative-Bitline Boosting (NBL) and the selective Error Correction Coding (ECC) to address bitcell failures at low voltage for both write- and read-critical corners. NBL is used to dynamically limit the number of boosted columns according to the targeted image/video quality. The selective ECC reuses the LSBs as check bits for the MSBs, providing significantly better energy efficiency compared to simple LSB dropping.

A 28nm CMOS 32kb SRAM test-chip exhibited 35% energy savings at iso-quality operating at a supply up to 220mV below a baseline voltage-scaled SRAM with less than 2% area penalty. Such advantages were found to be consistent across benchmarks and different tested dice. An H.264 video decoder was adopted as case study to show that the results on the SRAM as a component are highly representative of those at system level. Thus, the proposed approach permits to minimize the energy of SRAM for a given (dynamic) quality target, with benefits being largely independent of operating conditions.

ACKNOWLEDGEMENT

The authors kindly acknowledge the support of STMicroelectronics for chip fabrication. This work was partially supported by the grants from the Singaporean Ministry of Education MOE2014-T2-1-161 (“Error-tolerant VLSI fabrics with dynamic energy/quality for minimum energy”) and AcRF (”Sub-Cycle Error Correction for Resilient Ultra-Low Voltage VLSI Processing”, grant RG00003061), and the NSF Variability Expedition.

REFERENCES

[1] M.E. Sinangil, N. Verma, A.P. Chandrakasan "A Reconfigurable 8T Ultra-Dynamic Voltage Scalable (U-DVS) SRAM in 65nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 11, pp. 3163-3173, Nov. 2009.

[2] M. Alioto, Guest Editorial for the Special Issue on Ultra-Low-Voltage VLSI Circuits and Systems for Green Computing, *IEEE Trans. on Circuits and Systems – part II*, vol. 59, no. 12, pp. 849-852, Dec. 2012.

[3] K. Nii, Y. Tsukamoto, S. Ohbayashi, S. Imaoka, H. Makino, Y. Yamagami, S. Ishikura, T. Terano, T. Oashi, K. Hashimoto, A. Sebe, S. Okazaki, K. Satomi, H. Akamatsu, H. Shinohara, "A 45nm Low-Standby-Power Embedded SRAM with Improved Immunity Against Process and Temperature Variations," in *ISSCC Dig. Tech. Papers*, 2007, pp. 326-327.

[4] J. Chang, Y.-H. Chen, H. Cheng, W.-M. Chan, H.-J. Liao, Q. Li, S. Chang, S. Natarajan, R. Lee, P.-W. Wang, S.-S. Lin, C.-C. Wu, K.-L. Cheng, M. Cao, G. H. Chang, "A 20nm 112Mb SRAM in High-k Metal-Gate with Assist Circuitry for Low-Leakage and Low-VMIN Applications," in *ISSCC Dig. Tech. Papers*, 2013, pp. 316-317.

[5] K.A. Bowman, J. W. Tschanz, S. L. Lu, P.S. Aseron, M. M. Khellah, A. Raychowdhury, B. M. Geuskens, C. Tokunaga, C. B. Wilkerson, T. Karnik, V. K. De, "A 45 nm Resilient Microprocessor Core for Dynamic Variation Tolerance," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 194-208, Jan. 2011.

[6] M. Alioto, "Ultra-Low Power VLSI Circuit Design Demystified and Explained: A Tutorial," *IEEE Trans. on Circuits and Systems – part I* (invited), vol. 59, no. 1, pp. 3-29, Jan. 2012.

[7] H. Kaul, M. Anders, S. Mathew, S. Hsu, A. Agarwal, F. Sheikh, R. Krishnamurthy, S. Borkar, "A 1.45GHz 52-to-162 GFLOPS/W Variable-Precision Floating-Point Fused Multiply-Add Unit with Certainty Tracking in 32nm CMOS," in *ISSCC Dig. Tech. Papers*, 2013, pp. 182-184.

[8] J. Han, M. Orshansky, "Approximate Computing: An Emerging Paradigm for Energy-Efficient Design," in Proc. of *IEEE ETS'13*, Avignon (France), May 2013, pp.1-6.

[9] H. Esmaeilzadeh, A. Sampson, M. Ringenburg, L. Ceze, D. Grossman, D. Burger, "Addressing Dark Silicon Challenges with Disciplined Approximate Computing", in Proc. of *Dark Silicon 2012* (co-located with ISCA 2012), Portland (USA), June 2012, pp.1-2.

[10] D. Mohapatra, G. Karakonstantis, K. Roy, "Significance driven computation: a voltage-scalable, variation-aware, quality-tuning motion estimator," in Proc. of *ISLPED 2009*, San Francisco (USA), Aug. 2009, pp. 195-200.

[11] V. Chippa, A. Raghunathan, K. Roy, S. Chakradhar, "Dynamic effort scaling: Managing the quality-efficiency tradeoff," in Proc. of *DAC 2011*, New York (NJ), June 2011, pp. 603-608.

[12] I. J. Chang, D. Mohapatra, K. Roy, "A Priority-Based 6T/8T Hybrid SRAM Architecture for Aggressive Voltage Scaling in Video Applications," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 21, no. 2, 2011.

[13] K. Yi, S.-Y. Cheng, F. Kurdahi, A. Eltawil, "A Partial Memory Protection Scheme for Higher Effective Yield of Embedded Memory for Video Data," in Proc. of *ACSAC 2008*, Hsinchu (Taiwan), Aug. 2008, pp. 1-6.

[14] M. Cho, J. Schlessman, W. Wolf, S. Mukhopadhyay, "Reconfigurable SRAM Architecture With Spatial Voltage Scaling for Low Power Mobile Multimedia Applications," *IEEE Trans. on VLSI Systems*, vol. 19, no. 1, pp. 161-165, Jan. 2011.

[15] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, M. Alioto, "A 32kb SRAM for Error-Free and Error-Tolerant Applications with Dynamic Energy-Quality Management in 28nm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, 2014, pp. 24-25.

[16] J. Kwon, I. J. Chang, I. Lee, H. Park, and J. Park, "Heterogeneous SRAM cell sizing for low-power H.264 applications," *IEEE Trans. on Circuits and Systems – part I*, vol. 59, no. 10, Oct. 2012.

[17] A Raychowdhury, B. M. Geuskens, K. A. Bowman, J. W. Tschanz, S. L. Lu, T. Karnik, M. M. Khellah, V. K. De. "Tunable Replica Bits for Dynamic Variation Tolerance in 8T SRAM Arrays" *IEEE Journal of Solid-State Circuits*, vol. 46, no. 4, pp. 3163-3173, Apr. 2011.

[18] M. E. Sinangil, A. P. Chandrakasan "An SRAM using output prediction to reduce BL-switching activity and statistically-gated SA for up to 1.9× reduction in energy/access" in *ISSCC Dig. Tech. Papers*, 2013, pp. 318-319.

[19] N. Gong S. Jiang, A. Challapalli, S. Fernandes, R. Sridhar, "Ultra-Low Voltage Split-Data-Aware Embedded SRAM for Mobile Video Applications", *IEEE Trans. on Circuits and Systems – part II*, vol. 59, no. 12, pp. 883-887, Dec. 2012.

[20] N. Shibata, H. Kiya, S. Kurita, H. Okamoto, M. Tan'no, T. Douseki, "A 0.5-V 25-MHz 1-mW 256-kb MTCMOS/SOI SRAM for Solar-Power-Operated Portable Personal Digital Equipment," *IEEE Journal of Solid-State Circuit*s, vol. 41, no. 3, pp. 728-742, Mar. 2006.

[21] S. Mukhopadhyay, R. M. Rao, J.-J. Kim, C.-T. Chuang, "SRAM Write-Ability Improvement With Transient Negative Bit-Line Voltage," *IEEE Trans. on VLSI Systems*, vol. 19, no. 1, pp. 24-32, Jan. 2010.

[22] S. M. Jahinuzzaman, J. S. Shah, D. J. Rennie, M. Sachdev, "Design and Analysis of A 5.3-pJ 64-kb Gated Ground SRAM With Multiword ECC", *IEEE Journal of Solid-State Circuit*s, vol. 44, no. 9, pp. 2543-2553, Sept. 2009.

[23] G. Chen, M. Wieckowski, D. Kim, D. Blaauw, D. Sylvester, "A dense 45nm half-differential SRAM with lower minimum operating voltage," in Proc. of *ISCAS*, Rio de Janeiro (Brazil), May 2011, pp. 57-60.

[24] USC-SIPI Image Database, available at *http://sipi.usc.edu/database/?volume=misc*.

[25] Tsu-Ming Liu, Ting-An Lin, Sheng-Zen Wang, Wen-Ping Lee, Jiun-Yan Yang, Kang-Cheng Hou, Chen-Yi Lee, "A 125 μW, Fully Scalable MPEG-2 and H.264/AVC Video Decoder for Mobile Applications" *IEEE Journal of Solid-State Circuits*, vol. 42, no. 1, pp. 161-169, Jan. 2007.

[26] A. A. Muhit, M. R. Pickering, M. R. Frater, J. F. Arnold, "Video Coding using Elastic Motion Model and Larger Blocks," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 5, pp. 661-672, 2010.

[27] Xiph.org Video Test Media, available at *http://media.xiph.org/video/derf/*.

Figures

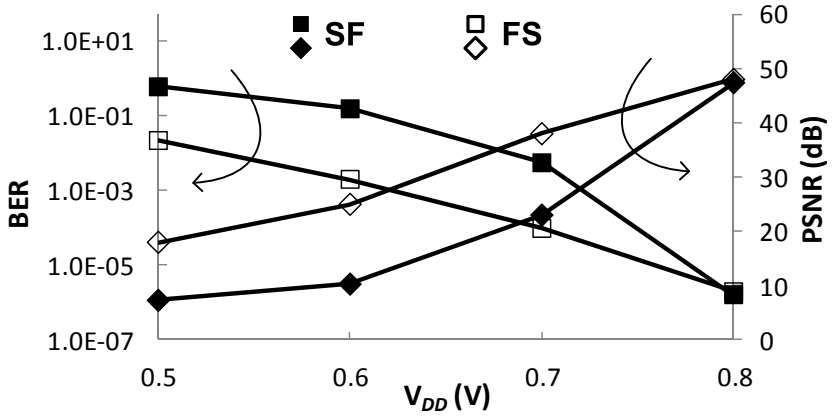


Fig. 1. Bit Error Rate (*BER*) and resulting quality (*PSNR*) vs. V_{DD} in an SRAM array at write-critical (SF) and read-critical (FS) process corner (temperature: 22 °C).

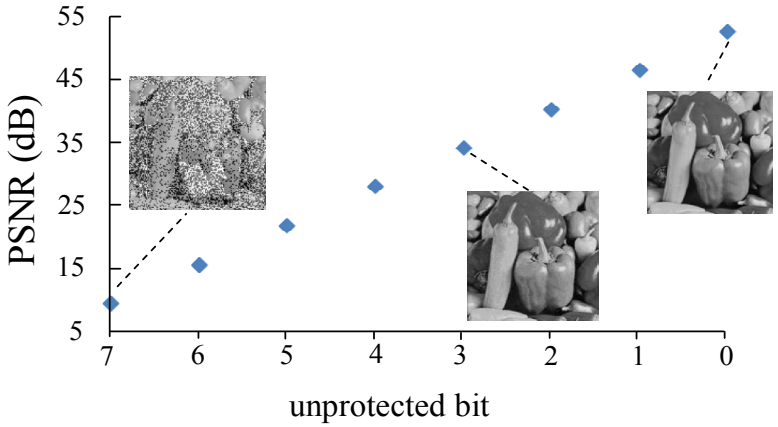


Fig. 2. Measured quality (*PSNR*) degradation due to errors occurring in a single bit position, under 8-bit grey-scale representation (28 nm 6T SRAM, V_{DD}=0.5 V, temperature: 22 °C).

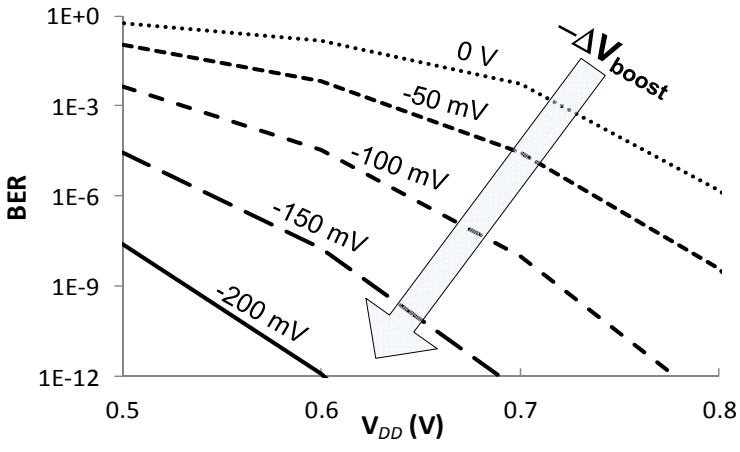


Fig. 3. Write Bit Error Rate (*BER*) vs. V_{DD} for various ΔV_{boost} (SF corner, temperature: 22 °C).

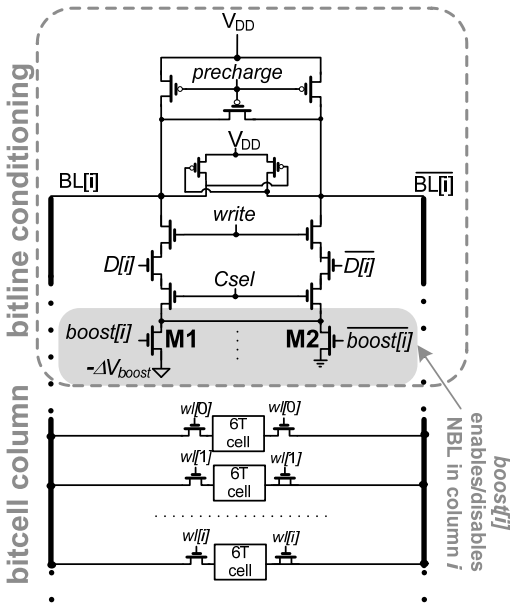


Fig. 4. Selective Negative Bitline boosting (NBL) scheme.

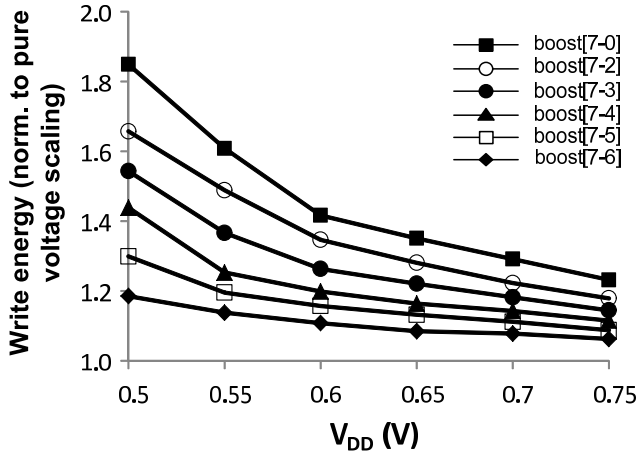


Fig. 5. Write energy for different boosting configurations (normalized to pure voltage scaling – i.e., no boosting).

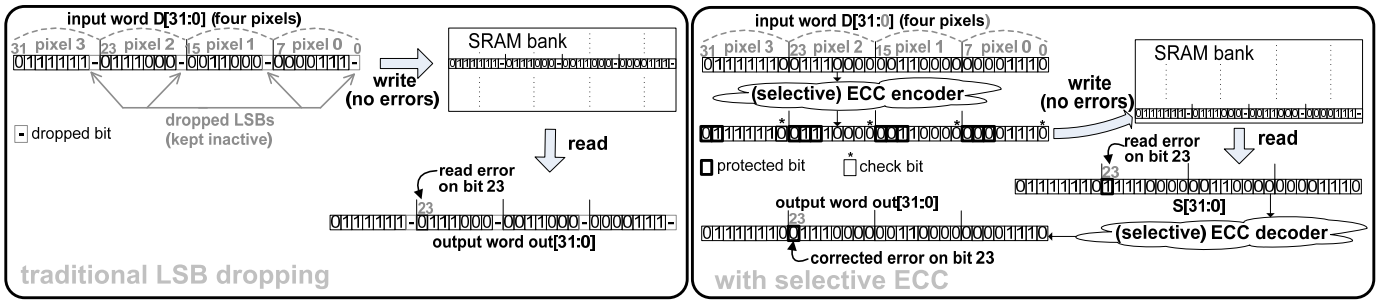


Fig.6. Operation of selective ECC and comparison with traditional LSB dropping scheme [7].

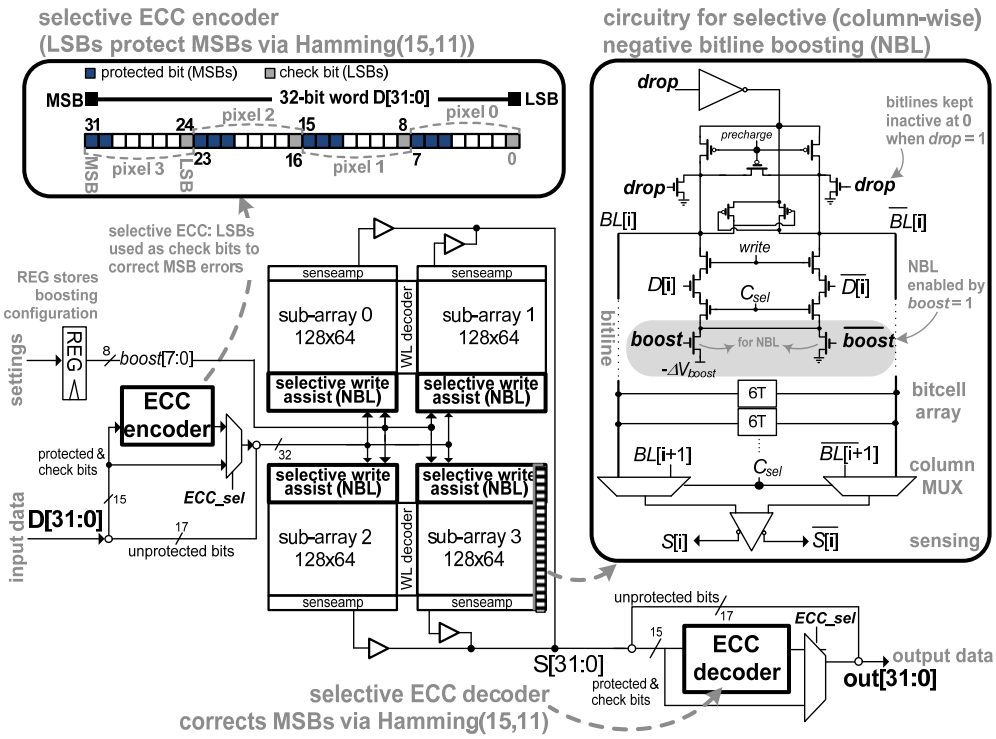


Fig.7. Proposed SRAM architecture that is reconfigurable for error-free and error-tolerant applications.

TABLE I. Comparison with prior art on circuits for error-tolerant applications

	Hybrid 8T/6T [12]	Heterogeneous sizing [16]	Dual supply [14]-[13]	LSB dropping [7]	Traditional ECC [22]	Proposed Selective NBL and ECC (this work)
same energy for MSB and LSB	YES	YES	NO	YES	YES	NO
bitcell array modification needed	YES	YES	YES	NO	YES	NO
dynamically adjustable PSNR	NO	NO	YES	NO	NO	YES
able to reduce V_{DD} at iso-PSNR	YES	YES	NO	NO	YES	YES
need for additional columns	NO	NO	NO	NO	YES	NO
energy penalty needed for extra bits	NO	NO	NO	NO	YES	NO

TABLE II. Testchip information

technology	28 nm
operating voltage	0.5 – 1 V
data retention voltage V_{min}	325 mV
leakage current (V_{DD}=V_{min}, T=22 °C)	11 µA
Area [µm X µm] (%)	
overall 32-kb SRAM array	252 X 202 (100%)
selective NBL	166 (0.3%)
selective ECC encoder	27.8 X 9.4 (0.5%)
selective ECC decoder	24.2 X 14.7 (0.7%)

TABLE III. Array yield (# of std deviations) vs. ΔV_{boost}

ΔV_{boost} (mV)	yield
21	σ
58	2σ
91	3σ
125	4σ
130	5σ

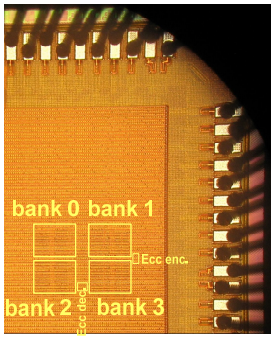


Fig.8. Die photograph.

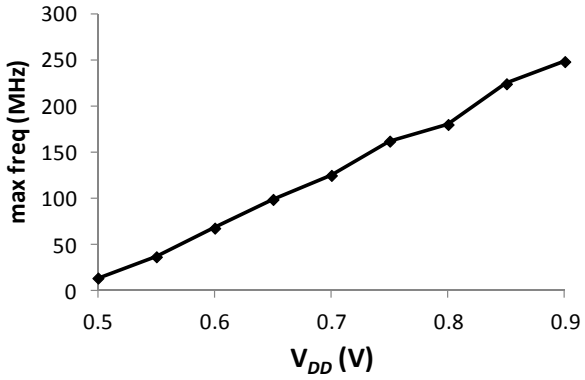


Fig.9. Measured SRAM maximum operating frequency vs V_{DD} (temperature: 22 °C).

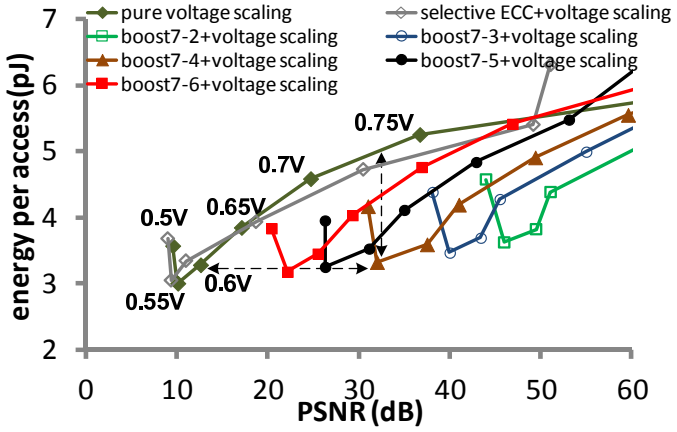


Fig.10. Energy versus quality for different configurations (write-critical corner, temperature: 22 °C).

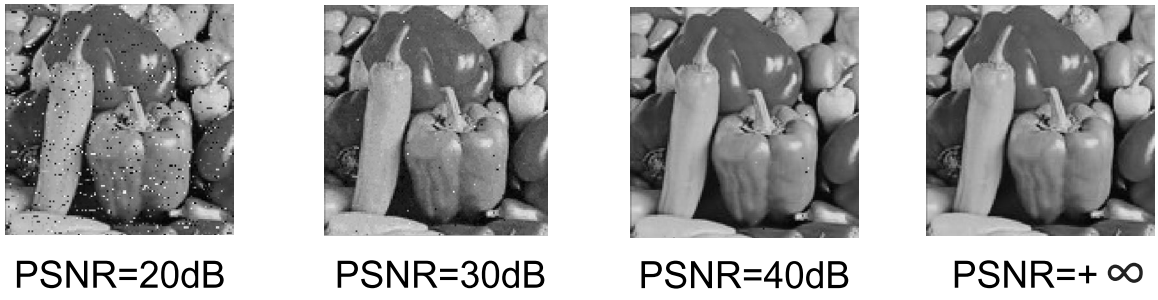


Fig.11. Sample images at different quality (PSNR).

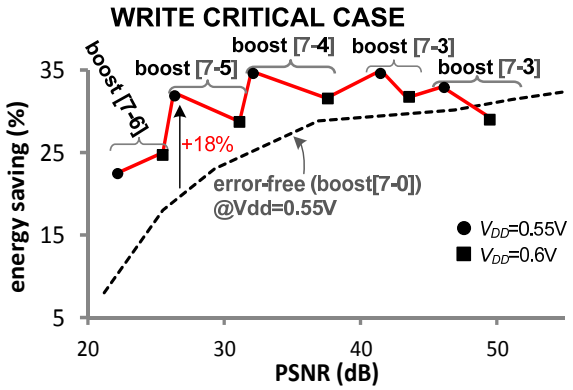


Fig.12. Energy saving of the proposed NBL boosting technique for different PSNR values under the corresponding energy-optimal configuration.

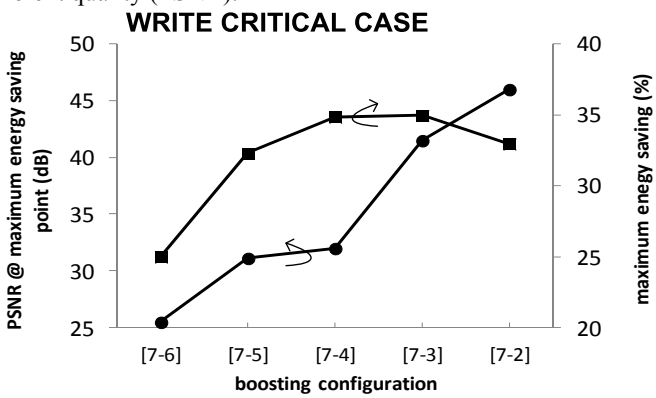


Fig.13. Energy saving over pure voltage scaling and PSNR versus energy-optimal configuration.

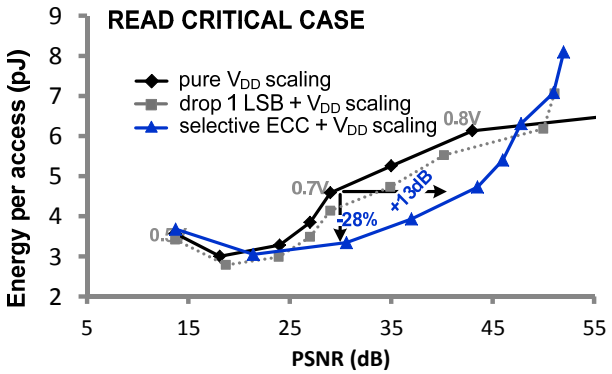


Fig.14. Energy versus quality for different configurations (read-critical corner, temperature: 22 °C).

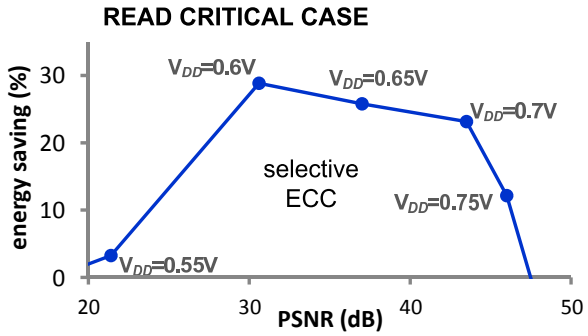


Fig.15. Energy saving of selective ECC over pure voltage scaling vs *PSNR* (read critical corner, T=22 °C).

TABLE IV
Error rate vs bit position (read critical corner, T= 22°C)

bit position	V_{DD}=0.7V		V_{DD}=0.6V		V_{DD}=0.55V	
	No ECC	ECC	No ECC	ECC	No ECC	ECC
7 (MSB)	0.4%	0.0%	1.1%	0.2%	4.2%	2.6%
6	0.2%	0.0%	1.0%	0.1%	3.9%	2.1%
5	0.4%	0.1%	1.0%	0.3%	3.4%	2.2%
4	0.4%	0.4%	1.3%	1.3%	3.8%	3.8%
3	0.3%	0.3%	1.0%	1.0%	3.5%	3.5%
2	0.3%	0.3%	1.1%	1.1%	3.4%	3.3%
1	0.5%	0.5%	1.2%	1.2%	3.5%	3.4%
0 (LSB)	0.4%	-	1.2%	-	3.9%	-

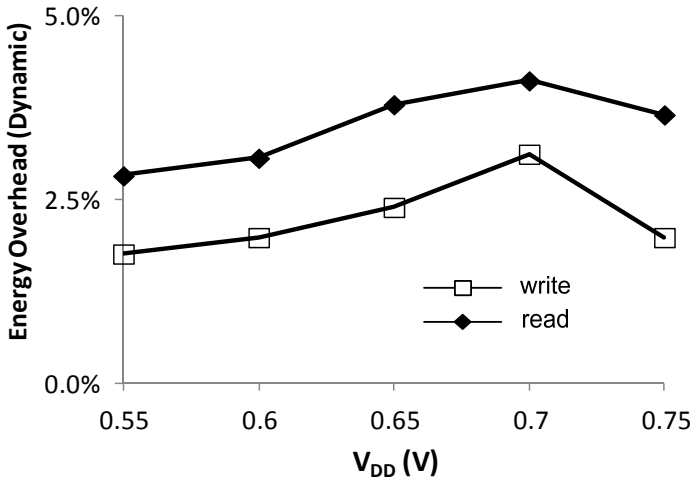
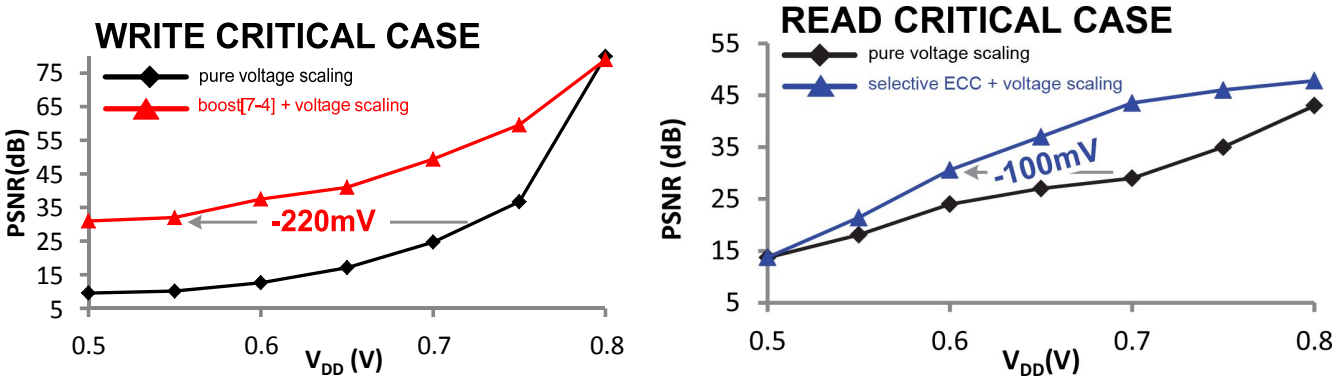
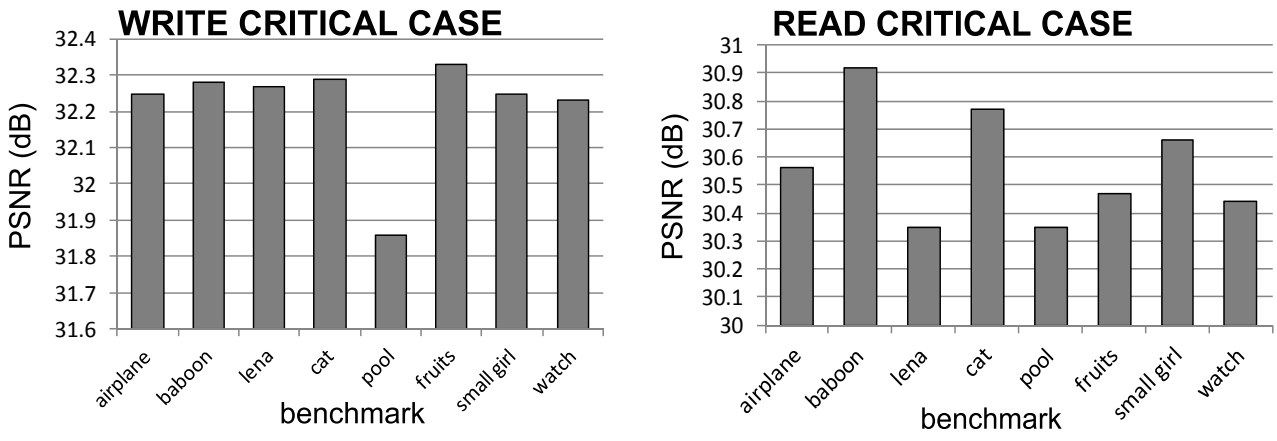


Fig.16. Dynamic energy overhead of selective ECC for a write (encoding) and read (decoding)operation.



(a) (b)

Fig.17. V_{min} reduction compared to pure voltage scaling of a) boosting [7-4] (write critical corner), b) selective ECC (read-critical corner) (temperature: 22 °C).



(a) (b)

Fig.18. PSNR values obtained for different image benchmarks at a) the write-critical corner, V_{DD}=0.55 V, boost[7:4] technique; b) the read-critical corner, V_{DD}=0.6V, selective ECC technique (T= 22°C, targeted PSNR≈30dB).

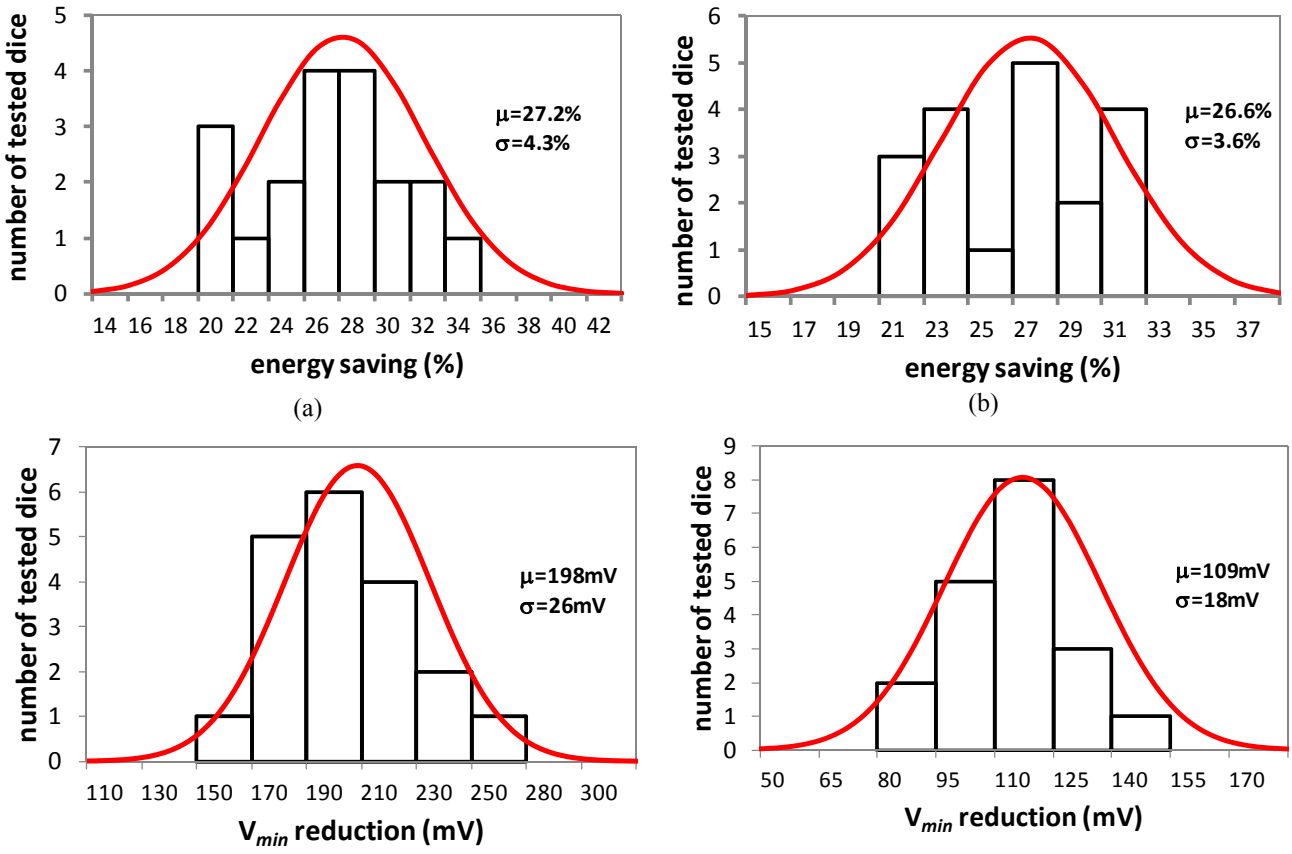


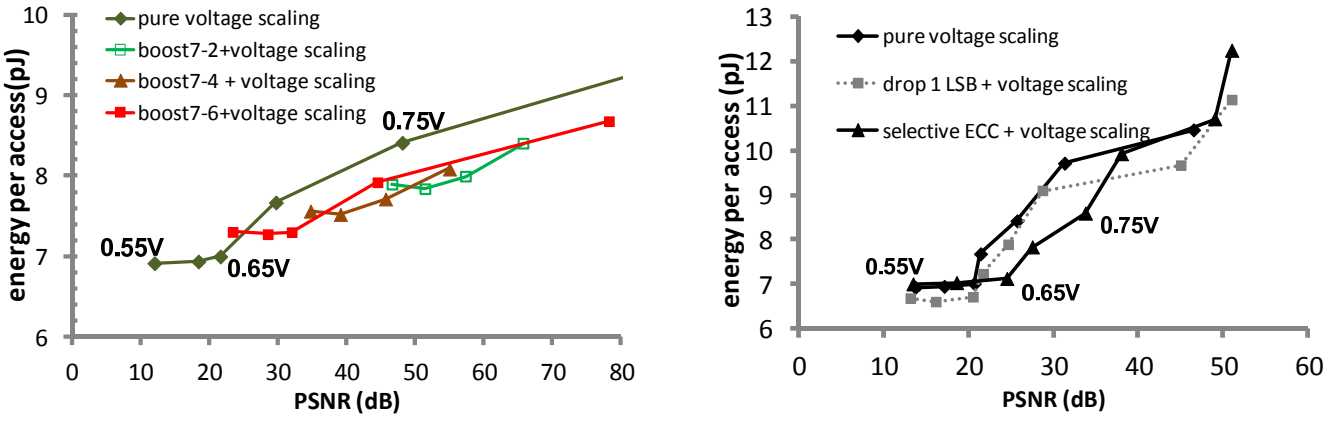
Fig.19. Results across multiple dice at T= 22ºC and targeted PSNR=30dB: energy saving versus die for a) write-critical corner with boosting [7-4]), b) read-critical corner with selective ECC. V_{min} reduction for c) write-critical corner with boosting [7-4]), d) read-critical corner with selective ECC

TABLE V. Measured average error rate in non-boosted columns (write-critical corner)

V_{DD}	# of errors (T=80°C)	# of errors (T=22°C)
0.7V	0.2%	1.0%
0.65V	2.5%	6.1%
0.6V	6.5%	13.5%
0.55V	23.1%	28.8%

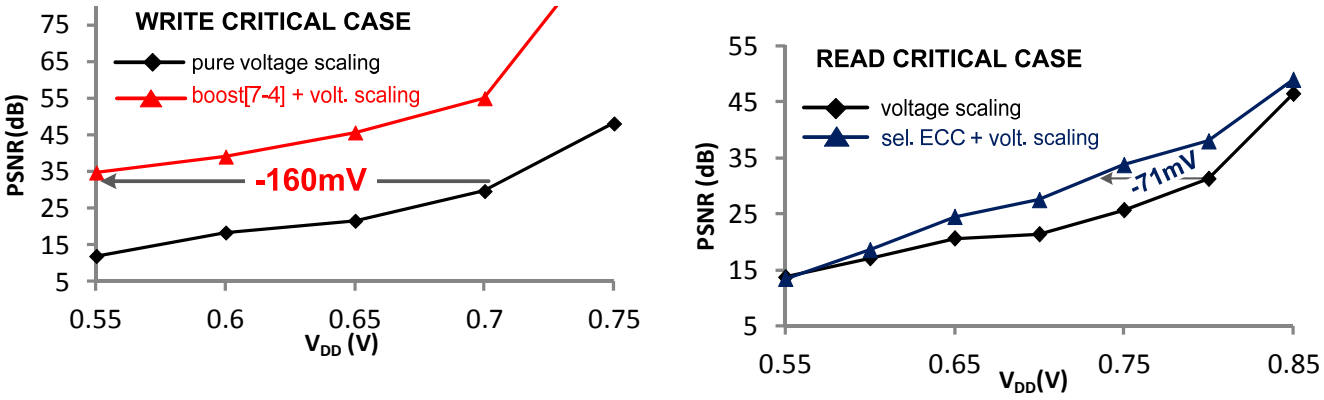
TABLE VI. Measured number of error ratevs. bit position (read-critical corner, T= 80°C)

Bit pos.	V_{DD}=0.7V		V_{DD}=0.6V		V_{DD}=0.55V	
	No ECC	ECC	No ECC	ECC	No ECC	ECC
7 (MSB)	2.1%	0.7%	6.1%	4.6%	13.4%	15.2%
6	2.4%	0.8%	5.5%	4.7%	14.2%	15.8%
5	1.8%	0.8%	5.6%	4.2%	14.0%	15.1%
4	2.2%	2.0%	5.5%	5.2%	13.7%	13.6%
3	1.9%	1.6%	6.1%	5.8%	14.5%	14.2%
2	2.0%	1.8%	5.0%	4.9%	13.5%	13.5%
1	2.0%	1.8%	5.1%	5.2%	13.4%	13.4%
0 (LSB)	1.7%	-	5.9%	-	13.8%	-



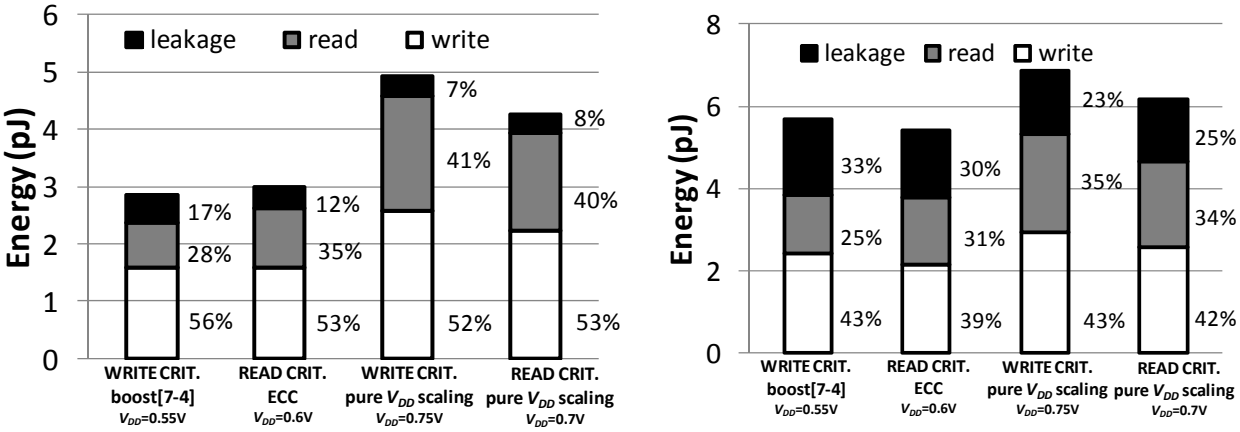
a) b)

Fig.20. Energy versus quality for different configurations: a) write-critical corner; b) read-critical corner T=80°C



(a) (b)

Fig.21.V_{min} reduction compared to pure voltage scaling of a) boosting [7-4] (write critical corner), b) selective ECC (read-critical corner) (temperature: 80°C).



(a) (b)

Fig.22. Energy breakdown for a) T=22°C, b) T=80°C(targeted PSNR≈30dB).

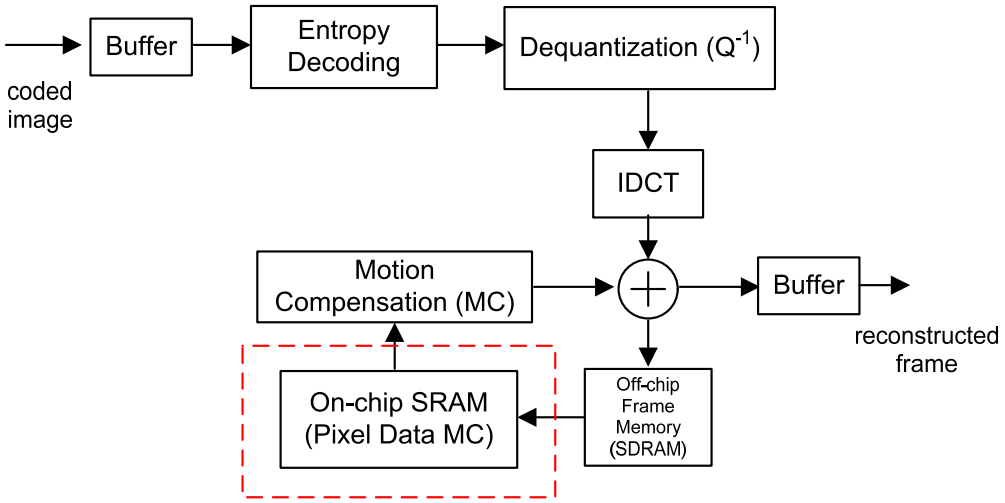
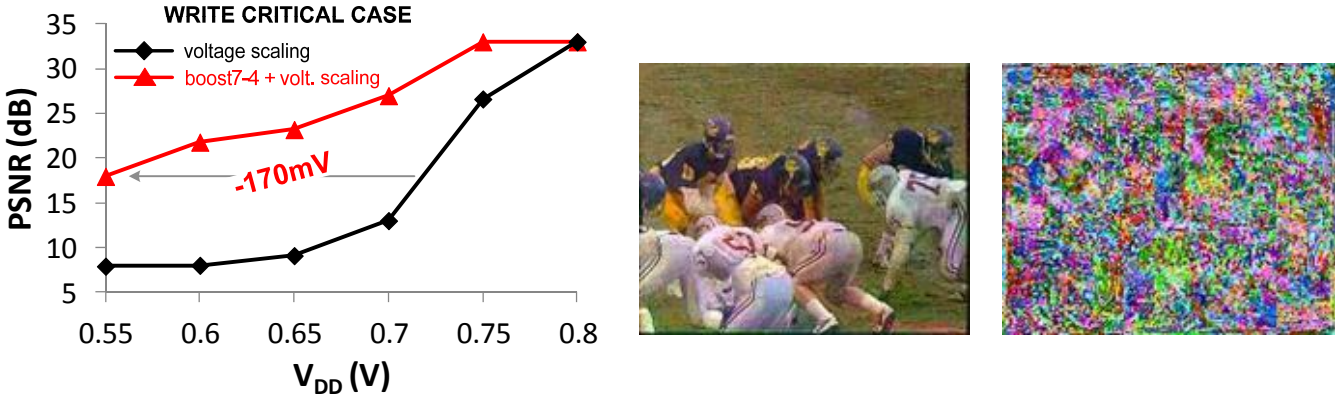
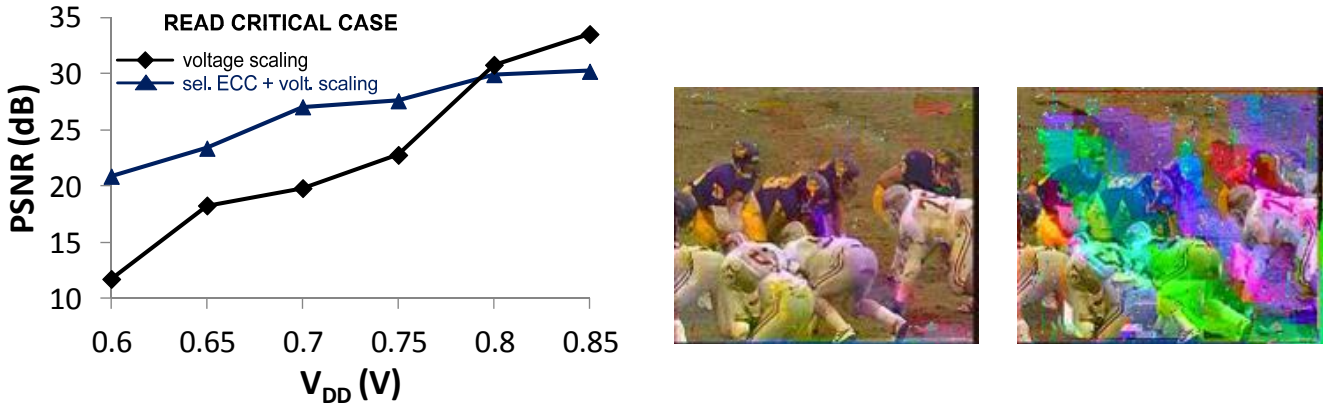


Fig.23. H.264 video decoder system overview (the proposed techniques are applied to the on-chip SRAM).



(a) (b) (c)

Fig.24. a) Output quality of H.264 decoder under boost[7-4] selective NBL and pure voltage scaling (PSNR>20dB),
b) reconstructed frame # 2 for boost[7-4] at V_{DD}=0.6V,
c) reconstructed frame # 2 for pure voltage scaling at V_{DD}=0.6V(write-critical corner,22°C).



(a) (b) (c)

Fig.25. a) Output quality of H.264 decoder under selective ECC and pure voltage scaling (target PSNR>20bB),
b) reconstructed frame # 2 for boost[7-4] at V_{DD}=0.6V,
c) reconstructed frame # 2 for pure voltage scaling at V_{DD}=0.6V (read-critical corner,22°C).