13.8 A 32kb SRAM for Error-Free and Error-Tolerant Applications with Dynamic Energy-Quality Management in 28nm CMOS

Fabio Frustaci[1,2], Mahmood Khayatzadeh[2], David Blaauw[2],
Dennis Sylvester[2], Massimo Alioto[3]

[1]University of Calabria, Rende, Italy,
[2]University of Michigan, Ann Arbor, MI,
[3]National University of Singapore, Singapore, Singapore

Voltage scaling is widely used to improve SRAM energy efficiency [1-2], particularly in mobile systems with tight power budgets. The resulting energy benefits are limited by the minimum voltage ensuring error-free operation, V_{min}, which has stagnated due to growing process variation in advanced technology nodes [3]. Error-tolerant applications and systems (e.g., multimedia) allow more aggressive voltage scaling by operating below V_{min}, which is acceptable if errors due to bitcell write/read failures do not perceptibly reduce application quality (e.g., image quality). Unfortunately, in traditional SRAMs bit error rate degrades rapidly for $V_{DD} < V_{min}$ [4], limiting energy gains. Under a given quality target, further energy reduction is possible through application-specific methods that exploit the features of data stored in a given application [4-5]. However, these approaches are not reusable across applications, and further the energy-quality trade-off is fixed at design time, which degrades energy savings in applications with lower quality targets and in chips near typical corner.

In this work, a new and highly flexible SRAM is developed for use in both error-free and error-tolerant applications, enabling a dynamic energy-quality trade-off. This work focuses on memories targeting video applications. The ideas can be generalized to different error-tolerant applications as well. The techniques are based on the observation that higher-order bits require more aggressive protection than lower-order bits. Two specific approaches are taken: 1) to address read stability, the array dynamically reconfigures lower-order bits to act as error-correcting code (ECC) bits to correct higher-order bits in the same word; 2) for write stability, the array selectively boosts bitlines of higher-order bits only. These techniques have low area overhead (2%) and are configurable in terms of aggressiveness to provide dynamic application-dependent adjustment of the energy/robustness trade-off. Energy is reduced by up to 35% based on measurements of a 28nm testchip.

Figure 13.8.1 shows the measured impact of supply-voltage reduction on bitcell error rate (BER) for a 28nm SRAM, and the resulting quality degradation of a 128×128-pixel grey-scale stored image, expressed as peak signal-to-noise-ratio (PSNR) [4]. Although all bit positions contribute equally to energy/access, quality is most strongly determined by the most-significant bits (MSBs), as shown by the PSNR increase in Fig. 13.8.1 when errors are disallowed in higher-order bit positions. In our SRAM, variation-resilient techniques are selectively introduced to improve the robustness of only those bitcells storing the most significant data. This reduces energy overhead and leads to more graceful quality degradation as bit errors mount at $V_{DD} < V_{min}$, enabling more aggressive voltage scaling. The selective robustness techniques can be independently applied to each bit position, hence our SRAM targets a wide range of energy/quality trade-offs, from low energy and low quality (protecting only MSBs) to error-free operation (protecting all bits for standard computation).

In the SRAM architecture (Fig. 13.8.2), negative bitline boosting (NBL) is used to reduce write failures in cells sharing the same column, while read failures are addressed through ECC, both of which are configurable. For example, in the error-free mode, NBL is enabled in all columns, and a traditional ECC equally protects all the bits within a 32b word D[31:0] (4 adjacent 8b pixels); in error-tolerant modes, NBL and ECC are enabled only in a desired subset of bit positions. As in Fig. 13.8.2, NBL is enabled (disabled) in columns having *boost*=1 (*boost*=0), with *boost* stored in a register, which entails an overhead of only one flip-flop every four columns (i.e., for 8b pixels in a 32b word) and two additional transistors per column for NBL enable. Column multiplexing (2:1) is used and controlled by C_{sel}. Similarly, *ECC_sel*=1 in Fig. 13.8.2 enables selective ECC, which corrects errors occurring in several MSBs in a pixel, as opposed to traditional ECC.

Previous work has adapted to lower quality targets by simply reducing bit width via dropping one or more lower-order bits [6]. This approach renders LSBs inactive, achieving a linear reduction in energy/access. In contrast, our approach re-uses LSBs as redundant bits that are then used by selective ECC to improve MSBs robustness. This enables further voltage scaling, yielding a quadratic reduction in energy/access. As described in Fig. 13.8.2, one LSB of each 8b pixel in the data word is used as a check bit. Selective ECC (a Hamming(15,11) code) protects three MSBs in each of pixels 0 to 2 and two MSBs of pixel 3, for a total of 11 bits protected in a 32b word. Only these MSBs (including check bits) are used as inputs to the ECC encoder. Fig. 13.8.3 shows an example of the selective ECC and traditional LSB dropping when a read error occurs on bit D[23] (i.e., the first MSB of pixel 2).

The above techniques are implemented in a 32kb SRAM in 28nm CMOS, comprising four 128×64 subarrays of traditional 6T cells. Negative bitline boosting voltage is set to –130mV to ensure writeability over 5σ, as appropriate for this 32kb array. The energy/quality trade-off of a test-chip near the SF corner (i.e., write critical, emulated by tuning WL voltage to skew the pull-up ratio) is shown in Fig. 13.8.4. The image testbench *peppers* (128×128 grayscale) is used. When scaling V_{DD}, selective NBL [7-4] on the first 4 MSBs reduces energy (voltage) by up to 35% (from 0.75 to 0.55V) compared to pure voltage scaling at the same quality. Other NBL schemes offer different energy/quality trade-offs: boosting [7-6] has the minimum advantage over pure voltage scaling (24%) due to its worse quality (PSNR = 25dB), while [7-2] has a 33% energy advantage due to the larger number of boosted bitlines and better quality (PSNR = 46dB). From Fig. 13.8.4, such advantage is consistently obtained within the range of practical PSNRs of ≥ 30 dB (see sample images in Fig. 13.8.5). Selective NBL also reduces energy by 18% compared to the error-free case. As expected, selective ECC does not provide significant benefit over pure voltage scaling, since it only corrects read failures (failures are mostly due to writes at SF corner). The same test-chip is used to emulate a read critical corner (FS) by tuning wordline voltage. In this case, using voltage scaling and selective ECC (Hamming(15,11) code), energy (V_{DD}) is reduced by 28% (from 0.7 to 0.6V) compared to pure voltage scaling at iso-quality. The core concept of using the dropped LSB to protect the MSB reduces energy by 19% through added voltage scaling compared to the simple case of keeping the dropped LSB inactive. As expected, selective NBL (omitted in Fig. 13.8.4) does not bring any energy advantage, as failures are mostly due to read.

Figure 13.8.5 shows the total energy advantage for the best combination of the schemes. In write-critical arrays, it is advantageous to progressively increase the number of boosted bitlines from [7-6] to [7-3] for higher PSNR targets, and the energy advantage over pure voltage scaling can be as high as 35% (28% on average) for practical PSNR ≥ 30dB. Similarly, in read-critical arrays, the energy saving enabled by ECC is as high as 28% (23% on average). The impact of process variation is evaluated across 19 dice for boosting [7-4] and PSNR = 30 dB. As in Fig. 13.8.6, average energy saving in write (read) critical case is 25% (27%), which is better (slightly worse) than the value of 20% (28%) obtained for the single chip measurements in Figs. 13.8.4 and 13.8.5. As in Fig. 13.8.6, our techniques significantly reduce V_{min} when operating in an error-tolerant mode compared to conventional voltage scaling. At a PSNR of 30dB, write-critical (read-critical) arrays can voltage scale by an additional 220mV (100 mV).

Acknowledgements:
The authors thank STMicroelectronics for chip fabrication. This work was supported in part by the NSF Variability Expedition and DARPA (agreement HR0011-13-2-0006).

References:

[1] J. Chang, et al., “A 20nm 112Mb SRAM in High-κ Metal-Gate with Assist Circuitry for Low-Leakage and Low-V_{MIN} Applications,” *ISSCC Dig. Tech. Papers*, pp. 316-317, 2013.
[2] H. Pilo, et al., “A 64Mb SRAM in 22nm SOI Technology Featuring Fine-Granularity Power Gating and Low-Energy Power-Supply-Partition Techniques for 37% Leakage Reduction” *ISSCC Dig. Tech. Papers*, pp. 322-323, 2013.
[3] M. Yabuuchi, et al., “A 45nm Low-Standby-Power Embedded SRAM with Improved Immunity Against Process and Temperature Variations,” *ISSCC Dig. Tech. Papers*, pp. 326-327, 2007.
[4] I. J. Chang, et al, “A Priority-Based 6T/8T Hybrid SRAM Architecture for Aggressive Voltage Scaling in Video Applications,” *IEEE TCSVT*, vol. 21, no. 2, pp. 101-112, 2011.
[5] M.E. Sinangil, A. Chandrakasan, “An SRAM Using Output Prediction to Reduce BL-Switching Activity and Statistically-Gated SA for up to 1.9× Reduction in Energy/Access,” *ISSCC Dig. Tech. Papers*, pp. 318-319, 2013.
[6] H. Kaul, et al., “A 1.45GHz 52-to-162GFLOPS/W Variable-Precision Floating-Point Fused Multiply-Add Unit with Certainty Tracking in 32nm CMOS,” *ISSCC Dig. Tech. Papers*, pp. 182-183, 2013.

978-1-4799-0920-9/14/$31.00 ©2014 IEEE

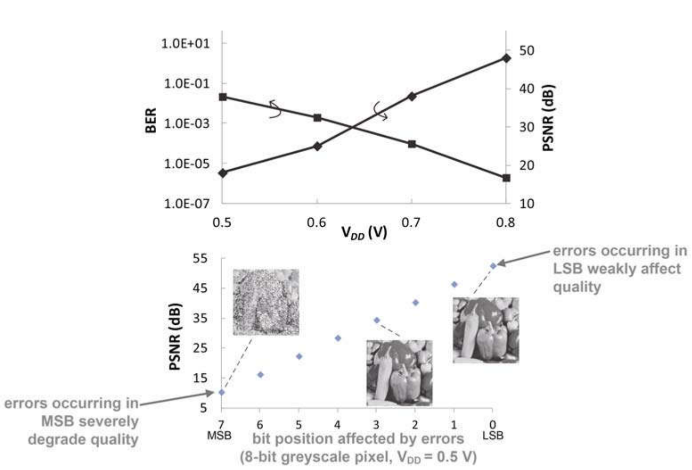


Figure 13.8.1: Aggressive SRAM voltage scaling rapidly degrades bit error rate (BER) and image quality (PSNR). Errors in MSBs impact PSNR more than LSBs (bottom).

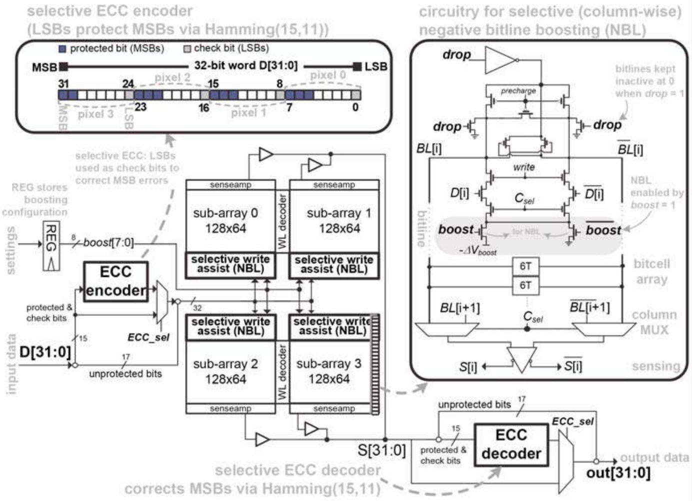


Figure 13.8.2: SRAM architecture (32b word, i.e., 4 8b pixels). MSBs of each pixel are protected via selective ECC (read) and selective NBL (write) for graceful quality degradation and lower energy.

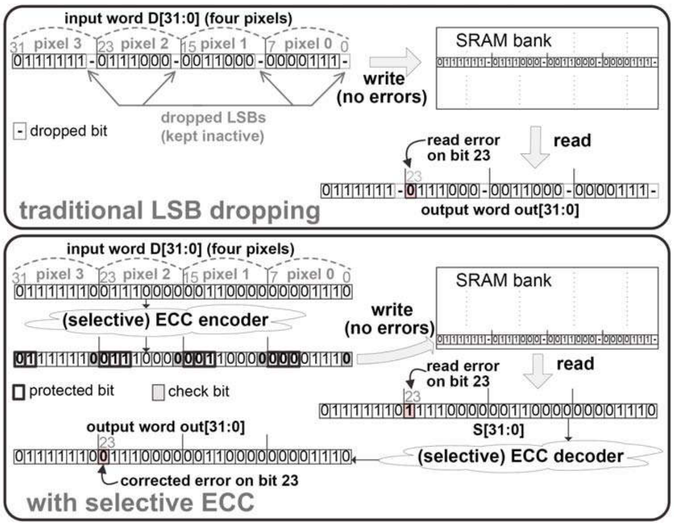


Figure 13.8.3: Operation of selective ECC (LSB employed as check bits of MSBs) as compared to traditional bit dropping.

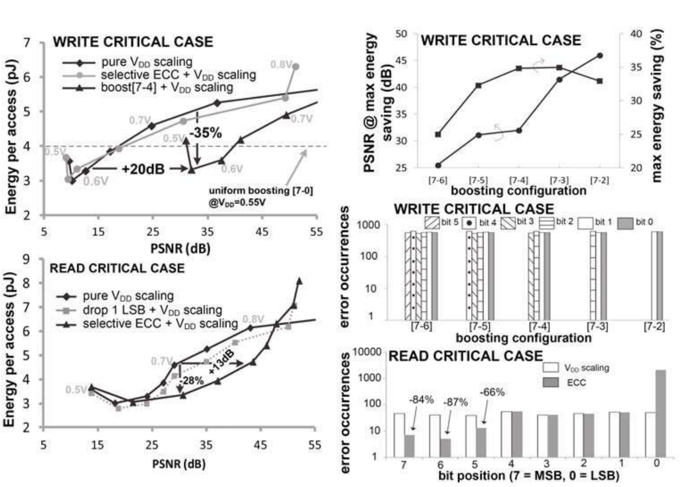


Figure 13.8.4: Measured energy/quality trade-off (left). Energy saving versus boosting configuration (top-right). Selective NBL and ECC suppress errors in MSBs (write: center-right, read: bottom-right).

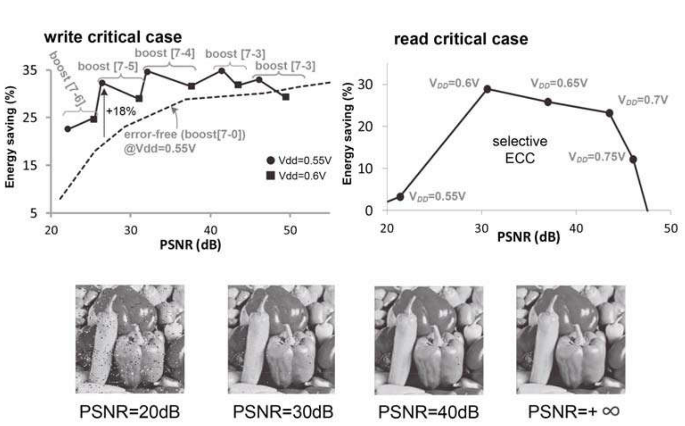


Figure 13.8.5: Measured energy saving in energy-optimal boosting configuration versus PSNR with respect to pure V_{DD} scaling for write-critical and read-critical cases.

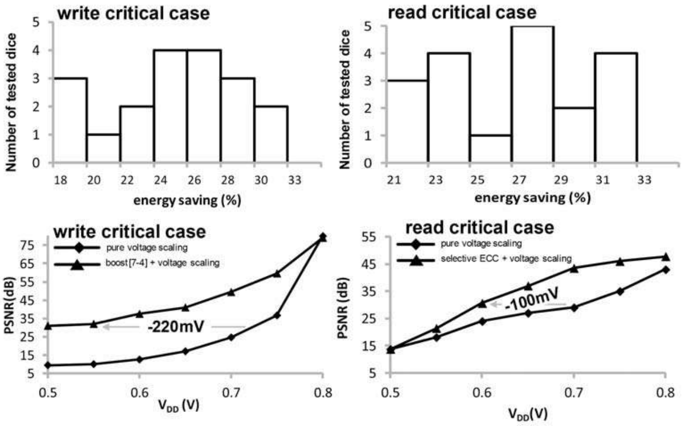
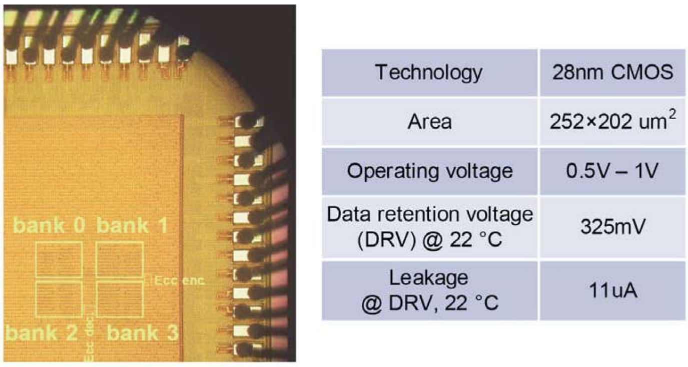


Figure 13.8.6: Energy savings on 19 dice for boost[7-4] (write critical, top-left) and selective ECC (read critical, top-right) with V_{DD} adjusted for PSNR=30dB. For write (read) critical case, average energy saving is 25% (27%). The techniques allow for more aggressive V_{DD} reduction at iso-quality (bottom): -220mV in the write-critical, -100mV in the read-critical case.



Technology	28nm CMOS
Area	252×202 um^2
Operating voltage	0.5V – 1V
Data retention voltage (DRV) @ 22 °C	325mV
Leakage @ DRV, 22 °C	11uA

Figure 13.8.7: Die micrograph of the 28nm test-chip and data.