Ultra-Low Voltage Split-Data-Aware Embedded SRAM for Mobile Video Applications

Na Gong, Shixiong Jiang, Anoosha Challapalli, Sherwin Fernandes, and Ramalingam Sridhar, *Senior Member, IEEE*

***Abstract*—This brief presents an ultra-low voltage split-data-aware 10T and 8T (SDA-10T-8T) embedded static random access memory (SRAM) design for MPEG-4 video processors. Without additional complex peripheral circuits, the proposed design enables a reliable operation at 0.36 V under process variation and aging effect. The experimental results based on 45-nm CMOS technology show that, as compared to conventional SRAM design, our proposed design can achieve a 95% reduction in active power, with no significant degradation in frame quality. In addition, the proposed design suppresses the leakage current effectively, thereby reducing the leakage induced bitline voltage drop rate from 1.54 mV/ns to 0.64 mV/ns at $V_{\rm dd} = 0.36$ V.**

***Index Terms*—Embedded, MPEG-4, negative bias temperature instability (NBTI), process variation, static random access memory (SRAM), ultra-low voltage.**

I. Introduction

RECENTLY, the growing popularity of powerful mobile devices has resulted in the exponential growth of demand for multimedia applications in these devices. Due to the intensive computation, these complex multimedia applications need highly frequent embedded memory accesses and are highly memory dependent. Hence, with embedded static random access memories (SRAMs) consuming a large amount of power, it limits the battery lifetime of mobile devices [1], [2].

Supply voltage scaling is one of the most effective techniques to reduce both leakage and dynamic power consumption [1]. However, designing embedded SRAM for ultra-low voltage operation is challenging because the noise margin of conventional SRAM deteriorates significantly due to process variation and the negative bias temperature instability (NBTI) aging effect, as shown in Fig. 1. In addition, the mobile video applications demand multi-megahertz performance (∼10 MHz for CIF/QCIF to 100 MHz for HD720 [3], [4]), which poses another challenge for ultra-low voltage memory. Also, the low area cost requirement of embedded SRAM is an important design concern.

In order to enable low-voltage operation, many SRAM designs have been presented for video applications. In [2], a hybrid 6T + 8T SRAM structure was proposed to achieve quality-area optimization at 600 mV. However, due to the write failure of 8T bitcells, this technique may not be suitable for ultra-low voltage SRAM design. In [3], a heterogeneous sizing scheme was presented to reduce the failure probability of conventional 6T bitcells, but it suffers from increased computation complexity. In [4], a subthreshold 7T SRAM design was introduced for video applications. In [5], a spatial voltage-scaling technique with optimal supply voltage was presented to achieve power-efficient embedded SRAM. However, all of these require complex peripheral circuits, resulting in large penalties in performance and layout area.

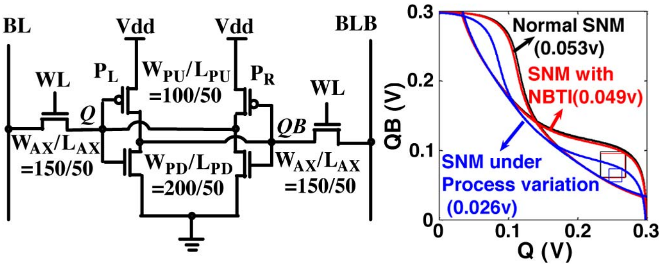


Fig. 1. Standard 6T SRAM and deteriorated static noise margin (SNM) due to process variation and the NBTI effect $(W_{\rm PU} : W_{\rm PD} : W_{\rm AX} = 1 : 2 : 1.5)$.

In this brief, we present a split-data-aware (SDA) 10T and 8T (SDA-10T-8T) embedded SRAM design for power-efficient mobile video applications. By exploring the nature of the pixel data, we use the following techniques to enable a reliable operation at 0.36 V under process variation and NBTI aging effect: 1) a hybrid 10T + 8T hybrid array to improve read static noise margin and reduce the read power simultaneously; 2) a SDA scheme to further increase the write margin and the write power efficiency; 3) a bit-truncation technique to achieve small area overhead. We apply our proposed design in one leading edge video compression system, the MPEG-4 decoder. In our analysis, we use a high-performance 45-nm FreePDK CMOS process to meet the multi-megahertz performance requirement of the video decoder. Also, based on predictive NBTI model [6], we calculate the threshold voltage $(V_{\rm th})$ shift due to aging effect after three years and include it in our simulations.

II. Case Study Overview

A. MPEG-4 Decoder

MPEG-4 is one of the most popular video codec standards [2] in multimedia communications. Fig. 2(a) shows the general block diagram of the MPEG-4 decoder. After entropy decoding, inverse quantization, and inverse transformation, the residual error of frames can be reconstructed based on the compressed video streams. The motion compensator uses the previous reconstructed frames stored in memory and the transmitted motion vectors to construct new frames. To reduce the

Manuscript received July 18, 2012; revised September 18, 2012; accepted October 27, 2012. Date of publication January 11, 2013; date of current version February 1, 2013. This brief was recommended by Associate Editor M. Alioto.

The authors are with the University at Buffalo (SUNY), Buffalo, NY 14260 USA (e-mail: rsridhar@buffalo.edu).

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

Digital Object Identifier 10.1109/TCSII.2012.2231018

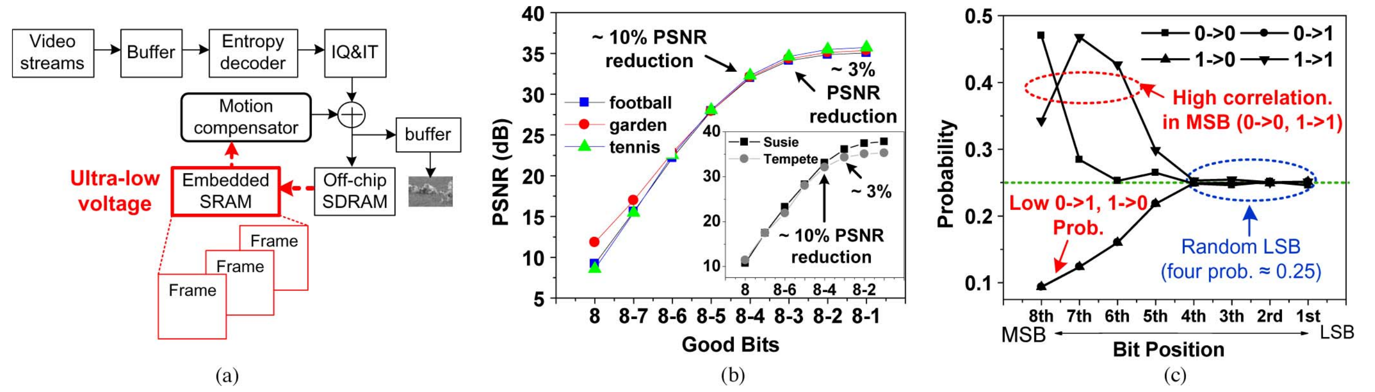


Fig. 2. MPEG-4 decoder and video data characteristics. (a) Block diagram of MPEG-4. (b) PSNR versus good bits. The numbers show the position of good bits. (c) Local correlation in *football* video sequences.

implementation cost, external synchronous DRAM (SDRAM) is used to form hierarchical SRAM/SDRAM architecture. Therefore, the previously decoded frame is stored in embedded SRAM, while the whole decoded frame data is often sent to off-chip SDRAM. Thus, the data is copied to SRAM when needed.

Due to the frequent accesses, embedded SRAM consumes large power consumption, which is the dominant contributor to the whole MPEG-4 decoder power [2]. Accordingly, ultra-low voltage embedded SRAM design is extremely important for power-efficient mobile video applications.

B. Pixel Data-Bit Characteristics

We use peak-signal-noise-ratio (PSNR) as the output quality metric, which is defined as [2], [3]

$$\mathrm{PSNR} = 20 \cdot \log_{10}\left(\frac{255}{\sqrt{\mathrm{MSE}}}\right) \qquad (1)$$

where MSE is the mean square error between the original videos (Org) and the degraded videos (Deg), expressed as

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[\mathrm{Org}(i,j) - \mathrm{Deg}(i,j)\right]^2. \qquad (2)$$

Fig. 2(b) shows the output quality of five known grayscale CIF video sequences (*football, garden, tennis, tempete, susie*) for different failure positions. We observe that, as the two least significant bits (LSBs) fail $(8 - 3$ bits are valid), the PSNR degradation is only $\sim$3%, but it becomes 10% with the failure of three LSBs. This is due to the larger contribution of higher order bits to the frame quality. Another important nature of pixel data in video applications is local correlation. Fig. 2(c) shows the switching probability of pixel data bits in *football* sequences. As shown, the switching activity is not distributed uniformly: for most significant bits (MSBs), they are highly correlated and have much lower switching probability than LSBs. For example, for the 8th bit, the probabilities of keeping "0" $(0 - > 0)$ and "1" $(1 - > 1)$ are 0.47 and 0.34, respectively. Alternately, the LSBs tend to store random data and have 0.5 switching probability.

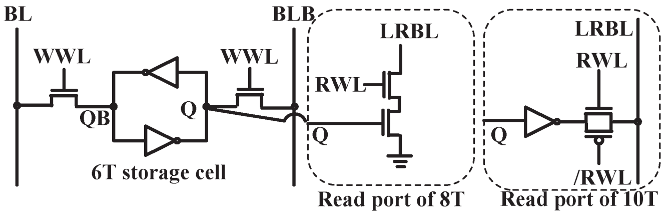


Fig. 3. Conventional 10T and 8T bitcells with a conventional 6T storage cell and separate read port.

III. Proposed Ultra-Low Power SRAM

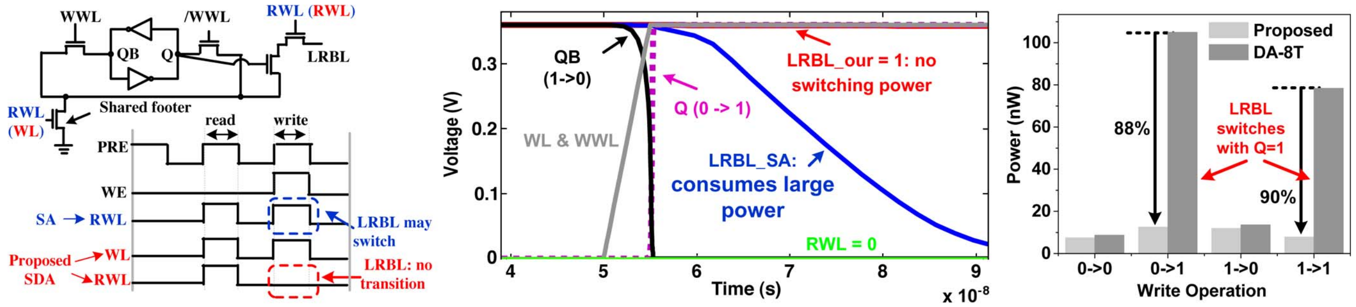
A. Hybrid Bitcell Array Design

In the proposed design, we use a 10T + 8T hybrid bitcell array to exploit their power efficiency in video applications. Fig. 3 shows the structure of conventional 8T bitcell and the 10T bitcell in [8]. We can see that these two cells isolate the read port from the conventional 6T storage cell, achieving superior read stability. As compared to the conventional 8T cell, the readout circuit of a 10T SRAM cell includes an inverter and a transmission gate, and it does not need a precharge circuit in bitlines. Therefore, the voltage on the bitline does not switch as the readout data bits are consecutive, suppressing the read power effectively. As discussed in Section II-B, MSBs in pixel data have a high probability of keeping "0" and "1," so we adopt 10T bitcells to store MSBs in pixel data. To reduce the area overhead, we use conventional 8T bitcells for LSBs.

However, both 10T and 8T bitcells suffer from write half-select disturb. To overcome this, a data-aware (DA) technique was proposed in [9], as shown in Fig. 4(a). During the write operation, the enabled row-based word line (RWL) and column-based DA write word line (WWL or /WWL) select a cell, and the stability of half-selected cells would not be affected. Therefore, the DA technique achieves a cross-point write structure, enabling a bit-interleaving architecture.

B. Proposed SDA Technique

The DA technique eliminates write half-select disturb, facilitating aggressive voltage scaling in memory. However, it suffers from large power consumption in the write operation. This is because the row-based shared footer is controlled by the RWL. During write operation, RWL is enabled to turn on the shared footer and the readout path is only controlled



(a) (b) (c)

Fig. 4. (a) Proposed SDA technique. (b) Waveform of LRBL in the write operation. (c) Power savings of proposed SDA-8T as compared to DA-8T.

by the stored data Q [see Fig. 4(a)]. Therefore, the floating high read bit-line (LRBL) may be discharged if Q turns on the read-out path, inducing large dynamic power, as shown in Fig. 4(b). To reduce the large write power, we proposed a SDA technique for 8T and 10T cells. As shown in Fig. 4(a), SDA isolates LRBL in the write operation by separating the control signal of shared footer (WL) from RWL: WL is enabled for both read and write operations while RWL is only enabled for read operation. Accordingly, SDA eliminates unnecessary transitions of LRBL and suppresses the power consumption [see Fig. 4(b)]. To implement the SDA scheme, the split WL and RWL can be easily obtained based on the write-enable signal in the existing word line decoder.

Fig. 4(c) compares the power consumption of proposed SDA-8T and DA-8T cells in write operations. As shown, the write power can be reduced up to 90% by the proposed SDA scheme. Note that the proposed SDA technique can be applied for other SRAM bitcells with conventional 6T storage cell and an isolated read port.

C. Bit-Truncation Technique

As discussed before, embedded SRAMs usually occupy a large portion of area in a video chip, and therefore the area cost of the embedded SRAM is an important design concern. However, as compared to conventional 6T bitcells, 8T and 10T bitcells in our proposed SDA-10T-8T scheme both consume a larger layout area. To reduce the area overhead, we adopt a bit-truncation technique based on the insignificant contribution of LSBs in pixel data, as discussed in Section II-B. This method has been applied widely in low-power video compression applications to reduce computation complexity [10]. Here, we use it to achieve area-efficient embedded SRAM design: the last two less important bits of each pixel are skipped and replaced with zeros. The simulation results in the following sections show that the bit-truncation technique results in lower power consumption, smaller area overhead, while delivering output quality with no significant degradation.

IV. Implementation

A. Optimal Number of 10T Cells

A key issue during the implementation of the proposed SDA-10T-8T SRAM is to select the number of 10T cells in an array. Based on conservative MOSIS deep submicrometer design rules [11], we designed the layout of SDA-8T and SDA-10T cells, as shown in Fig. 5. We can see that, compared to a conventional 6T cell (1662.5×400 nm^2, not shown), the areas of a SDA-10T and SDA-8T cell are increased to 3025×400 nm^2 and 2225×400 nm^2, resulting in $\sim$82% and 34% area overhead, respectively.

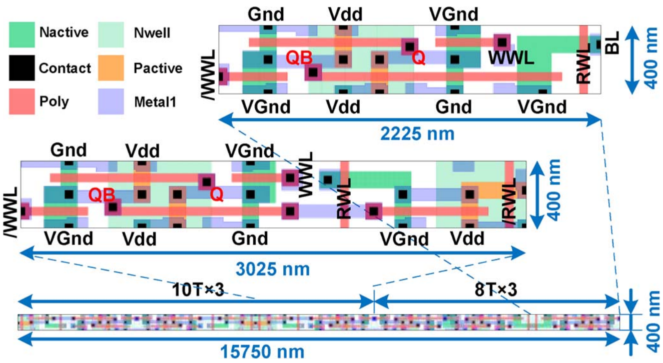


Fig. 5. Layout design of proposed SDA-10T-8T SRAM.

Therefore, we can express the area overhead of a SDA-10T-8T array as

$$\begin{aligned}\text{Area\_Overhead} &\cong \frac{1.82 \times n + 1.34 \times (N - n) - N}{N}\\ &= \frac{48n + 34N}{N}\% \end{aligned} \quad (3)$$

where n is the number of 10T cells and N is the total number of cells in an array. As n increases, the power consumption would be reduced due to the higher power efficiency of 10T cells. However, at the same time, the area overhead becomes larger. Therefore, determining n is an area-power optimization problem. Here, we define a new quality metric Power Efficiency per Area overhead metric (PEA) for the proposed design:

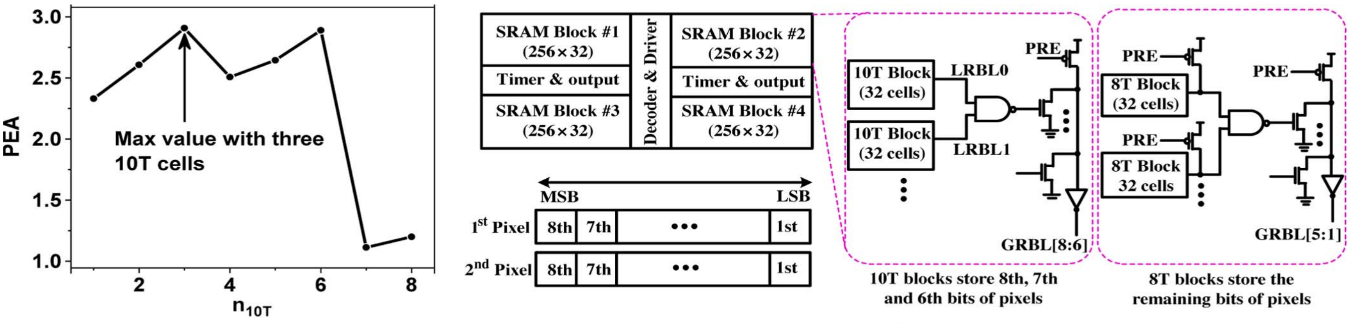
$$\text{PEA} = \frac{\Delta\text{Power}(n)}{\text{Area\_Overhead}} \quad (4)$$

where power reduction $[\Delta\text{Power}(n)]$ indicates the power improvement with n 10T cells as compared to only 8T cells.

Hence, the optimization problem can now be formulated as

$$n_{opt} = \underset{0 \leq n \leq 8}{\arg\max}(\text{PEA}). \quad (5)$$

Based on the above equations, we plotted PEA as a function of n, as shown in Fig. 6(a). n_{opt} was found to be three. Therefore, as compared to the conventional 6T design, the area overheads of the proposed 8-bit SDA-10T-8T design and 6-bit SDA-10T-8T design are 52% and 18.5%, respectively.



(a) (b)

Fig. 6. (a) Optimal number of 10T cells. (b) Array architecture of the proposed SRAM.

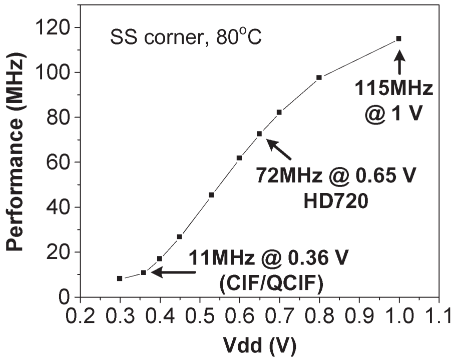


Fig. 7. Performance of the proposed SDA-SRAM design.

B. Array Architecture of the Proposed SRAM

Fig. 6(b) shows the array architecture of the proposed hybrid SRAM, where the total array size is 32 kbit and there are four blocks. A hierarchical bitline scheme (local RBL and global RBL) is applied to reduce the delay time. To further enable high performance in ultra-low voltage operation, we determine that the number of bit-cells per LBL is 32. Based on the above analysis, for each 8-bit pixel data, the first three MSBs $(8_{\rm th} - 6_{\rm th})$ are stored in 10T cells and the rest bits are stored in 8T bits.

V. Simulation Results

A. Performance

Fig. 7 shows the performance simulation results for the proposed SDA-10T-8T design at the SNSP-NBTI-80 °C corner. It is shown that, despite the worst process corner and NBTI aging effect, the proposed SDA-10T-8T design shows 11-MHz performance at 360 mV$V_{\rm dd}$, meeting the requirement of operation frequency of CIF/QCIF video format. Also, the proposed design can operate at 72 MHz as $V_{\rm dd}$ is 0.65 V, which successfully delivers high-quality HD720 format.

B. Output Quality

Also, we use 50 frames of three known grayscale CIF video sequences: *football, garden*, and *tennis*, to verify the output quality based on the proposed SRAM scheme. The frame size in our simulation is 352 × 240. In order to observe the video-quality degradations during the low-voltage operations, we first performed 10 000 Monte Carlo simulations to obtain the failure probabilities of SRAM bitcells for different SRAM schemes with local $V_{\rm th}$ variation in the worst global process corner. Then, we assumed the failed bits to be located across the memory cells based on the failure probabilities according to a uniform distribution, introducing embedded memory failures to the decoding process. Finally, we capture the video frames on the MPEG-4 decoder side. Due to space limitations, we only present the results of the *football* clip, as shown in Fig. 8. We can see that the conventional 6T SRAM results in significant degradation of frame quality at 360 mV. Alternatively, our proposed SDA-SRAM scheme can deliver output quality with no significant degradation.

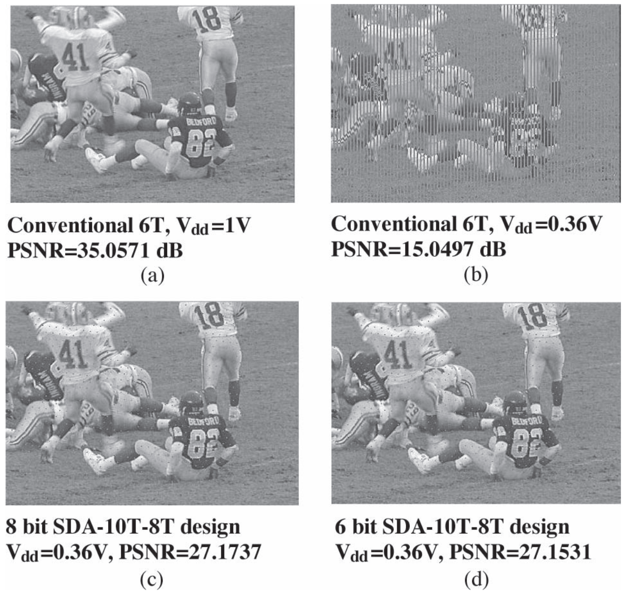

Conventional 6T, V_{dd}=1V
PSNR=35.0571 dB
(a)
Conventional 6T, V_{dd}=0.36V
PSNR=15.0497 dB
(b)
8 bit SDA-10T-8T design
V_{dd}=0.36V, PSNR=27.1737
(c)
6 bit SDA-10T-8T design
V_{dd}=0.36V, PSNR=27.1531
(d)

Fig. 8. Output quality.

C. Active Power Consumption

We use the following model to estimate the overall active power consumption including both dynamic and leakage power of embedded SRAM:

$$P = P_w + P_r \tag{6}$$

where $P_w(P_r)$ is the power consumption during write (read) operation and can be expressed as

$$P_w = \sum_{i,j\in(0,1)} [F(i,j)\cdot P_w(i,j)]$$
$$P_r = \sum_{i,j\in(0,1)} [F(i,j)\cdot P_r(i,j)] \tag{7}$$

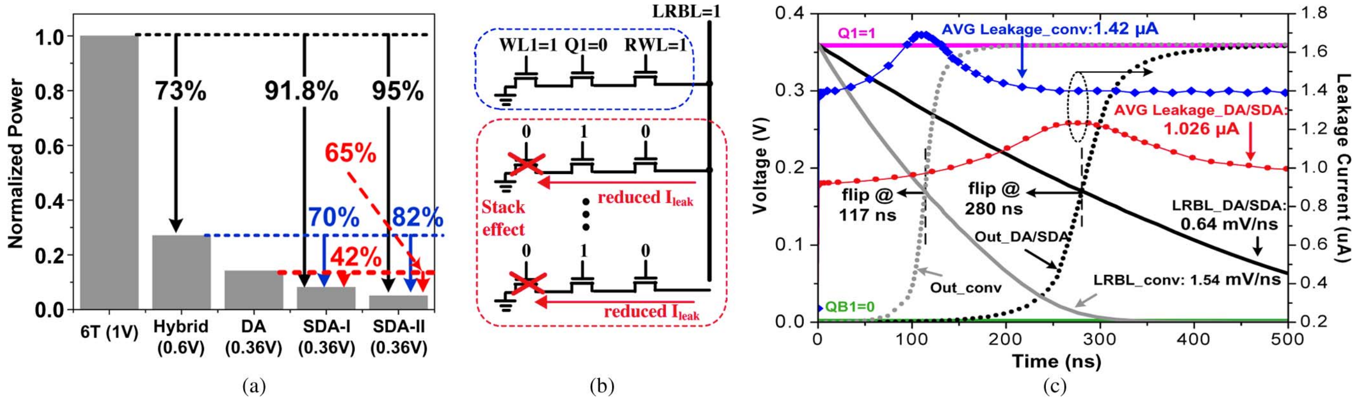


Fig. 9. (a) Active power consumption. Hybrid: 6T + 8T design in [2]; SDA-I: 8-bit SDA-10T-8T; SDA-II: 6-bit SDA-10T-8T. (b) Leakage current with worst-case data pattern. (c) Leakage-induced voltage drop with 32 cells on local read bit-line.

where i and j are old and new values, respectively. $F(i, j)$ indicates the probability of switching i to j, and it is extracted from the frame data in the decoding process [see Fig. 2(c)].

Here, we estimated the write power and read power in TT corner, 80 °C, and the frequency was 10 MHz. Fig. 9(a) shows the power saving of our proposed technique over a standard SRAM design. It can be seen that significant power saving can be achieved with our technique. As compared to the conventional 6T design with $V_{\mathrm{dd}} = 1$ V, 91.8% and 95% active power reduction can be obtained for the proposed 8-bit and 6-bit SDA-10T-8T (with the bit-truncation technique), respectively. In addition, additional 70% and 82% savings can be achieved over the hybrid design in [2] for the proposed two designs, respectively. As also shown in Fig. 9, due to the smaller write power, the proposed 8-bit and 6-bit SDA-10T-8T schemes reduce the overall active power by 42% and 65% as compared to the DA technique, respectively.

D. Leakage Current

In ultra-low voltage operation, the leakage current is not only a power issue for SRAM. More importantly, the high leakage current of unselected cells degrades the $I_{\mathrm{off}}/I_{\mathrm{on}}$ and causes an undesirable voltage drop in bitline. With the worst-case data pattern [see Fig. 9(b)], the leakage current may induce sensing failure. In the DA design and the proposed SDA-10T-8T design, the footers connected to the unselected cells are all disabled. Due to the stack effect, the leakage current can be suppressed effectively. Fig. 9(c) shows the simulation results at the FNFP -0.36 V $-$ 80 °C corner. We can see that, when there are 32 cells connected to a single bitline, the average leakage current (AVG Leakage_DA/SDA) is reduced from 1.42 μA to 1.026 μA as compared to the conventional 8T design (AVG Leakage_conv). At the same time, the leakage caused RBL voltage drop rate is reduced from 1.54 mV/ns to 0.64 mV/ns. Even though the DA scheme shows the similar low leakage characteristics, our proposed SDA scheme achieves a considerable reduction in both active power and leakage power.

VI. Conclusion

An ultra-low voltage embedded SRAM design has been presented for low-power mobile video applications. Based on the nature of pixel data, we developed an SDA technique and an 10T + 8T hybrid array for achieving a low V_{dd}. In addition, the bit-truncation technique was used for a small area overhead and further power savings. Simulation results demonstrate that the proposed design achieves 95% power savings as compared to the conventional SRAM.

References

[1] 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.

[2] I. J. Chang, D. Mohapatra, and K. Roy, “A priority-based 6 T/8 T 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.

[3] 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.

[4] J. S. Wang, P.-Y. Chang, T.-S. Tang, J.-W. Chen, and J.-I. Guo, “Design of subthreshold SRAMs for energy-efficient quality-scalable video applications,” *IEEE Trans. Emerging Sel. Topics Circuits Syst.*, vol. 1, no. 2, pp. 183–192, Jun. 2011.

[5] N. Gong, S. Jiang, A. Challapalli, M. Panesar, and R. Sridhar, “Variation-and-aging aware low power embedded SRAM for multimedia applications,” in *Proc. IEEE SoCC*, Sep. 2012, pp. 21–26.

[6] PTM Model. [Online]. Available: http://www.eas.asu.edu/~ptm

[7] M. Qazi, M. E. Sinangil, and A. P. Chandrakasan, “Challenges and directions for low-voltage SRAM,” *IEEE Des. Test Comput.*, vol. 28, no. 1, pp. 32–43, Jan./Feb. 2011.

[8] H. Noguchi, Y. Iguchi, H. Fujiwara, Y. Morita, K. Nii, H. Kawaguchi, and M. Yoshimoto, “A 10 T non-precharge two-port SRAM for 74% power reduction in video processing,” in *VLSI Symp. Tech. Dig.*, Mar. 2007, pp. 107–112.

[9] Y.-W. Chiu, Y.-Y. Lin, M.-H. Tu, S.-J. Jou, and C.-T. Chuang, “8 T single-ended sub-threshold SRAM with crosspoint data-aware write operation,” in *Proc. IEEE ISLPED*, Aug. 2011, pp. 169–174.

[10] T. Xanthopoulos and A. P. Chandrakasan, “A low-power DCT core using adaptive bitwidth and arithmetic activity exploiting signal correlations and quantization,” *IEEE J. Solid-State Circuits*, vol. 35, no. 5, pp. 740–750, May 2000.

[11] MOSIS Deep Design Rules. [Online]. Available: http://www.mosis.com/