This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2019.2958668, IEEE Transactions on Circuits and Systems II: Express Briefs

Energy-Efficient Data-Aware SRAM Design Utilizing Column-based Data Encoding

Anh Tuan Do, *Member, IEEE*, Seyed Mohammad Ali Zeinolabedin, *Member, IEEE*, and Tony Tae-Hyoung Kim, *Senior Member, IEEE*

***Abstract*— This paper presents an ultra-low power SRAM utilizing a column-based data encoding scheme for power reduction. The proposed scheme is particularly beneficial in applications like bio-signal and image processing where neighboring data have similar values. The proposed technique generates write data through bit-wise comparison, which leads to a larger number of "0s". To utilize this, a data-aware bitline pre-charge scheme is proposed to minimize the write power for "0". In addition, a PVT-tracking bias generator compensates for the read bitline leakage to improve the sensing margin. A 32Kb SRAM in 65nm CMOS technology shows successful operation down to 0.36 V with the power of 0.37 µW and the maximum frequency of 0.25 MHz. The minimum energy is 0.3 pJ/access at 0.5 V.**

Keywords—SRAM; data-aware; ultra-low power; bitline leakage.

I. Introduction

Ultra-low power and energy SRAMs have been highly demanded in many emerging applications such as Internet-of-Things (IoT), wearable devices, etc [1]. Since SRAMs dominate power consumption and silicon area, various techniques for ultra-low power consumption have been developed [2]. One of the most popular techniques is to use the sub- or near-threshold operation [1]. However, this leads to various SRAM design issues such as lower cell stability, poor write margin, and reduced sensing margin [3]. The employment of decoupled SRAM cells improves the cell stability during read operation by isolating internal cell nodes from read bitlines [3, 4]. Other read and write assist techniques such as collapsed power supply [2], boosted wordline [5], negative bitline (BL), hierarchical bitlines structures [6], and sense amplifier redundancy [7] have also been reported to improve the reliability of ultra-low power SRAM operation.

Recently, application-specific SRAM design has gained attention to further reduce the read and write power [8, 9, 12, 13, 14]. For example, the SRAM in [8] employs 8T cells for MSBs but 6T cells for LSBs for supply voltage scaling without significant increase in the error rate. In [9], the SRAM leverages on high similarities in nearby cell data to predict read output and thus reduce read power. Similarly, the SRAM in [12] utilizes the similarity of neighboring cells and generates pseudo-random write data. This removes the worst and best cases of read bitline sensing. The SRAM in [13] checks the number of "1"s in the write data and flips them if "1" is dominant. This allows the SRAM to store more "0"s, which consumes less read power by avoiding read bitline discharging. The ReRAM-based TCAM in [14] reduces the search energy by realizing that only the hit rows discharge.

In this work, we further utilize the features of the data in bio-signal and image processing applications [8, 9, 10, 12] where data is largely biased in most operations. Several techniques are proposed for power reduction such as (i) column-based data encoding, (ii) data-aware pre-charge, and (iii) PVT-tracking bias generation.

II. Proposed 8T Data-Aware SRAM (8T-DAS)

A. 8T SRAM Cell

This work employs an 8T SRAM cell (Fig. 1(a)) for low-voltage operation [3]. To minimize the leakage current, higher threshold voltage (HVT) transistors are used in the 6T cell structure while standard threshold voltage (SVT) transistors are used in the read port (i.e. M7-M8). Even though this cell has no read disturbance, bitline disturbance in the half-selected cells during write operation still limits the worst case cell stability [15]. Therefore, the write access transistors must be sized carefully to avoid unwanted data flipping at target voltage (e.g. 0.4 ~ 0.5 V in this work).

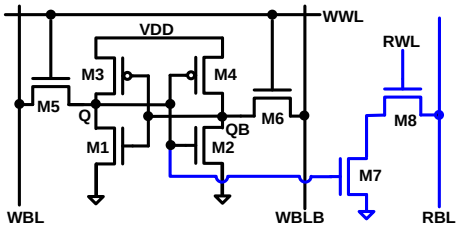

Fig. 1. 8T SRAM cell schematic.

B. Data-aware WBL pre-charge scheme

In conventional SRAMs, write bitlines (WBL and WBLB) are pre-charged to V_{DD} as shown in Fig. 2(a). In each write operation, either WBL or WBLB is discharged, which makes write power constant regardless of write data. However, the proposed SRAM will have much more "0s" than "1s" in the target applications [8, 9, 10, 12], which provides an optimization opportunity. Fig. 2(b) illustrates the proposed write driver and its operation. Since most of the write data is "0", WBL and WBLB are pre-charged to ground and V_{DD} for removing dynamic power during writing "0". It is obvious that the power for writing "1" is doubled since both WBL and WBLB switch. However, we can still achieve overall power saving because of the write data biased to "0". The saving in the write power of the proposed pre-charge scheme depends on the percentage values of "0"s in the write data. As expected, the breakeven point is formed at 50%. Therefore, it is critical to maintain the overall percentage of "0"s higher than 50%, which is explained in the following section.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2019.2958668, IEEE Transactions on Circuits and Systems II: Express Briefs

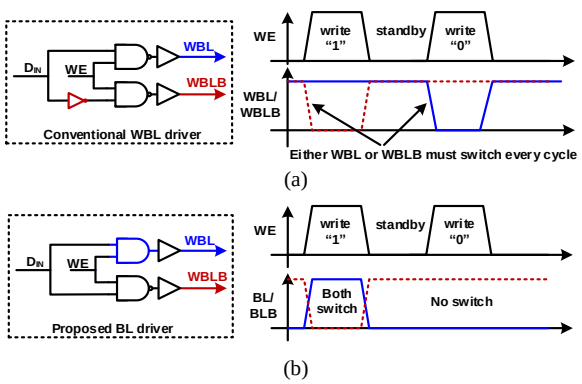

Fig. 2. (a) Conventional write driver and its operation and (b) proposed data-aware write driver and its operation [16].

C. *Proposed column-based data encoding and recovery*

In most image and video processing applications [8, 9, 11], a block of data is written into an SRAM sequentially. Therefore, nearby SRAM cells are likely to store similar or biased data because of the nature of the image and video signals. Data compression can be utilized to further reduce the amount of data for lower power. However, it can still generate biased data. In SRAMs, biased column data represent either a best case or a worst case in terms of bitline sensing. In this work, biased column data is converted into a data pattern similar to the best case. The key principle is as follows. First, the first word of a data block is written directly into an SRAM. Subsequently, next word is compared with the previous word in a bitwise manner to generate encoded write data. If two consecutive bits are identical, the encoded bit becomes "0". Otherwise, the encoded bit of "1" is generated. When read, the original data can be easily recovered through the sequential bitwise comparison.

Fig. 3 illustrates the detailed operation of the proposed column-based encoding and recovery scheme using a data block of 3 words (W_1, W_2, and W_3). In the 1st write, W_1 is directly written into the SRAM (i.e. $W'_1 = W_1$) as shown in Fig. 3(a). After that, W_2 is bit-wise compared to W_1 and the difference (i.e. W'_2) is written into the SRAM. For example, the first bits (highlighted in green in Fig. 3(a)) are identical and thus $W'_{2,1}$ becomes "0". However, the 9th bits are different (highlighted in yellow in Fig. 3(a)), which makes $W'_{2,9}$ "1". This flow is repeated until the last word of the data block is encoded and written into the SRAM. During read (Fig. 3(b)), W'_1 is read out directly since it is the same as the original word (W_1). After that, W'_2 is read and recover W_2 by comparing W'_2 with W_1. For example, since the first bit of W'_2 is "0", it indicates that $W_{2,1} = W_{1,1}$. However, $W_{2,9}$ is "1" since $W'_{2,9}$ is different from $W_{1,9}$ and $W_{1,9}$ is "0".

If an input block has many similar consecutive words, the encoded words will contain mostly "0"s. For example, images including large backgrounds or objects are likely to have highly biased MSBs. Consequently, the encoded MSBs will contain mostly "0"s. In an extreme example where all words of a block are the same, only the first word stores actual data whereas the rest of the words store "0"s. Fig. 4 summarizes the effect of the proposed scheme using a database of 200 images [10]. In the original database, the average percentage of "0"s is 49.5%. It becomes 75.8% after applying the proposed encoding scheme. Therefore, the proposed encoding scheme can reduce the

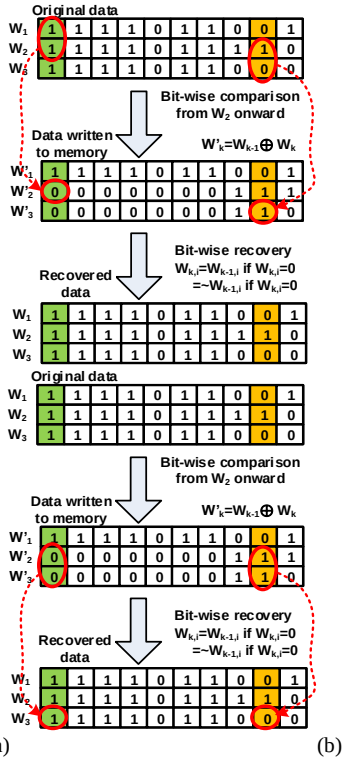

Fig. 3. Proposed column-based data encoding scheme: (a) write operation and (b) recovery operation through decoding.

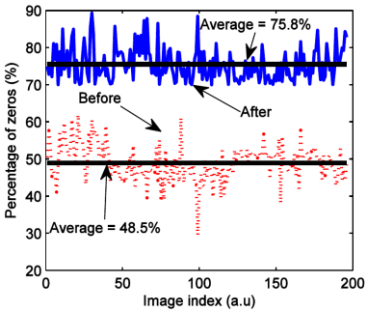

Fig. 4. Simulated percentage of "0s" using a database of 200 images in [10].

overall write power when combined with the data-aware WBL pre-charge scheme. Note that the proposed encoding scheme is beneficial when the original data is biased to either "1" or "0" enough to generate a large number of "0"s after encoding. If the original data is relatively random, the encoded data will include more "1"s than "0"s, which consumes more power. Therefore, this scheme is more beneficial to MSBs while LSBs can still employ normal SRAM write/read operations. Implementation of column-based data encoding/decoding

The proposed column-based data encoding and decoding circuits are shown in Fig. 5. They support two operation modes, the normal mode (ME = "0") and the sequential mode (ME = "1"). In the normal mode, ME selects the original data D_{inT} and transfer it to the WBL driver (Fig. 5(a)). During read, D_{norm} directly coming from the read bitline (RBL) is selected as the final output D_{out}. In this mode, no data encoding occurs and the SRAM behaves as a conventional 8T SRAM. In the sequential mode, $D_{seq.}$ is selected by ME as write data and transferred to the WBL driver. The encoding circuit uses two DFFs to store the current (D_{inT}) and the previous ($D_{in(T-1)}$) input data, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2019.2958668, IEEE Transactions on Circuits and Systems II: Express Briefs

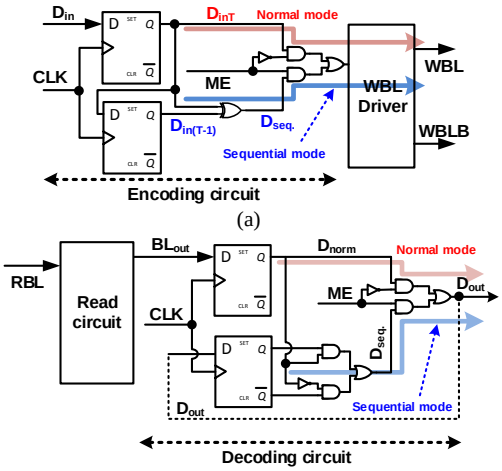

Fig. 5. (a) Column-based data encoding circuit and (b) decoding circuit [16].

computes $D_{seq.}$ through a XOR gate. Thus, if $D_{inT} = D_{in(T-1)}$, $D_{seq.}$ becomes "0" and is written into WBL. If $D_{inT} \neq D_{in(T-1)}$, "1" is written into WBL. During read, the decoding circuit compares the current RBL output (D_{norm}) with the previous RBL output (D_{out}) to retrieve the original data (D_{seq}) as shown in Fig. 5(b). Note that the first data is not encoded in the sequential mode. Thus, ME is only activated from the second word onwards. This can be done easily by delaying the mode enable signal by one clock cycle using a DFF. The area overhead of the proposed data encoding/decoding is only dependent on the data bus width for the proposed encoding and decoding. Therefore, it is insignificant at higher memory density. In this implementation, the area and power overheads are less than 1% and 3%, respectively when compared with the conventional scheme.

E. Bitline leakage compensation and data-aware read

RBL sensing margin reduces substantially at ultra-low voltage operation due to the exponential degradation in the on-current to off-current ratio of the read port. A large number of cells in a RBL increases RBL leakage, which affects RBL sensing substantially [4]. Fig. 6 illustrates the impact of supply voltage scaling on RBL sensing using two worst cases (Fig. 6(a)). It can be observed that RBL sensing is successful at 0.4 V while it fails at 0.25 V due to the negative sensing margin (Fig. 6(b)). Various techniques have been proposed to tackle the RBL leakage issue, including our controller-based optimal bias voltage generator for compensating for RBL leakage [11]. However, the controller-based techniques require periodic calibration for tracking the variations in the operating condition, and lead to relatively large power and area overheads.

In this work, we propose a simpler technique as described in Fig. 7. Unlike the technique in [11] using two replica columns with worst case data patterns for reading "0" and "1", and additional control circuits, the proposed scheme uses only one dummy column containing "1"s and "0"s equally to estimate the average RBL leakage (I_{leak}). The dummy column has one diode-connected PMOS transistor for converting the compensating leakage into bias voltage (V_{bias}). The size of the PMOS transistor is determined carefully so that I_{leak} can hold RBL at a "high" level during reading "0" and RBL can be discharged to a "low" level when reading "1". The dummy column automatically tracks the changes in I_{leak} caused by

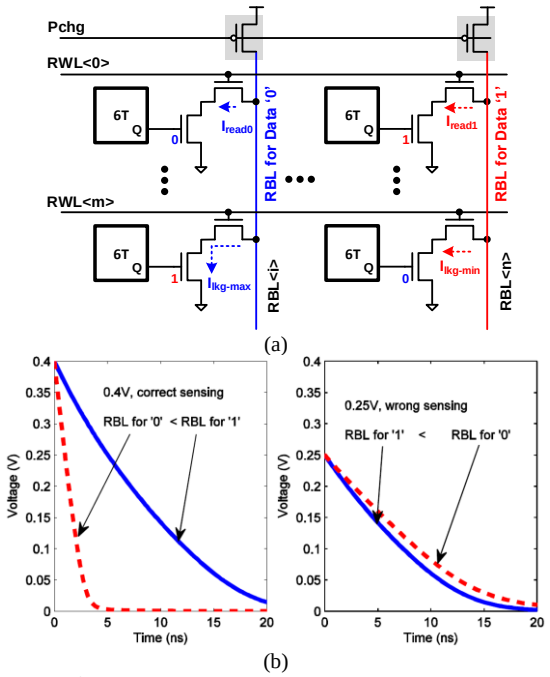

Fig. 6. Impact of bitline leakage on bitline sensing: (a) worst case scenarios for bitline sensing and (b) simulated bitline waveforms at 0.4 V and 0.25 V.

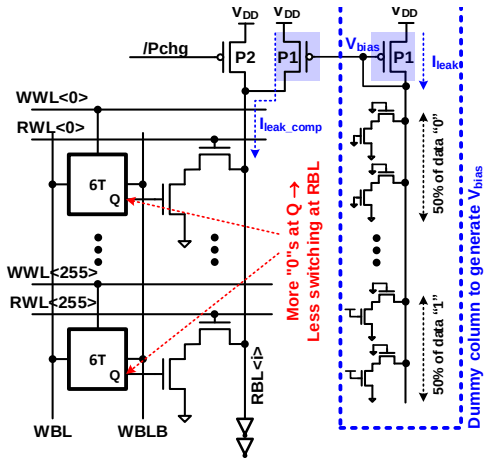

Fig. 7. Proposed RBL read circuit with a PVT-tracking biasing voltage.

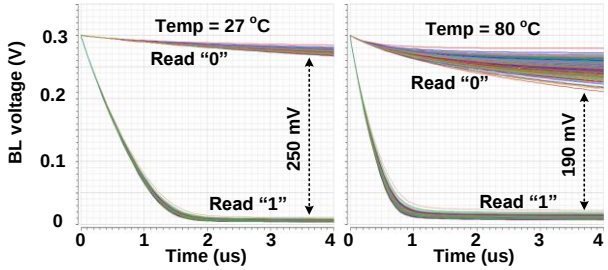

Fig. 8. 1k Monte-Carlo simulation waveforms of RBL discharge when reading "0" and "1" at 27 °C and 80 °C [16].

process, temperature, and voltage variations. I_{leak} is mirrored and used in the main array to partially compensate for the actual RBL leakage current. Fig. 8 presents Monte-Carlo simulation results showing that the proposed leakage compensation technique achieves the sensing margin of 190 mV at 0.3 V. Note that RBL for "0" is kept near V_{DD} because of the pull-up I_{leak} while RBL for "1" is discharged close to ground. Compared to the prior art [11], it is simpler, consumes less power, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2019.2958668, IEEE Transactions on Circuits and Systems II: Express Briefs

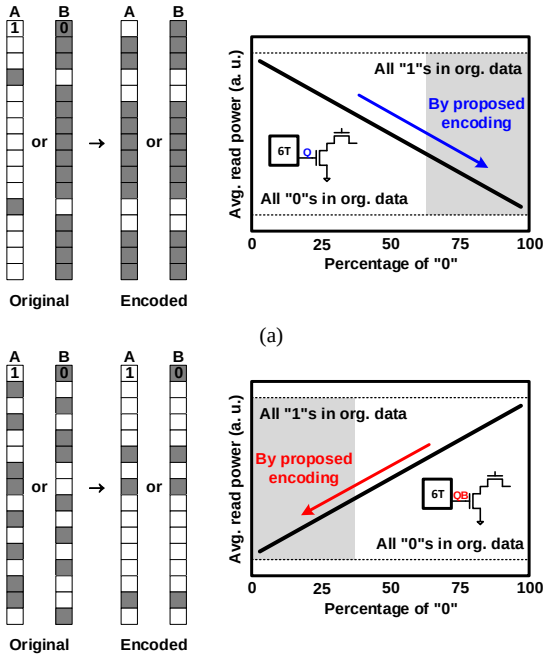

Fig. 9. Effect of original data on average read power per column: (a) when the original data is biased and (b) when the original data is random.

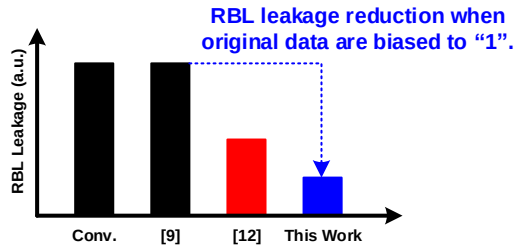

Fig. 10. RBL leakage in state-of-the-art SRAMs. The actual reduction depends on the degree of the data bias.

occupies less silicon area while delivering similar amount of sensing margins .

In addition to write power, RBL switching is another major dynamic power component in SRAMs. As explained in II-C, the proposed column-based data encoding increases the probability of "0" in the column data. Thus, the read port of the employed 8T SRAM cell can be configured in the way of minimizing the RBL switching probability. Therefore, the internal node Q of each cell is connected to the gate of the pull-down transistor of the read port as depicted in Fig. 8. Since the majority of the cells in a column store "0", RBL remains at a "high" level in most of the read operations. This reduces the RBL swi tching and consequently reduces the read power. If the original column data is highly random, the number of "1"s in the encoded write data will be much more than 50%. In this case, the read port can be connected to QB instead of Q, which is more likely to occur in the columns storing LSBs. Fig. 9 summarizes the above two cases.

This read port configuration also reduces RBL leakage, which is bebeficial when reading "1". Fig. 10 illustrates the RBL leakage in several state-of-the-art SRAMs and this work when the original data is biased. In [11], the number of "1"s is around 50% of the cell count in each column by the randomization technique. However, the proposed encoding scheme decreases the number of "1"s below 50%, showing the lowest RBL leakage among compared works. Like the overall write power, the actual RBL leakage and the overall read power saving depends on how much the original column data are biased.

III. Test Chip Measurement Results

A 32 Kb (256×128) SRAM test chip was fabricated in 65 nm CMOS technology. Fig. 11(a) shows the test chip microphotograph. The proposed techniques have the area overhead less than 1%, which will be smaller at higher density. The power overhead of the added circuits is ~3% of the conventional SRAM power, which is also negligible compared to the saved power. Sample consecutive read/write waveforms confirming that the successful SRAM operation at 0.36 V is presented in Fig. 11(b).

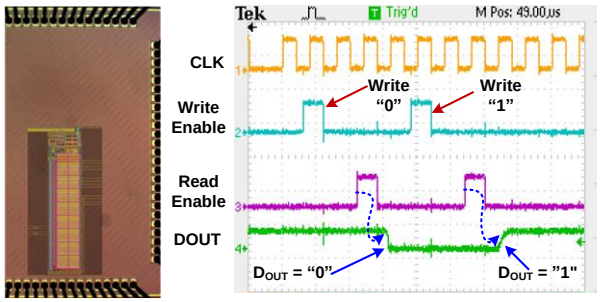

Fig. 11. (a) test chip die photo and (b) sample measured consecutive write and read waveforms of the proposed SRAM at 0.36 V and room temperature [16].

Fig. 12 summarizes the maximum operating frequencies, the average power, and the average energy at different supply voltages, using 20 randomly chosen image data from the 200 images in [10]. The average number of "0"s from those images is ~80%. The maximum frequency of the SRAM at 0,36 V is 250kHz (Fig. 12(a)). Note that the speed degradation compared to our previous version [11] is negligible since both the bitline leakage and the compensation current decrease. The proposed scheme can be disabled if faster access is required, which will lose the energy reduction shown in Fig. 12(c). The SRAM consumes only 0.37 µW at 0.36 V with the leakage power of 0.15 µW (Fig. 12(b)). The minimum energy of the SRAM is only 0.3 pJ/access at 0.5 V (Fig. 12(c)). Compared to our previous version in [11] where the minimum energy point is 0.4 V, this work forms the minimum energy point at 0.5 V. This is because of the significant dynamic power reduction achieved by the proposed techniques. Fig. 12(c) also compares the energy of the proposed SRAM with the conventional SRAM. It can be seen that the proposed SRAM saves energy significantly compared to one operating in the normal mode.

Table I summarize the performance of the proposed design in comparison with prior arts. While it is difficult to directly compare all factors between all designs, it is clear that the proposed design offers the lowest access energy with 9.2 aJ/b, which is about 3× lower than our previous implementation using similar process node. Note that the speed degradation compared to our previous version [11] is negligible since both the bitline leakage and the compensation current decrease. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSII.2019.2958668, IEEE Transactions on Circuits and Systems II: Express Briefs

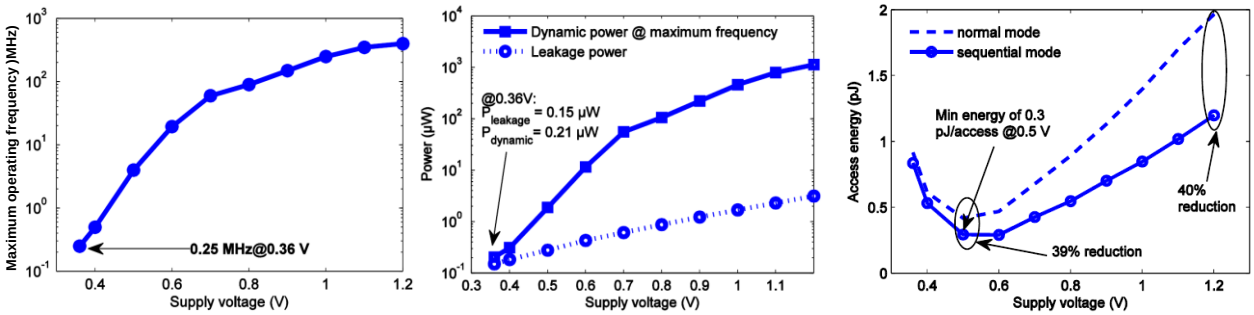

Fig. 12. Measurement results: (a) maximum operating frequency at room temperature, (b) measured dynamic and leakage power in sequential mode with 80% of data "0" at room temperature, and (c) measured energy per access [16].

TABLE I: COMPARISON WITH PRIOR ARTS

	A-SSCC2012 [12]	TVLSI2008 [13]	TCSVT2011 [8]	JSSC2014 [9]	JSSC2016 [11]	**This work**
Technology	65 nm	90 nm	65 nm	65 nm	65 nm	**65 nm**
Density	128 kb	68 kb	32 kb	32 kb	32 Kb	**32 Kb**
Cell structure	8T	8T	6T + 8T	10T	8T	**8T**
Cell size	N.A	3.15 × 0.76 µm^2	N.A.	N.A.	2.6 × 0.52 µm^2	**3 × 0.52 µm^2**
V_{DDmin}	0.37 V	1.0 V	0.6 V	0.52 V	0.2 V	**0.36 V**
Access time or Op. Freq.	N.A	100 MHz	N.A.	N.A.	2.5 µs (0.2 V)	**250 KHz (0.36 V)**
Normalized Energy	162 aJ/b	219.8 aJ/s	20.6 aJ/b	N.A.	31 aJ/b	**9.2 aJ/b**
Data-aware technique	No	Majority logic	Hybrid cells	Prediction	Randomization	**Encoding**

proposed scheme can be disabled if faster access is required, which will lose the energy reduction shown in Fig. 12(c). It is also obvious that this work consumes less energy compared to [11] regardless of the supply voltage due to the data encoded towards bitline leakage minimization.

IV. CONCLUSIONS

This paper presents an ultra-low power SRAM for bio-signal and image/video processing systems. The column-based data encoding scheme transforms the original data with high spatial correlations into differential data so that "0"s are dominant in the memory. The proposed novel WBL pre-charge scheme and the data-aware read scheme further reduce the write power and the read power. The SRAM test chip achieves the minimum energy of 0.3 pJ/access at 0.5 V, which is 40% lower when compared to the normal mode.

REFERENCES

[1] A. Wang and A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 310-319, Jan. 2005.

[2] M. Yamaoka, N. Maeda, Y. Shinozaki, Y. Shimazaki, K. Nii, S. Shimada, K. Yanagisawa, and T. Kawahara, "90-nm process-variation adaptive embedded SRAM modules with power-line-floating write technique," *IEEE Journal of Solid-State Circuits*, vol. 41, pp. 705-711, Mar. 2006.

[3] L. Chang, R. K. Montoye, Y. Nakamura, K. A. Batson, R. J. Eickemeyer, R. H. Dennard, W. Haensch, and D. Jamsek, "An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 956-963, Apr. 2008.

[4] T.-H. Kim, J. Liu, and C. H. Kim, "A voltage scalable 0.26V, 64kb 8T SRAM with Vmin lowering techniques and deep sleep mode," in *Proc. IEEE Custom Integrated Circuits Conference*, pp. 407-410, 2008.

[5] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation," *IEEE Journal of Solid-State Circuits*, vol. 42, pp. 680-688, Mar. 2007.

[6] B.-D. Yang and L.-S. Kim, "A low-power SRAM using hierarchical bit line and local sense amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1366-1376, June 2005.

[7] N. Verma and A. P. Chandrakasan, "A 65nm 8T Sub-Vt SRAM Employing Sense-Amplifier Redundancy," in *Proc. IEEE International Solid-State Circuits Conference*, pp. 328-606, 2007.

[8] I. J. Chang, D. Mohapatra, and K. Roy, "A Priority-Based 6T/8T Hybrid SRAM Architecture for Aggressive Voltage Scaling in Video Applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, pp. 101-112, Feb. 2011.

[9] M. E. Sinangil and A. P. Chandrakasan, "Application-Specific SRAM Design Using Output Prediction to Reduce Bit-Line Switching Activity and Statistically Gated Sense Amplifiers for Up to 1.9 Lower Energy/Access," *IEEE Journal of Solid-State Circuits*, vol. 49, pp. 107-117, Jan. 2014.

[10] Available: http://decsai.ugr.es/cvg/dbimagenes/

[11] A. T. Do, Z. C. Lee, B. Wang, I. J. Chang, X. Liu, and T. T. H. Kim, "0.2 V 8T SRAM With PVT-Aware Bitline Sensing and Column-Based Data Randomization," *IEEE Journal of Solid-State Circuits*, vol. 51, pp. 1487-1498, June 2016.

[12] Y. Sinangil and A. P. Chandrakasan, "An embedded energy monitoring circuit for a 128kbit SRAM with body-biased sense-amplifiers," in *Proc. IEEE Asian Solid-State Circuits Conference*, pp. 69-72, 2012.

[13] H. Fujiwara *et al.*, "Novel video memory reduces 45% of bitline power using majority logic and data-bit reordering," *IEEE Trans. on VLSI Systems*, vol. 16, pp. 620–627, Jun. 2008.

[14] M.Imani, S. Patil, and T. S. Rosing, "Approximate Computing Using Multiple-Access Single-Charge Associate Memory," *IEEE Trans. On Emerging Topics in Computing*, vol. 6, pp. 305-316, Mar. 2018

[15] K. Honda, K. Miyaji, S. Tanakamaru, S. Miyano, and K. Takeuchi, "Elimination of half select disturb in 8T-SRAM by local injected electron asymmetric pass gate transistor," in *Proc. IEEE Custom Integrated Circuits Conference*, pp. 1-4, Sept. 2010.

[16] A. T. Do, S. M. A. Zeinolabedin and T. Kim, "A 0.3 pW/Access 8T Data-Aware SRAM Utilizing Column-based Data Encoding for Ultra-Low Power Applications," in *Proc. IEEE Asian Solid-State Circuits Conference*, pp. 173-176, Nov. 2016