A Selective Bit Dropping and Encoding Co-Strategy in Image Processing for Low-Power Design in DRAM and SRAM

IEEE Publication Technology, *Staff, IEEE,*

***Abstract*—A novel and efficient way of image processing is proposed in this paper, which fully exploits the features of DRAM (Dynamic Random Access Memory) and SRAM (Static Random Access Memory) as well as the human visual system. The proposed strategy first approximates and encodes the image to effectively reduce the number of bit-'1' in the original pixel data, then the processed data is pushed into the off-chip DRAM, and later written into the on-chip SRAM for further computation. Since the storage power consumption of DRAM is proportional to the number of bit-'1', while the write power consumption of SRAM is linear relative to the switch probability and the square of the supply voltage, fewer bits-'1' in the processed pixel data will decrease the power consumption of DRAM and SRAM, yet accompanied by negligible influence on output quality as our proposed method has the advantages of both approximate computation and error compensation. Thus, a tradeoff is finally achieved between storage power consumption and output quality. In the experimental simulations, 39.8% power reduction for DRAM and 25.9% write power reduction for SRAM have been achieved. Regarding output quality, Discrete Cosine Transform (DCT), quantization, inverse quantization and inverse DCT (IDCT) are employed to process the approximated data. The simulations shows 3.36 dB losses in Peak-Signal-Noise-Ratio(PSNR). Based on this strategy, an approach of priority-based reduction in supply voltage for insignificant pixel data is introduced to achieve further reduction in power consumption for SRAM. Undoubtedly, a lower supply voltage will increase the probability of read errors from SRAM. However, with our proposed approximate coding strategy, the output quality is barely impacted by the lower supply voltage.**

***Index Terms*—Approximate storage, embedded system, DRAM, SRAM, image processing, low power design.**

I. Introduction

With the large-scale application and development of new computer vision and artificial intelligence in recent years, applications like high-definition images and videos are gradually being ported to embedded devices such as mobile phones, tablets and wearable devices. As a result, there are higher demands on the computation and storage of large-scale data, especially in storage as it accounts for more than 90% of the power consumption [1], [2] in common image and video applications. The storage power consumption is generally divided into two parts: off-chip storage DRAM and on-chip storage SRAM. In typical image processing, as shown in Fig. 1, the raw image data will first be sent to off-chip storage DRAM, followed by on-chip SRAM for further computation, and some outputs of the computation will be sent back to DRAM. Due to battery capacity and the slow development of new storage materials, reducing the power consumption of DRAM and SRAM storage in existing embedded systems is a serious concern.

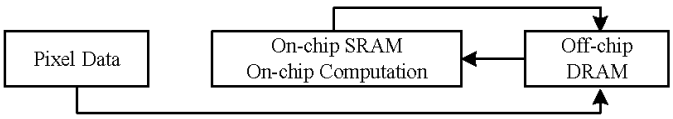

Fig. 1. Dataflow of typical image processing.

Lowering the amount of image data is one of the normal ideas for reducing storage power. The most well-known approach is image compression [3], which implements a compression algorithm to decrease the amount of image data before transferred and stored in the off-chip storage DRAM. However, this approach has two major defects: firstly, compression algorithm is complex to implement, such as adaptive bit-width compression [4], a dictionary-based fixed length coding scheme [5], especially considering the power consumption of embedded devices is limited; Secondly, the system modification and overhead required to integrate the compression algorithm into the existing image and video processing system are significant.

A different idea is to build on existing algorithms to reduce power consumption by changing the storage structure, or by preprocessing the raw image data. Some of the common solutions include data encoding, data dropping and supply voltage reduction, etc. In contrast to precisely computed circuit systems, image and video processing takes advantage of the fact that the human visual system is more sensitive to low frequency than high frequency [6], [7]. For example, during the common lossy decompression of JPEG images, the human eye is comparably less sensitive in distinguishing the results of output when the PSNR is above 30 dB. In other words, a completed accurate storage method is not necessary in practical applications. Therefore, the data in storage system can be divided into important and non-important parts depending on the perceptual sensitivity of the human eye. More specifically, for a pixel value, the information in the high-bit part is more crucial than those in the low-bit part [8]. Based on this feature, through exploring the characteristics of DRAM and SRAM, the pixel data can be selectively approximated or encoded in advance in order to provide a corresponding solution to reduce storage power consumption.

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

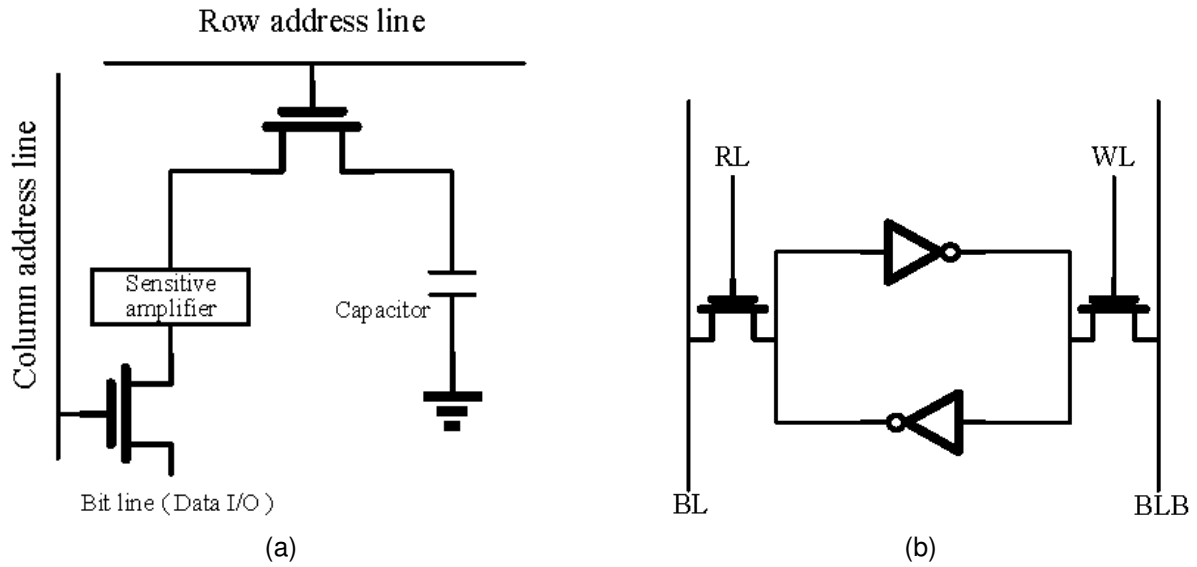

Fig. 2. Circuit scheme of bitcell in DRAM and SRAM. (a) circuit scheme of one bitcell in DRAM. (b) circuit scheme of one bitcell in SRAM.

This paper proposed a method for preprocessing raw image data with selective bit dropping and approximate encoding strategies, as well as an approach of priority-based reduction in supply voltage for insignificant pixel data to achieve further reduction in power consumption. The proposed data could preserve the most significant information of the image data effectively, and has a limited impact on output quality. A tradeoff has been achieved between power savings and output quality as 39.8% power reduction for DRAM and 25.9% write power reduction for SRAM are shown in the extensive simulations, with only 3.36 dB losses in PSNR compared to accurate processing. The technique for lowering supply voltage for insignificant pixel data can be used in addition to the strategy to reduce storage power consumption. The results from the subsequent experimental data demonstrated that it has a negligible effect on output quality.

The structure of the paper is as follows: Section II gives an introduction of SRAM and DRAM. The proposed strategy is described in Section III. Section IV presents the experiment results. The paper is concluded in the final section.

II. Power Analysis of SRAM and DRAM

Both DRAM and SRAM account for a large percentage of the power overhead in circuit systems, where the read/write and refresh operations consume the most of the power consumption for DRAM and SRAM. Fig. 2(a) [9] shows the basic cell of a single-bit DRAM, which is composed of three basic modules: two NMOS transistors, a sensitive amplifier and a capacitor to store data. When the read/write operation is performed, the row address line and column address line are first activated to “1”, and the NMOS transistors are turned on. The sensitive amplifier is then used to connect the bit line to the capacitor. When the data is “1” or “0”, the capacitor will be charged and discharged according to the read/write operation. It should be noted that when the data is stored as “1”, the charge stored on the capacitor will deplete over time due to the capacitor’s leakage current, so the entire DRAM needs to be dynamically refreshed in regular intervals to replenish the leaked charge to ensure the accuracy of the stored data. While no charge replenishment is required during the refresh process if the original stored data is “0”, i.e. no additional dynamic power consumption is generated when the data is stored as “0”. Therefore, as stated in [9], [10], the storage power consumption of DRAM is proportional to the number of bit-‘1’. In order to reduce the storage power of DRAM, Dr. Song et al. [11] proposed a multi-level refresh frequency for approximate storage. Likewise, the authors kept the refresh frequency of the significant bits constant and reduced the refresh frequency of the insignificant bits to varying degrees based on the principle that the high bits of the original image data cannot be contaminated and the low bits can be approximated. The main problem with this approach is the need to modify the original DRAM refresh control system. It is impossible to ignore the additional overhead caused by this modification process. Moreover, approximate storage with complete bit dropping method for original stored data to decrease the power consumption of off-chip DRAM is a common approach. Nevertheless, complete bit dropping method could generate excessive loss of raw data information and require real-time dynamic monitoring of the output quality [12]. In this context, data encoding from [13] proposed an efficient approach to reducing the number of bit-‘1’ for DRAM storage. However, bandwidth utilization is a key problem due to additional flags and information.

Compared with DRAM, the storage process of SRAM is relatively simple. The power consumption of SRAM can be divided into three parts: write power, read power and leakage power. Fig. 2(b) [14] shows the basic cell of a single-bit SRAM. This single-bit memory structure consists of 6 MOS transistors (6T structure), where the two inverters in the middle constitute the positive feedback while the two NMOS transistors on both sides serve as the write and read interfaces of the memory cell. When the write operation is performed,

the WL and RL are first activated to "1", and the two NMOS transistors are turned on. Through the BL, BLB and the two NMOS transistors, the input data will be written into the inverters. Two situations occur in this case. If the written data is opposite to the currently saved data, the intermediate inverters with positive feedback loop will be forcibly flipped to the desired value, during which both inverters are charged, discharged and consume energy. The write operation will not generate flipped power if the data value being written matches the data value kept in the prior state. As for read power and leakage power, [14], [15] pointed out that the value of write power is 3.3X larger than that of read power, and the leakage power consumption is only a small fraction of SRAM power dissipation. In summary, data write power is the main source of power consumption for on-chip SRAM. More specifically, most of the power consumption occurs when the transistors of the bit-cell in SRAM switch between the on and off state. Therefore, lots of research has been done in the past to reduce write power consumption. For the on-chip memory cell circuit, the energy overhead is proportional to several parameters as shown in Eq. (1):

$$\boldsymbol{E} \propto \boldsymbol{\alpha C V_{DD}^2 f} \tag{1}$$

Where $\boldsymbol{E}$ is the energy overhead, $\boldsymbol{\alpha}$ is the switch probability (switches between bit-'1' and bit-'0'), $\boldsymbol{C}$ is the effective capacitance, $\boldsymbol{V_{DD}}$ is the supply voltage and $\boldsymbol{f}$ is the operating frequency [16]. In a practical circuit design, the circuit equivalent capacitance is mainly related to the circuit structure, and it is extremely complicated to improve from the design. Although reducing the operating frequency may seem effective for large-scale data operations, this will reduce the circuit's performance as there is no guarantee of real-time data processing in the circuit system [17], which is unacceptable in practical image processing applications. Since energy has a squared relationship with supply voltage, reducing the voltage is an effective method [18]. However, when reducing the supply voltage, the circuit's ability to run at a high frequency decreases and the susceptibility to circuit noise increases. The circuit will have timing errors if the operating frequency is forced to be constant. In short, lowering the supply voltage may result in an incorrect circuit's logic output. A mixed approximate storage scheme can therefore be adopted in light of the fact that the high bits of the original image data cannot be contaminated and the low bits can be approximated. The normal supply voltage can be used for the important part of the pixel data, while the lower supply voltage can be applied for non-important part. Moreover, since SRAM write power consumption is proportional to the switch probability, reducing the $\boldsymbol{\alpha}$ value in Eq. (1) is also considered an efficient way to achieve power consumption.

From the analysis above, reducing the number of '1' is critical for saving DRAM power consumption, whereas decreasing the switch probability $\boldsymbol{\alpha}$ in Eq. (1) is an effective approach that can also be utilized jointly with a mixed voltage approximate storage scheme for saving power consumption of SRAM.

III. Co-Strategy of Selective Bit Dropping and Encoding with Supply Voltage Reduction

As previously indicated, the power consumption of DRAM could be effectively reduced if most of the bits stored in the original image are bit-'0'. More crucially, the probability of switching in the entire SRAM write operation is simultaneously reduced when the image data with this feature is cached from off-chip DRAM to on-chip SRAM for further computation. Therefore, SRAM's storage power consumption can also be effectively reduced. Thus, we can focus on the preprocessing of original image data and no more design overhead or large modification to the circuit system are needed. The encoding and complete bit dropping mentioned above are two effective methods to reduce the number of bit-'1' in the raw image data. However, the encoding method introduces additional flags, reducing the bandwidth availability of off-chip storage significantly. Meanwhile, complete bit dropping method leads to excessive data losses. In this paper, we try to optimize both methods and combine the results to obtain our approximate storage method. The proposed method is as follows:

Based on the fault tolerance of image video, the last bit of a pixel value (8 bits per pixel as illustrated) is used as the encoding flag bit, which means that we do not introduce extra bit-flags. Our initial focus is on the high seven bits of the pixel data. From high to low order, the high seven bits are designated as bit_7, bit_6, bit_5, bit_4, bit_3, bit_2, and bit_1. These seven bits are divided into two parts: the high part, which is from bit_7 to bit_{8-K}, and the low part, which is from bit_{7-K} to bit_1, $K \in [1, 7]$. K value denotes the number of pixel bits in the high part. In the case of K = 7, the number of pixel bits in the low part is 0. When K = 1, the number of pixel bits in the high part is 1, and the number of pixel bits in the low part is 6.

After capturing the raw data of the image, as the low bits in the pixel data are insignificant, for the low part bits, which is from bit_{7-K} to bit_1 ($K \in [1, 7]$), we proposed an approximate compensation dropping method to keep the first bit- '1' from high order to low order in bit_{7-K} to bit_1, and then set the other bits to zero. Unlike the complete dropping method, this method retains the highest bit with logical value '1' and provides various degrees of error compensation for different degrees of data approximation. When the number of dropped bits is higher, this method will be more advantageous. Since more bits in low part of the pixel will be set to zero, the power consumption of DRAM could be reduced effectively. Moreover, the switch probability $\boldsymbol{\alpha}$ in Eq. (1) could also be reduce to achieve further power savings in SRAM.

Simultaneously, as high bits are important in pixel data, the original data must be stored accurately. A flip encoding method is proposed here to reduce the number of bit-'1' in the high part. Fig. 3 shows the proposed encoding scheme. First, the number of bit- '1' in the high part is counted. When the number of bit- '1' is greater than K/2, all the bits in the high part are flipped, and the flag bit(the last bit of the pixel as we described before) is enforced as '0'. When the number of bit- '1' is smaller than K/2, the data in the high part remains

unchanged with the flag bit will be set as '1'. Using the last bit as the flag information has no effect on the bit width of the data. Instead, it can effectively reduce the number of bit-'1' in the high part. The flag bit is utilized to determine whether to flip the high part during the decoding operation. If it is '0', then flip it; if it is '1', keep it constant. This simple-to-implement and easy-to-integrate decoding operation does not modify the current image decoding system, making it an efficient method to reduce the number of bit-'1'.

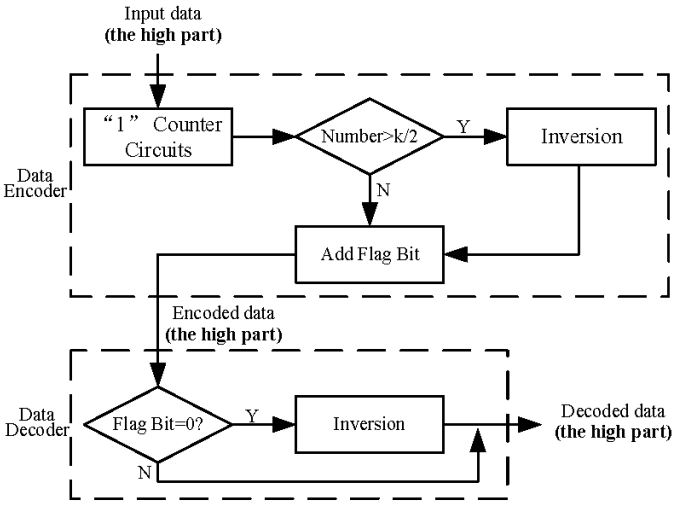

Fig. 3. Encoding scheme for the high part.

As illustrated in Fig. 4, with K = 4 as an example, the high part contains four bits from bit_7 to bit_4 and the low part contains three bits from bit_3 to bit_1. As shown in the diagram, if the number of bit-'1' is greater than 2, it will be flipped and the corresponding flag bit is set to '0'. If the number of '1' is less than or equal to 2, the flip will not occur and the corresponding flag bit will be set to '1'. As for the low part, it is clear that the data after the first bit-'1' position is cleared.

It can be seen that with smaller value K, more bits will be assigned to the low part, which means that more bits could be set to bit-'0' to achieve power savings. Meanwhile, the larger amount of bit-'0' in the low part of the pixel data indicates that more information will be lost in the image, resulting in poor output quality. However, it should be noted that the output quality requirements for image processing during real-world applications are various. As shown in Fig. 5, the designer can set an output quality threshold for a specific image application. For example, PSNR is a common output quality evaluation index that could adjust the size of the K value to determine whether the output quality threshold is satisfied. The minimum value of K is obtained if the output quality threshold is satisfied, so as to obtain the minimum storage power consumption. It should be noted that if the output K value of Fig. 5 is 8, which is outside the range of K, the output quality requirement cannot be met when K = 7. At this point, accurate processing should be taken and no data preprocessing can be done. If the output K value is 0, which is also outside the range of K, it is clear from the above analysis that the number of bits in the high part is 0, the

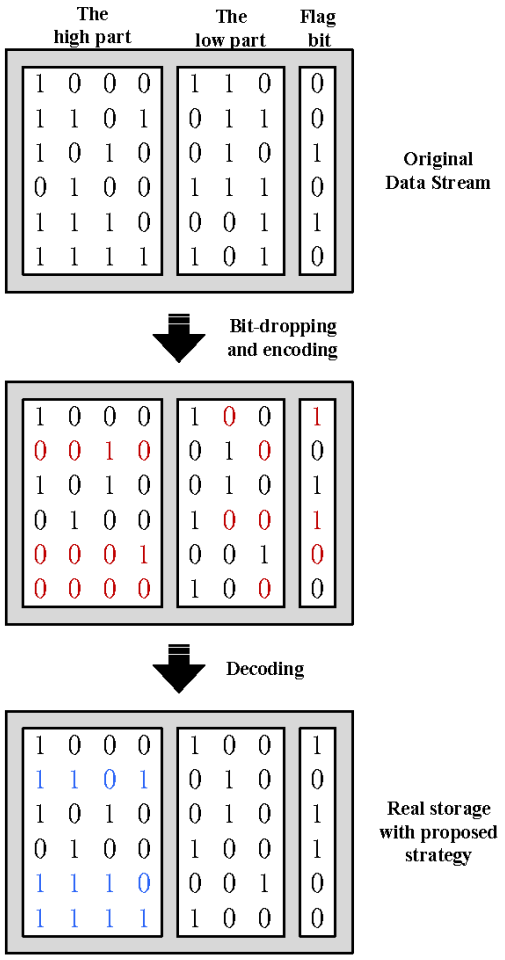

Fig. 4. Illustration of the proposed strategy.

encoding scheme will be ignored, and as a result the resources of the flag bit are wasted.

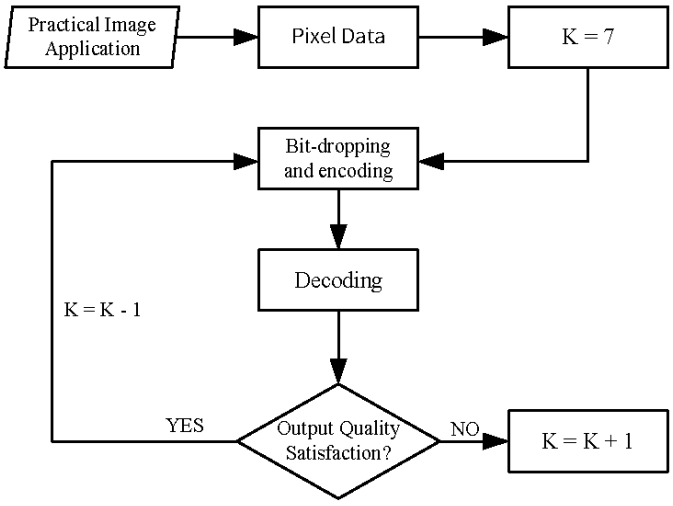

Fig. 5. Proposed data analysis method.

Based on the above analysis, it is clear that the proposed method could achieve the tradeoff between storage power consumption and output quality while being more efficient than

the approaches previously mentioned. As the method focuses solely on image data preprocessing, the storage structure is not modified and the method is simple to implement, requiring no system modifications. More importantly, the strategy can also be utilized in conjunction with an approximate mixed voltage storage scheme. The read errors from SRAM due to the reduced supply voltage could "recover" the missing pixel information. The result will be demonstrated in later simulations.

IV. Experimental Simulations

In this section, the proposed strategy is modeled through C++. As shown in Fig. 6, the pixel data are preprocessed with different approximations using different K values. According to the typical data flow process in image processing applications, the processed pixel data is pushed into DRAM and SRAM, where the number of bit-'1' per pixel for storage in DRAM and the switch probability for storage in SRAM are simulated and analyzed. In terms of output quality, the preprocessed pixel data is first decoded before being pushed into the simulation of the common JPEG and MPEG-2 codec core algorithms, and the PSNR value of the final output image is used as a metric for output quality.

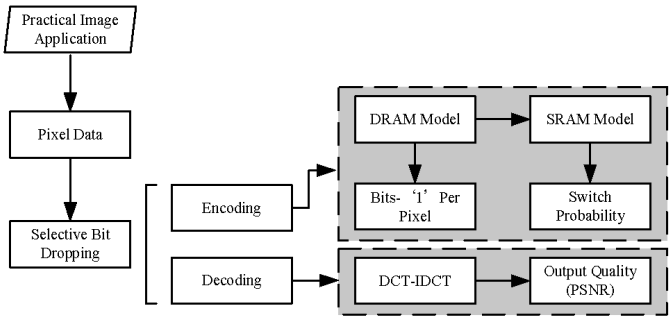

Fig. 6. Flow chart for experimental simulation.

A. Simulation results for the number of bit-'1' in DRAM and the switch probability in SRAM

Since the image preprocessing does not change the off-chip and on-chip memory systems and image processing algorithms, it can be easily integrated into the front end of the image processing algorithm. The image preprocessing algorithm including selective bit dropping and encoding has been implemented in C++. After preprocessing with different values of K, real image data featuring a resolution of 256*256 is sent to the DRAM model, and the simulation result of storage power consumption in DRAM is shown in Fig. 7. The measure of storage power consumption is the average number of bit-'1' per pixel Navg. The pixel data without preprocessing, i.e., the accurate storage method, is used as the baseline for comparison with the approximate preprocessing method. It can be seen that Navg in DRAM for yet-to-be-preprocessed pixel data is 4.22, which is close to 50%. When K = 7, only the flag bit (as the encoded flag information) at the end of the pixel data is lost, but Navg decreases to 2.81, reducing the DRAM storage power consumption by 33.4%. This shows the efficiency and effectiveness of the encoding scheme. More notably, if the requirement for the output quality is not so high, the value of K can be reduced even further. As illustrated in the analysis above, the value of Navg decreases significantly as K decreases.

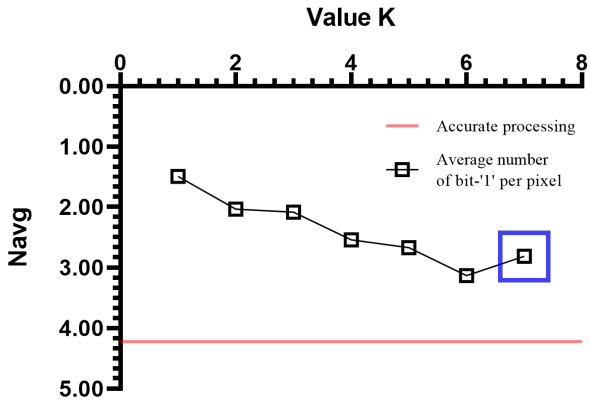

Fig. 7. Average number of bit- '1' per pixel Navg for different value K.

The image data is then sent to the SRAM model for further computation. Due to the high resolution of the image, it is not possible to send all of the data to the on-chip SRAM at once. As the capacity of the common SRAM is very limited and the size of SRAM affects the switch probability, two sizes of SRAM are used, which is 64k-bits and 128k-bits respectively. As shown in Fig. 8, the 128k-bits size SRAM has an overall significantly lower switch probability than that of 64k-bits size SRAM. The effect of distinct K values on the switch probability is essentially the same as that of DRAM, where the switch probability for yet-to-be-preprocessed image data is close to 50%, and decreases with smaller K value. Thus, the storage power consumption of SRAM could be significantly reduced. However, this comes at the expense of output quality losses, and the following part will focus on the impact of image preprocessing on output quality.

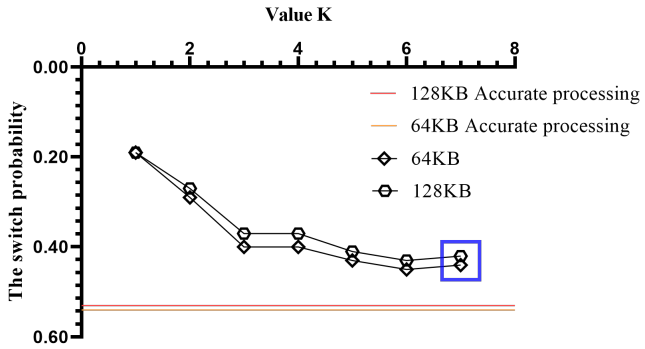

Fig. 8. The switch probability for different value K.

B. The evaluation of output quality with the proposed strategy

After analyzing the power consumption of DRAM and SRAM, the image data is processed with a decoder, which can

be easily implemented into the integrated image processing algorithm. DCT, quantization, inverse quantization and IDCT in C++ are employed to process the approximated data after decoding. For the generality of the experimental results, 1 million images are used for the simulation. The final average PSNR with different K values are shown in Fig. 9, where it can be seen that the output quality decreases with smaller K value. Therefore, according to the requirements of different image applications, dynamic adjustment of off-chip and off-chip storage power consumption can be achieved based on output quality needs. When K = 4, the proposed strategy in the paper has a PSNR loss of around 3.36 dB (compared to the accurate processing), and 39.8% power reduction for DRAM and 25.9% write power reduction for SRAM can be achieved. Additionally, as for the complete dropping method shown in Fig. 9, in which all the bits in the low part are set to be '0'. It can be seen that the complete dropping method has an overall significantly lower PSNR value than that of our proposed strategy. As the K decreases, the output quality decreases sharply, when K = 4, the corresponding PSNR is 33.51 dB, which means that 10% reduction than our proposed strategy. This shows the effectiveness and efficiency of our approximation approach for error compensation. Fig. 10 provides some examples when K = 4, the appearance of the image processing are not significantly different from that of the accurate processing results. However, as K increases, the PSNR value decreases and the image becomes substantially noisier.

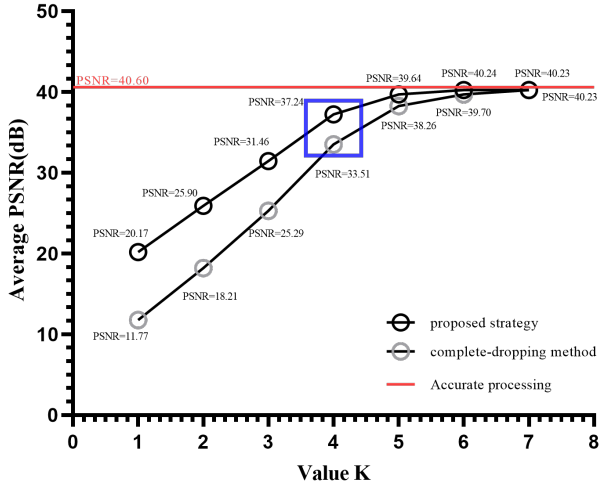

Fig. 9. Average PSNR with different value K.

C. The proposed strategy with a priority-based reduction in supply voltage

According to Eq. (1), the power consumption for SRAM is proportional to the supply voltage. As a result, lots of researches have been done in the past on reducing supply voltage for lower power dissipation. In this section, the proposed strategy is utilized jointly with an approximate mixed voltage storage scheme. More specifically, as shown in Fig. 11, the

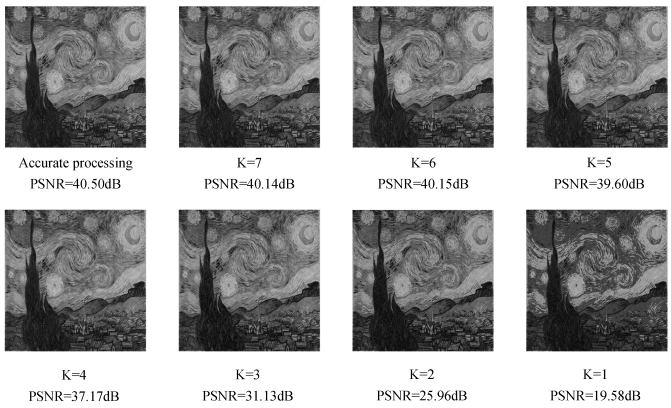

Fig. 10. PSNR value of images with proposed strategy.

normal supply voltage is used for the accuracy of the high part as well as the flag bit, while the lower supply voltage is applied for the approximation of the low part.

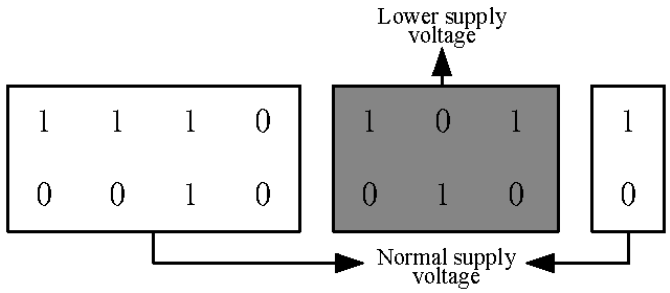

Fig. 11. Selective reduction in supply voltage.

Theoretically, further reduction in power consumption for SRAM is attained. The evaluation of power saving can be expressed as follows:

$$\boldsymbol{\eta} = (1 - \boldsymbol{P}_{\text{low}}/\boldsymbol{P}) * 100\% \tag{2}$$

Where $\boldsymbol{\eta}$ is the evaluation of power saving, $\boldsymbol{P}$ is the energy overhead with normal supply voltage, and $\boldsymbol{P}_{\text{low}}$ is the energy overhead with selectively reduced supply voltage. In the case of K = 4, the normal supply voltage is applied to the high part of 4 bits and the flag bit, while the lower supply voltage is applied to the low part of 3 bits. Thus, $\boldsymbol{P}$ and $\boldsymbol{P}_{\text{low}}$ can be respectively replaced through Eq. (2). Therefore, the evaluation of power saving can be calculated as follows:

$$\begin{aligned}\boldsymbol{\eta} &= [1 - (\frac{5}{8}\boldsymbol{\alpha C V_{DD}^2 f} + \frac{3}{8}\boldsymbol{\alpha C V_{\text{low}}^2 f})/\boldsymbol{\alpha C V_{DD}^2 f}] * 100\% \\ &= (\frac{3}{8} - \frac{3}{8}\frac{\boldsymbol{V}_{\text{low}}^2}{\boldsymbol{V}_{\boldsymbol{DD}}^2}) * 100\% \end{aligned} \tag{3}$$

Where $\boldsymbol{\eta}$ is the evaluation of power saving, $\boldsymbol{V_{DD}}$ is the normal supply voltage, while $\boldsymbol{V_{low}}$ is the lower supply voltage.

Undoubtedly, a lower supply voltage increases the probability of read errors from SRAM. Many researchers conducted in the past have presented the relationship between supply voltage and read error probability, as shown in Fig. 12 [19]. With lower supply voltage, the probability of read error

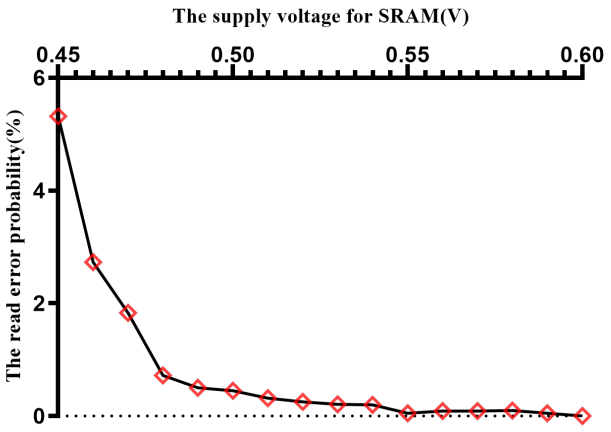

Fig. 12. SRAM approximation error model.

TABLE I
PSNR VALUE OF IMAGES WITH DIFFERENT SUPPLY VOLTAGE

Supply voltage(V)	Read error probability(%)	PSNR(dB)
0.60	0	37.31
0.59	0.05	37.28
0.58	0.10	37.27
0.57	0.09	37.27
0.56	0.09	37.27
0.55	0.05	37.28
0.54	0.20	37.22
0.53	0.21	37.22
0.52	0.25	37.10
0.51	0.32	37.18
0.50	0.45	37.16
0.49	0.50	37.17
0.48	0.72	37.08
0.47	1.83	36.75
0.46	2.73	36.50
0.45	5.32	35.80

increases. Hence, after the introduction of selectively reduction in supply voltage, a stochastic probability model is built in c++ to analyze the impact of selective reduced supply voltage on output quality. As shown in Table I, the output quality PSNR does not decrease linearly with increasing read error probability; instead, an anti-correlation may occur. This is because some pixel information was lost during our approximation process, however, subsequent read errors may "recover" the missing pixel information. Therefore, this strategy not only further reduces the storage power consumption, but also shows surprising results in terms of output quality.

V. Conclusion

In this paper, we presented a selective bit dropping and encoding co-strategy in image processing for low-power in DRAM and SRAM. Based on the characteristic of human visual system, the pixel data is selectively approximated and encoded in advance in order to provide a corresponding solution to reduce storage power consumption for both DRAM and SRAM, and the cost of the output quality and the system modification is negligible. The encoding strategy has shown its effectiveness and efficiency. At the same time, the selective bit dropping strategy archives the improvement on PSNR (compared to the complete dropping method). Our co-strategy decreases the number of bit-'1' in original image data and contributes to the tradeoff between storage power consumption and output quality. More importantly, the approximate mixed voltage storage scheme has also been vertified its effectiveness and efficiency. Therefore, our co-strategy can be utilized in conjunction with other adaptive methods in the future work and we believe the proposed co-strategy could also provide the tradeoff between storage power consumption and output quality in various applications.

References

[1] Y. Chen, X. Yang, F. Qiao, J. Han, Q. Wei, and H. Yang, "A multi-accuracy-level approximate memory architecture based on data significance analysis," in *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2016, pp. 385–390.

[2] D. Culler, J. P. Singh, and A. Gupta, *Parallel computer architecture: a hardware/software approach*. Gulf Professional Publishing, 1999.

[3] W.-Y. Chen, L.-F. Ding, P.-K. Tsung, and L.-G. Chen, "Architecture design of high performance embedded compression for high definition video coding," in *2008 IEEE International Conference on Multimedia and Expo*, 2008, pp. 825–828.

[4] J. Zhu, L. Hou, W. Wu, R. Wang, C. Huang, and J. Li, "High performance synchronous drams controller in h.264 hdtv decoder," in *Proceedings. 7th International Conference on Solid-State and Integrated Circuits Technology, 2004.*, vol. 3, 2004, pp. 1621–1624 vol.3.

[5] H. Gao, F. Qiao, and H. Yang, "Lossless memory reduction and efficient frame storage architecture for hdtv video decoder," in *2008 International Conference on Audio, Language and Image Processing*, 2008, pp. 593–598.

[6] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video processing and communications*. Prentice hall Upper Saddle River, NJ, 2002, vol. 1.

[7] F. Qiao, N. Zhou, Y. Chen, and H. Yang, "Approximate computing in chrominance cache for image/video processing," in *2015 IEEE International Conference on Multimedia Big Data*, 2015, pp. 180–183.

[8] V. K. Chippa, D. Mohapatra, A. Raghunathan, K. Roy, and S. T. Chakradhar, "Scalable effort hardware design: Exploiting algorithmic resilience for energy efficiency," in *Design Automation Conference*, 2010, pp. 555–560.

[9] N. Zhou, F. Qiao, H. Yang, and H. Wang, "Low-power off-chip memory design for video decoder using embedded bus-invert coding," in *2011 Tenth International Symposium on Autonomous Decentralized Systems*, 2011, pp. 251–255.

[10] Y. Joo, Y. Choi, H. Shim, H. G. Lee, K. Kim, and N. Chang, "Energy exploration and reduction of sdram memory systems," in *Proceedings of the 39th Annual Design Automation Conference*, ser. DAC '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 892–897. [Online]. Available: https://doi.org/10.1145/513918.514138

[11] S. Liu, K. Pattabiraman, T. Moscibroda, and B. G. Zorn, "Flikker: Saving dram refresh-power through critical data partitioning," *SIGPLAN Not.*, vol. 46, no. 3, p. 213–224, mar 2011. [Online]. Available: https://doi.org/10.1145/1961296.1950391

[12] J. Miao, "Modeling and synthesis of approximate digital circuits," Ph.D. dissertation, 2014.

[13] H. Shim, Y. Joo, Y. Choi, H. G. Lee, and N. Chang, "Low-energy off-chip sdram memory systems for embedded applications," *ACM Trans. Embed. Comput. Syst.*, vol. 2, no. 1, p. 98–130, feb 2003. [Online]. Available: https://doi.org/10.1145/605459.605464

[14] 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, no. 2, pp. 101–112, 2011.

[15] Z. Liu, Y. Song, M. Shao, S. Li, L. Li, S. Ishiwata, M. Nakagawa, S. Goto, and T. Ikenaga, "A 1.41w h.264/avc real-time encoder soc for hdtv1080p," in *2007 IEEE Symposium on VLSI Circuits*, 2007, pp. 12–13.

[16] V. Gupta, D. Mohapatra, A. Raghunathan, and K. Roy, "Low-power digital signal processing using approximate adders," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 1, pp. 124–137, 2013.

[17] J. Kwon, I. J. Chang, I. Lee, H. Park, and J. Park, "Heterogeneous sram cell sizing for low-power h.264 applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 59, no. 10, pp. 2275–2284, 2012.
[18] N. Gong, S. Jiang, A. Challapalli, S. Fernandes, and R. Sridhar, "Ultra-low voltage split-data-aware embedded sram for mobile video applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 12, pp. 883–887, 2012.
[19] Q. Li, P. Dong, Z. Yu, C. Liu, F. Qiao, Y. Wang, and H. Yang, "Puncturing the memory wall: Joint optimization of network compression with approximate memory for asr application," in *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2021, pp. 505–511.