# A Multi-Accuracy-Level Approximate Memory Architecture Based on Data Significance Analysis

Yuanchang Chen, Xinghua Yang, Fei Qiao*, Jie Han+, Qi Wei and Huazhong Yang
Tsinghua National Laboratory for Information Science and Technology
Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, P. R. China
+Dept. of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada
{yc-chen14,yang-xh11}@mails.tsinghua.edu.cn, jhan8@ualberta.ca+, {qiaofei*,weiqi,yanghz}@tsinghua.edu.cn

***Abstract*—Approximate memory is a promising technology for emerging recognition, mining and vision applications. These applications require the processing of large volumes of data to achieve energy-efficiency with negligible accuracy loss. This paper proposes a multi-level approximate memory architecture based on data significance analysis. In this architecture, a memory array is divided into several separated banks with different predefined accuracy levels. A key novelty of this work is the design of a memory controller that distributes data to the memory banks according to the results of data significance analysis. When applied to a DCT (Discrete Cosine Transform) processing module, the proposed approximate memory controller can achieve over 60% power saving with on-chip memory model of multiple supply voltage SRAM banks, at the cost of a marginal output PSNR (Peak Signal to Noise Ratio) degradation of 3.34 dB.**

***Keywords*-Approximate Memory Architecture; Data Significance Analysis; Low Power Design.**

## I. Introduction

Emerging recognition, mining and vision applications process large volumes of unstructured data, hence they place a significant demand on the memory subsystem of a modern computing platform, especially on the power consumption of such systems [1]. These applications exhibit a property of inherent error-resilience, which can be attributed to several factors: (i) the input data are noisy and redundant; (ii) they employ statistical and iterative computation patterns that possess a self-healing nature; (iii) there are a range of acceptable outputs other than a unique golden output [2]. Approximate memory is a promising technique for achieving energy-efficiency in these applications due to the inherent redundancy in their input and intermediate signals.

Process parameter variations have become a significant issue with the evolution of technology. However, a conventional worst-case design would incur a large power consumption. Relaxing design constraints and allowing appropriate approximation can lead to low power operations with negligible degradation in output quality. Approximate computing allows for approximation in computation as long as output quality meets user requirements. It is thus considered a promising approach for energy-efficient design. In many image processors, memories consume a large portion of total power [3]. It is potentially interesting to exploit approximate techniques in memory to reduce total power consumption.

Approximate memory employs a data-oriented approximation technique. Previous studies of approximate memory design can be divided into several categories, including approximate on-chip and off-chip memory design, approximations in off-chip memory access and approximate storage in emerging devices. A typical approximate on-chip memory design is a priority-based 6T/8T hybrid SRAM, in which significant energy is saved by aggressive voltage scaling [4] [5]. Another study investigates an unreliable embedded DRAM, in which improvement is obtained in memory availability and retention power through relaxation of the worst-case cell criterion [6]. For on-chip memory design, Flikker [7] and Sparkk [8] explores lowering the refresh rate in DRAM to trade off energy at the cost of retention errors. For off-chip memory access, ApproxMA reduces access power consumption by dynamic precision scaling [9]. In [10], spatial correlation of image chrominance is used in an approximate chrominance cache to reduce off-chip memory access. Moreover, a few prior efforts have focused on emerging device to trade off write/read errors for energy, such as a quality configurable spintronic memory [1] and a multi-level PCM (Phase-Change Memory) [11] [12].

There are two related studies that are most relevant to our effort. One is a voltage-scalable and process-variation resilient hybrid architecture using a mixed array of 6T and 8T SRAM bit cells [4] [5]. It is implemented in a preferential storage policy that the higher order luminance bits are stored in robust 8T bit cells, while the lower order luminance are stored in conventional 6T bit cells. In this work, 6T and 8T bit cells are integrated together to form the hybrid array. Since the 8T bit cell has two word lines, the reading and writing timing sequence and layout of the hybrid array must be altered accordingly. This hybrid array is only applicable for image luminance data and its precision is not configurable. This feature prevents other data from using this memory in different tasks of the processor. The other relevant effort is a quality-configurable STT-MRAM

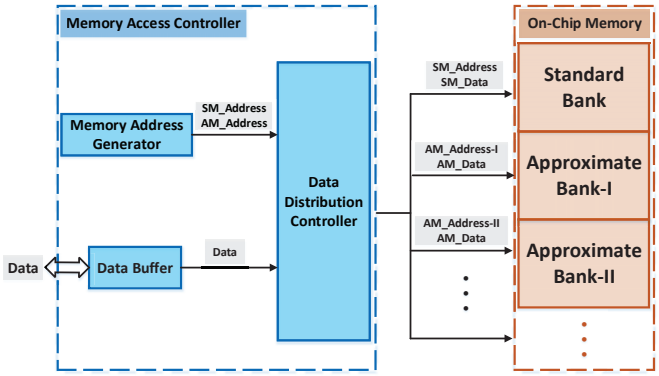


Figure 1. Multi-Accuracy-Level Approximate Memory Architecture.

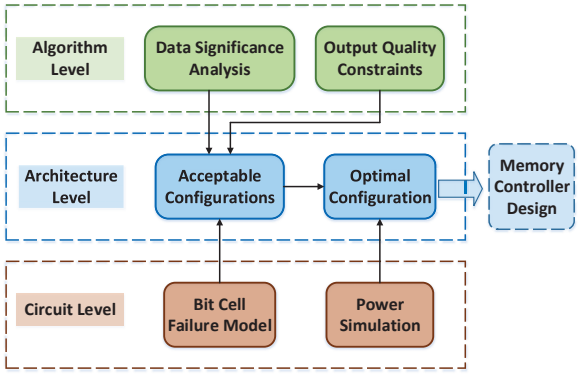


Figure 2. Design Flow Based on Proposed Architecture.

(Spin Transfer Torque Magnetic Random Access Memory) array, in which data can be stored to various levels of accuracy based on the application requirement [1]. In this quality-configurable memory, each read and write operation can be performed at different predefined levels of accuracy. Read/write quality and configurability can be achieved by some peripherals to modulate the bit cell read current or write duration. This mechanism of precision configuration is based on the characteristics of STT-MRAM bit cell. However, it cannot be applied to other memories such as CMOS based SRAM and embedded DRAM.

Both of these studies develop a device-to-architecture modeling framework, which is closely related to the characteristics of devices. However, a memory is a data-related device. It is more interesting to exploit designs based on a data significance analysis. In this paper, we propose a multi-accuracy-level approximate memory architecture based on processing data analysis. Memory cell arrays in the proposed architecture are divided into several separated banks at different predefined accuracy levels, thus data can be stored at various levels of accuracy. The key innovation of this work is a memory controller that can distribute data to these banks according to the results of data significance analysis. We subsequently validate the effectiveness of the proposed architecture with multiple supply voltage SRAM banks to obtain an energy-efficient design for DCT (Discrete Cosine Transformation) processing.

The rest of this paper is organized as follows. Section II presents the framework of a multi-level approximate memory architecture based on data significance analysis. The design methodology for the proposed architecture is described in Section III. In Section IV, we validate the effectiveness of proposed architecture with SRAM banks using multiple supply voltage to obtain an energy-efficiency design for DCT processing. Finally, Section V concludes the paper.

## II. Proposed Architecture

The proposed multi-accuracy-level approximate memory architecture is shown in Figure 1. The memory cell array in the proposed architecture is divided into several separated banks, including a standard memory bank and some approximate memory banks. Sensitive data will be stored in the standard memory bank, while resilient data will be stored in the approximate memory banks at different accuracy levels. The control operations are conducted by a memory controller, which consists of a memory address generator, a data buffer and a data distribution controller. The address generator is used to generate addresses mapped to specific accuracy levels, while the data distribution controller splits a multi-bit data into several parts and allocates them to different banks based on precision requirements.

The proposed architecture differs from previous designs in two aspects. (i) The memory cell array in the architecture is divided into several separate banks, thus each bank can be independently designed. Different banks are grouped together by some interface circuitry, so the structure of underlying circuits do not need to be changed. (ii) The memory banks at various accuracy levels utilize a detailed mapping from an address to a precision level, thus a fine-grained precision configuration can be implemented by the address distribution. The extent of approximation can be controlled by the memory controller, which is designed according to the data significance analysis at the algorithm level. It is worth mentioning that this mechanism of precision configuration is available for almost all digital memories.

## III. Design Methodology

We develop a modeling framework from both algorithm and device levels to the architecture level. Figure 2 shows a cross-layer design methodology, including the designs at algorithm, architecture and circuit levels. In the early design stage, a data significance analysis is conducted to distinguish between sensitive and resilient data; it is determined what degree those resilient data can be approximated to. Then the failure probability model and power consumption of

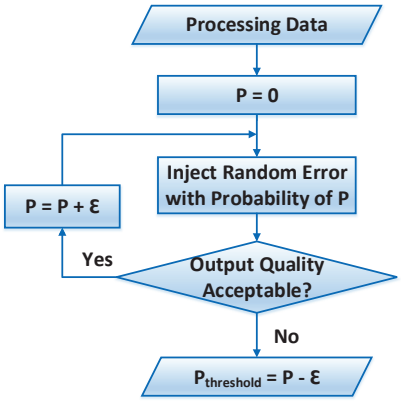


Figure 3. Data Significance Analysis for Massive Regular Data.

memory bit cells can be obtained by a series of circuit level simulations. According to the results of data significance and failure probability analysis of memory bit cells, we can obtain some configurations with acceptable output quality. The performance of each acceptable design can be estimated rapidly by power simulation and the most energy-efficient configuration among those acceptable designs can be identified. Accordingly, an approximate memory controller is then designed for the most energy-efficient configuration.

### A. Data Significance Analysis

Most applications require the use of both sensitive data that cannot be processed with approximation and resilient data that may be approximated to some degree. Some highly sensitive data, including instruction sets and control logic, are of most significance. No error can be tolerated for these highly sensitive data. However, it is a great challenge to distinguish between sensitive and resilient data and to determine to what degree those resilient data can be approximated. A data significance analysis may be one of the solutions for this challenge. There are two major methods for data significance analysis, which are discussed in detail next.

For massive regular data such as those for image luminance and chrominance, an effective method is to inject faults with an error probability to each bit. This is based on the premise that failure bits are distributed uniformly in approximate memories. If the output quality is acceptable, then increase the injected error probability until the output quality falls below the user requirements. Otherwise, decrease the error probability to find out the threshold of error probability that can be tolerated at these resilient data. Figure 3 shows the overview of a data significance analysis for massive regular data. For some irregular data such as those for spectrum components, it is not suitable to inject evenly-distributed errors. A fine-grained analysis is necessary for these irregular bits. By flipping only one bit at a time and observing its effect on output quality, we can figure out which bit is more significant according to the output quality degradation. It is an useful way to compare the significance between two bits.

### B. Memory Cell Failure Model and Power Simulation

COMS based memories, including SRAM and embedded DRAM, have served the industry for several decades. In recent years, emerging memories, such as STT-MRAM, PCM and RRAM (Resistive Random Access Memory), also gain extensive attention because of their potential as future on-chip memories. The failure mechanisms are usually different for different types of memories. For example, SRAM and embedded DRAM bit cells experience possible failures under process variations. PCM and other non-volatile memories work by storing an analog value and quantizing it to a digital value. The same pulse may lead to a different state in these non-volatile memories when applied to a different cell or to the same cell at a different time, which may result in read and write failures. In general, we assume that failure bits are uniformly distributed in approximate memories and a failure probability can be used to describe the reliability of an approximate memory array.

On the other hand, a fast and reliable estimation of power consumption is imperative to evaluate the performance of a design at the application level. Therefore, it is necessary to perform power simulations (including static, read and write power simulation) of memory bit cells.

### C. Storage Scheme

A storage scheme (or memory configuration), is usually a guide for memory designers. Storage details are clearly specified, including how many banks would be grouped together in the multi-accuracy-level approximate memory, the capacity and address range of each bank, and the storage location of each bit in the application. In the proposed architecture, the memory controller is designed for the optimal storage scheme (or memory configuration) to achieve the most energy-efficient design. Therefore, how to obtain the optimal configuration is crucial. It is largely an optimization problem and we need to determine the optimization target and constraints according to the application requirements. The solution of the optimization problem will result in the optimal configuration.

## IV. Simulation Results

To show the effectiveness of the proposed methodology, a design example of DCT processing with multiple supply voltage SRAM banks is considered.

### A. Data Significance Analysis

DCT is frequently used in signal and image processing, especially in lossy compression. In this section, we analyze the data significance of image luminance in DCT processing and use PSNR (Peak Signal to Noise Ratio) as a metric

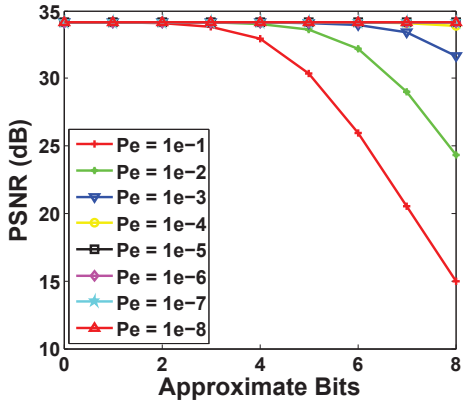


Figure 4. Data Significance Analysis of Luminance Bits.

of output quality. It is assumed that image luminance can be approximated to some extent and it is stored in an approximate memory before DCT processing. The PSNR between the original and reconstructed images is used as a metric to evaluate the degradation of output quality. In our simulation using MATLAB, the reconstructed image is obtained in 4 steps, including DCT, quantization, inverse quantization and IDCT (Inverse Discrete Cosine Transform).

There are two premises about our simulation: (i) Failure bits are uniformly distributed in approximate memories; (ii) MSBs (More Significant Bits) contribute more to output quality than LSBs (Less Significant Bits) [4]. In this simulation, input data (image luminance) are divided into 2 sections: MSBs without approximation and LSBs with a degree of approximation. A range of failure probabilities are considered in the fault injection into specific LSBs for analyzing the data significance of image luminance. Figure 4 shows the variation of output quality (PSNR) with different error injection rates (from $10^{-8}$ to $10^{-1}$). It is obvious that image luminance bits are not sensitive data. They can be approximated to some extent, thus it is suitable for them to be stored in an approximate memory. We can observe that the output quality is acceptable (PSNR>30 dB) when the injected error probability is less than $10^{-3}$, even if all the luminance bits are stored in approximate banks. For an error probability of $10^{-1}$, up to 5 LSBs of image luminance can be stored in approximate banks to obtain an acceptable output quality.

### B. Memory Cell Failure Model and Power Simulation

SRAM is the most commonly used on-chip memory, which utilizes bistable latching circuitry to store a bit [13]. The structure of a standard 6T-SRAM bit cell is presented in Figure 5.

Voltage over-scaling (VOS) has been known as an effective approach for lowering power dissipation due to the quadratic dependence of dynamic power on supply voltage [14] [15]. Unfortunately, memory failure probability

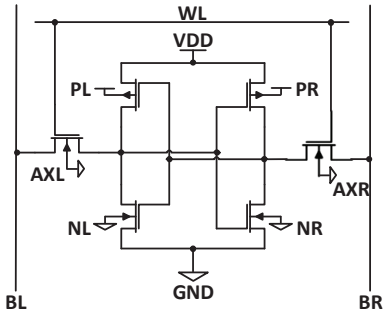


Figure 5. SRAM Bit Cell.

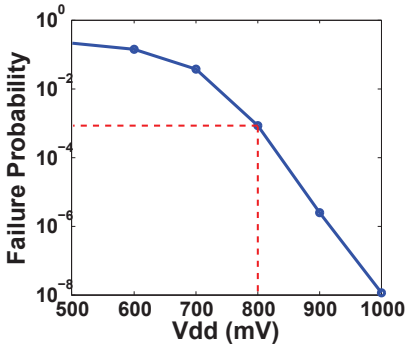


Figure 6. Failure Probability vs. Vdd of a Single 6T SRAM Cell.

increases considerably as supply voltage is scaled down, which imposes a lower bound on the supply voltage [16]. Approximate memory is an effective solution, and it implies that we can further scale down the operating voltage of some memory banks to achieve more significant energy-efficiency gains. In this subsection, a 6T-SRAM bit-cell is implemented in 65nm CMOS process. The failure probability and power consumption of the SRAM arrays with different supply voltage are obtained by HSPICE simulations.

*1) Failure Probability in 6T SRAM Cell:* Since failure probability is used to describe the reliability of an approximate memory, we consider the bit cell failure model of 6T-SRAM in [17] and obtain a quick estimate of SRAM failure probability. Figure 6 shows the failure probability of SRAM bit cells with different supply voltages (from 0.5 V to 1.0 V). We observe that the failure probability decreases when supply voltage (Vdd) increases. Given a supply voltage (Vdd) higher than 800mV, the failure probability of 6T-SRAM bit cells is less than $10^{-3}$. Considering the results in Figures 4 and 6, all the image luminance bits can be stored in an SRAM array with a supply voltage equal to or higher than 800 mV. In contrast, the SRAM array experiences a quite significant failure probability increase over $10^{-1}$ when the supply voltage is lower than 600 mV.

*2) Power:* Power simulation is required to obtain a fast and reliable energy consumption estimate of a specific design. In this subsection, simulations are performed to

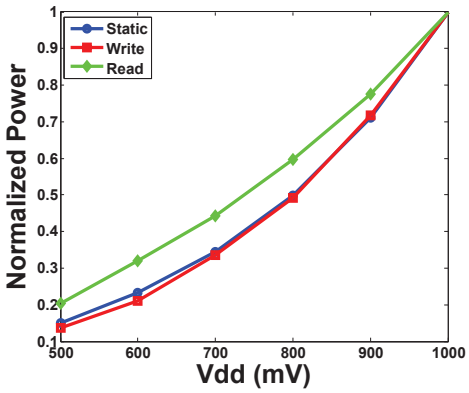


Figure 7. Normalized Power with Vdd of 6T SRAM Cell.

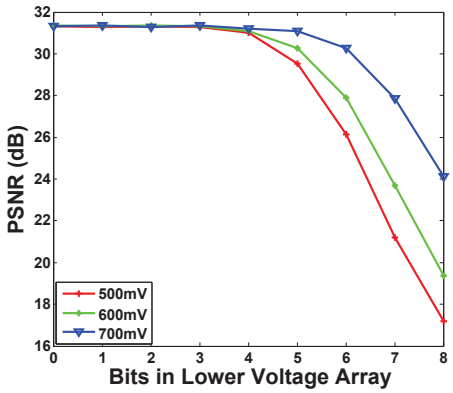


Figure 8. Output Quality (PSNR) with Different Approximate Memory Configurations.

obtain read, write and static power of an SRAM bit cell with different supply voltages (from 0.5 V to 1.0 V). Figure 7 shows the normalized power of an SRAM bit cell, in which read, write and static power increase sharply with supply voltage (Vdd). At a supply voltage of 800 mV, the SRAM array consumes about 40% less power compared to the case when the supply voltage is 1V. When the supply voltage scales down to 500 mV, it achieves more than 80% power reduction. The results of power simulations indicate that it is useful to scale down the supply voltage of SRAM array to achieve an energy-efficient design.

### C. Storage Scheme

Considering data significance and failure probability simultaneously, it is clear that all the luminance bits can be stored in an approximate SRAM array at 800 mV, because the SRAM array experiences an error probability less than $10^{-3}$ at the supply voltage of 800 mV. To make full use of the fault-tolerant capacity of these luminance bits and achieve maximum energy-efficiency, we adopt a dual-accuracy-level approximate storage scheme in which MSBs are stored in the SRAM array at 800 mV while LSBs are

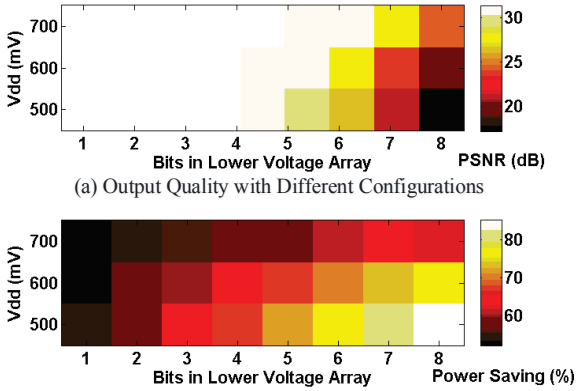


Figure 9. Output Quality and Power Saving with Different Approximate Memory Configurations. In the above configurations, MSBs are stored in an SRAM array at 800 mV, LSBs are stored in a lower voltage array.

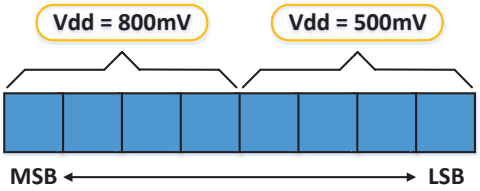


Figure 10. The Most Energy-Efficient Scheme for DCT Processing.

stored in the SRAM array at a lower supply voltage. Figure 8 shows the variation of output quality (PSNR) with various storage schemes. Both the variation of output quality and power saving with different storage schemes can be found in Figure 9. It is generally believed that an acceptable output quality is equivalent to the criteria that PSNR is higher than 30 dB. Among those acceptable schemes (when PSNR>30 dB), the most energy-efficient storage scheme is presented in Figure 10. In this scheme, 4 MSBs of image luminance are stored in an SRAM array at 800 mV, while 4 LSBs are stored in an SRAM array at 500 mV.

In the design of our multi-accuracy-level approximate memory, an SRAM array is divided into two banks: bank I at the supply voltage of 800 mV and bank II at the supply voltage of 500 mV. The memory controller is designed to fit this storage scheme. A memory address generator in the controller will generate two types of addresses for the two aforementioned banks. The data distribution controller will split an 8-bit luminance data into two parts, then save the 4 MSBs to bank I and the other 4 bits to bank II.

Simulation results show that we can achieve significant power saving with the proposed storage scheme, up to 60.02%, 68.56% and 67.52% power reduction for read, write and static power consumption respectively. Figure 11 shows the output quality of DCT processing with precise memory and the proposed multiple voltage SRAM. As shown in the figure, the output quality of our design is acceptable


Figure 11. Comparison of DCT Processing Output Quality.

compared with the standard output quality. In spite of aggressive voltage over-scaling, the PSNR of the latter is smaller than the former by 3.34 dB.

## V. Conclusion

Approximate memory is a promising technology for emerging recognition, mining and vision applications that process large volumes of data to achieve energy-efficient design with negligible accuracy loss. In this paper, a multi-accuracy-level approximate memory architecture based on data significance analysis is proposed. In this architecture, a memory array is divided into several separated banks at different predefined accuracy levels. The key innovation of this work is a memory controller that distributes data to these banks according to the results of data significance analysis. The simulation result based on DCT processing shows that, with SRAM banks using multiple supply voltages, the proposed design can achieve over 60% power saving compared to a standard SRAM array at the cost of 3.34 dB output quality (PSNR) degradation.

## References

[1] A. Ranjan, S. Venkataramani, X. Fong, K. Roy, and A. Raghunathan, "Approximate storage for energy efficient spintronic memories," in *52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2015, pp. 1–6.

[2] V. K. Chippa, S. T. Chakradhar, K. Roy, and A. Raghunathan, "Analysis and characterization of inherent application resilience for approximate computing," in *50th ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2013, pp. 1–9.

[3] C.-P. Lin, P.-C. Tseng, Y.-T. Chiu, S.-S. Lin, C.-C. Cheng, H.-C. Fang, W.-M. Chao, and L.-G. Chen, "A 5mw mpeg4 sp encoder with 2d bandwidth-sharing motion estimation for mobile applications," in *IEEE International Solid State Circuits Conference (ISSCC)*, 2006, pp. 1626–1635.

[4] I. J. Chang, D. Mohapatra, and K. Roy, "A voltage-scalable & process variation resilient hybrid sram architecture for mpeg-4 video processors," in *46th ACM/IEEE Design Automation Conference (DAC)*, 2009, pp. 670–675.

[5] C. I. Joon, M. Debabrata, and R. Kaushik, "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.

[6] S. Ganapathy, A. S. Teman, R. Giterman, A. P. Burg, and G. Karakonstantis, "Approximate computing with unreliable dynamic memories," in *International New Circuits And Systems Conference (NEWCAS)*, 2015.

[7] S. Liu, K. Pattabiraman, T. Moscibroda, and B. G. Zorn, "Flikker: saving dram refresh-power through critical data partitioning," *ACM SIGPLAN Notices*, vol. 47, no. 4, pp. 213–224, 2012.

[8] J. Lucas, M. Alvarez-Mesa, M. Andersch, and B. Juurlink, "Sparkk: Quality-scalable approximate storage in dram," in *The Memory Forum*, 2014, pp. 1–9.

[9] Y. Tian, Q. Zhang, T. Wang, F. Yuan, and Q. Xu, "Approxma: Approximate memory access for dynamic precision scaling," in *Proceedings of the 25th edition on Great Lakes Symposium on VLSI*. ACM, 2015, pp. 337–342.

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

[11] J. Nelson, A. Sampson, and L. Ceze, "Dense approximate storage in phase-change memory," *ASPLOS Ideas & Perspectives*, 2011.

[12] A. Sampson, J. Nelson, K. Strauss, and L. Ceze, "Approximate storage in solid-state memories," *ACM Transactions on Computer Systems (TOCS)*, vol. 32, no. 3, p. 9, 2014.

[13] S. Skorobogatov, "Low temperature data remanence in static ram," *University of Cambridge Computer Laborary Technical Report*, vol. 536, p. 11, 2002.

[14] D. Mohapatra, G. Karakonstantis, and K. Roy, "Significance driven computation: a voltage-scalable, variation-aware, quality-tuning motion estimator," in *Proceedings of the ACM/IEEE international symposium on Low power electronics and design (ISLPED)*, 2009, pp. 195–200.

[15] P. K. Krause and I. Polian, "Adaptive voltage over-scaling for resilient applications," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2011, pp. 1–6.

[16] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Statistical design and optimization of sram cell for yield enhancement," in *IEEE/ACM International conference on Computer-aided design (ICCAD)*, 2004, pp. 10–13.

[17] M. Saibal, M. Hamid, and R. Kaushik, "Modeling of failure probability and statistical design of sram array for yield enhancement in nanoscaled cmos," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 12, pp. 1859–1880, 2005.