Approximate Flash Storage: A Feasibility Study

Amir Rahmati, Matthew Hicks, Atul Prakash
University of Michigan
{rahmati,mdhicks,aprakash}@umich.edu

Abstract

Approximate storage is an essential part of any approximate computing architecture. In this work, we provide a platform for conducting experiments on flash storage targeted at approximate computing applications. Using this platform, we perform exploratory experiments that show the potential of flash memory as an approximate storage. Our results uncover the correlation in reliability between memory cells in the same sector and show that flash memory can operate with minimal error, acceptable for many approximate applications, at 65% of its original voltage.

1. Introduction

Flash technology is seeing rapid adoption across all tiers of computer systems ranging from data centers to home computers, embedded devices, and sensors [8]. Many of these devices run applications that do not require precise results and can benefit from trading some of their precision with energy saving, performance gain, or increased hardware lifetime [1]. In recent years, *approximate computing* has created a new branch of research in computer architecture research that examines these trade-offs.

The potential benefits of using approximate volatile memory has been studied both in SRAM [2] and DRAM [3, 5] technologies. Sampson et al. [7] studied the concept of approximate non-volatile memory. They proposed a mechanism based on reducing the number of programming pulses, and wear out management to increase performance and lifetime of PCMs and broached the applicability of their method for flash memories. Tseng et al. [11] examined the effects of undervolting flash memory and designed an energy saving scheme that adjusts voltage to the minimum value which allows error free functionality.

Motivated by these works and the growing ubiquity of flash storage, we examine the feasibility of using external flash memory as an approximate storage by lowering the supply voltage to the flash chips during its operations. To achieve this, we design and implement an approximate flash storage platform that allows researchers to study effects of undervolting on functionality of flash storage drives and design schemes that can take advantage of its characteristics. Using this platform we perform a series of initial experiments that evaluate

(i) Effect of lowering input voltage on successful writes.

(ii) Correlation between cell position and its resistance to low voltage writes.

(iii) Effect of write iterations on success of low voltage writes in flash memory.

Our results verify the feasibility of using flash memory as an approximate storage and motivate future work on developing approximate storage systems.

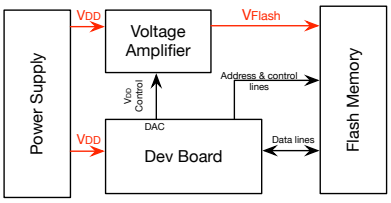

Figure 1. Flash experimental platform. The development board controls input voltage of flash through use of an operational amplifier while running experiments.

2. Flash Experimental Platform

To enable research on approximate storage systems, we developed an approximate flash storage platform. Figure 1 presents the block diagram of our setup. We use an EFM32 Wonder Gecko Starter Kit [9] to control a SST39SF040 flash chip [4]. The flash chip cells are arranged in $128 \times 4KB$ sectors with $1\frac{Byte}{address}$. The development board controls a $TLV4110$ operational amplifier [10] that allows us to have fine-grained control over input voltage of flash chip.

We developed libraries for communication between the development board and flash chip that implements full range of write, read, and erase functions available on the flash chip and allows us to run all operations in low voltage. We are releasing these libraries along with our design and pin diagrams to promote future research.[1]

3. Running Flash at Low Voltage

The most accessible approach for gaining power saving in current flash storage systems is to decrease the input voltage to the chip. To assess the effect of undervolting in flash memories, we performed a series of experiments ranging from $2.5V$ to $5V$ with steps of $10mV$. At each voltage, we performed a full write on the flash memory and recorded the cells that were successfully written. Each experiment was preceded by a chip erase to ensure minimal effect on the chip from previous experiments.

Figure 2 presents the cumulative distribution function of successful writes across different voltages. At $2.90V$, more than 90% of the cells are successfully written and at $2.94V$ (65% of the minimum recommended operational voltage of $4.5V$) all but 12 cells are successfully written in our flash chip. These results open up the possibility of an approximate storage system which can trade-off between data precision and energy consumption based on the application and power availability.

To study the potential correlations in location of memory cells and their minimum write voltage, we recorded the minimum volt-

[1] Artifacts will become available upon presentation of the work.

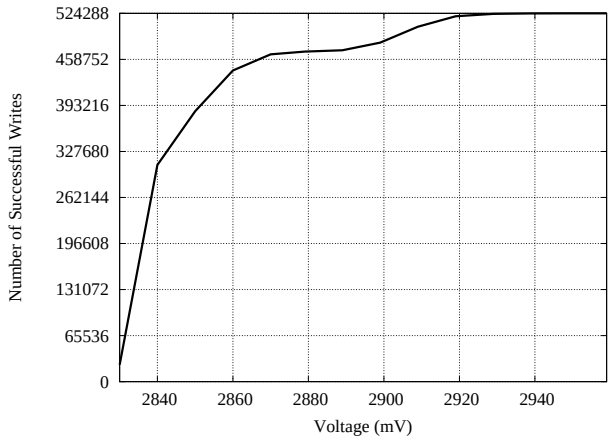

Figure 2. CDF of number of successful writes per voltage. All but 12 cells are successfully written at $2.94V$ (65% of minimum recommended operational voltage of $4.5V$)

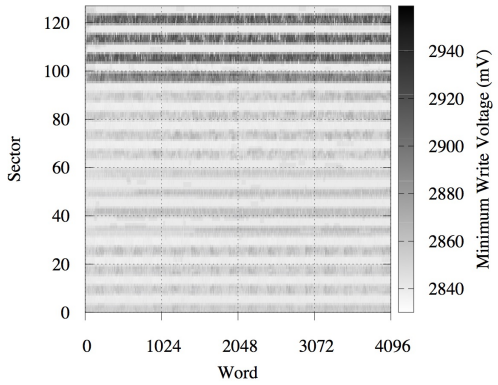

Figure 3. Heat map of minimum write voltage for the flash chip. Minimum write voltage of cells in a sector closely match each other.

age at which each memory address is successfully written. Figure 3 presents the heat map of these values across the memory chip. Our results show similarity in minimum write voltage on cells in the same memory sector. These results have interesting implications for approximate systems as these systems can potentially assign a reliability value to each of their sectors and adjust their write locations in accordance with their precision requirement, energy availability, and memory reliability profile.

4. Effect of Iteration

Multiple low voltage writes in a cell is an approach for reducing the number of errors when writing at low voltage. This approach was first introduced by Salajegheh et al. [6] and used on internal flash of embedded devices as an energy saving mechanism. We explored the effect of multiple writes on correctness of data in approximate flash by conducting 100 writes on each memory address at low voltage. After each write, a read operation was performed to check for success or failure of the write.

Figure 4 presents the effect of multiple writes on number of successful writes in a sector of memory across different voltages.

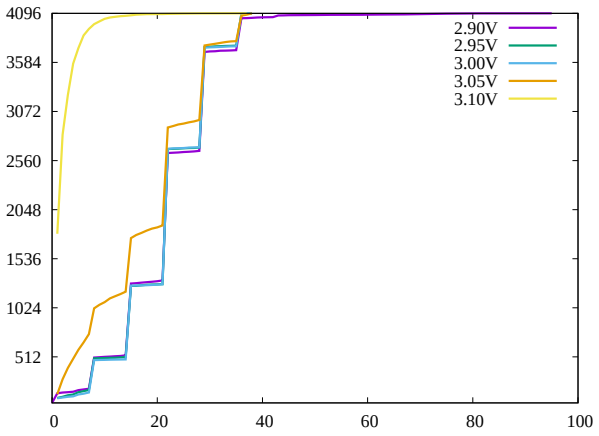

Figure 4. CDF of number of successful writes per different number of iterations for a sector of flash memory.

Increased number of iterations directly correlates with number of successful writes in the memory sector. Similarly, as the write voltage increases, the number of successful writes at a lower iteration also increases. This effect appear to happen in distinctive steps at lower voltage.

5. Conclusion & Future Directions

In this paper, we provided blueprint for an experimental flash storage platform that can be used to design and test approximate storage schemes on real hardware. Using this platform, we performed initial experiments examining effect of write at low voltage and write iterations on correctness of written values in flash memory. Our results showed that an approximate flash memory can perform with minimal error at 65% of its recommended voltage.

Although our initial results motivates the use of flash memory as approximate storage, more extensive experiments need to be performed to model the potentials and limitations of this approach. Effect of additional factors such as temperature and wear out also needs to be analyzed to create a clear picture about the system. Based on these results, we motivate the design of an approximate storage system that provides maximum energy saving and performance gain in an approximate computing architecture.

Acknowledgments

The authors would like to thank Earlence Fernandes and Justin Paupore for their feedbacks on the project.

References

[1] H. Esmaeilzadeh, A. Sampson, L. Ceze, and D. Burger. Architecture support for disciplined approximate programming. In *ACM SIGPLAN Notices*, 2012.

[2] A. Kumar, J. Rabaey, and K. Ramchandran. SRAM supply voltage scaling: A reliability perspective. In *International Symposium on Quality of Electronic Design (ISQED)*, 2009.

[3] S. Liu, K. Pattabiraman, T. Moscibroda, and B. G. Zorn. Flikker: saving DRAM refresh-power through critical data partitioning. *ACM SIGPLAN Notices*, 2012.

[4] Microchip Technology. 4 Mbit (x8) Multi-Purpose Flash - SST39SF040. `http://ww1.microchip.com/downloads/en/DeviceDoc/25022B.pdf`.

[5] A. Rahmati, M. Hicks, D. Holcomb, and K. Fu. Refreshing thoughts on DRAM: Power saving vs. data integrity. In *Workshop on Approximate Computing Across the System Stack (WACAS)*, 2014.

[6] M. Salajegheh, Y. Wang, K. Fu, A. Jiang, and E. G. Learned-Miller. Exploiting half-wits: Smarter storage for low-power devices. In *FAST*, 2011.

[7] A. Sampson, J. Nelson, K. Strauss, and L. Ceze. Approximate storage in solid-state memories. *ACM Transactions on Computer Systems (TOCS)*, 2014.

[8] SanDisk. The flash-transformed data center: Flash adoption is growing across the enterprise. June 2014.

[9] Silicon Labs. EFM32WG-STK3800 Wonder Gecko Starter Kit User's Guide. `http://www.silabs.com/Support%20Documents/TechnicalDocs/efm32wg-stk3800-ug.pdf`.

[10] Texas Instruments. TLV4100 Family of High Output Drive Operational Amplifiers with Shutdown. `http://www.ti.com/lit/ds/symlink/tlv4110.pdf`.

[11] H.-W. Tseng, L. M. Grupp, and S. Swanson. Underpowering nand flash: Profits and perils. In *Design Automation Conference (DAC), 2013 50th ACM/EDAC/IEEE*, pages 1–6. IEEE, 2013.