International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 4, No. 1, 2017, pp.163-175.

ISSN 2454-390X

International Academic Journal of Science and Engineering

www.iaiest.com

Recent Advances in Nonvolatile Memory

Zafar Ali
Department of Computer Sciene, Virtual University of Pakistan

Abstract

Nonvolatile memory is making significant advances off-late. Earlier nonvolatile memory was 'Read Only' type. Present breakthroughs added 'Write' access to this type of memory. Market demands many attractive features on a single product. There is a trade-off for optimized features. An over-all optimization between hardware and software is essential.

Keywords: RAM, ROM, flash, write time , data retention time, reliability.

Introduction

Memory elements play a critical role in computer architecture. The famous law of Moore 1965, that predicted more and more component integration with time as can be observed from fig. 1 is still obeyed. For memory elements the present integration trend is faster than that predicted by Moore's law.

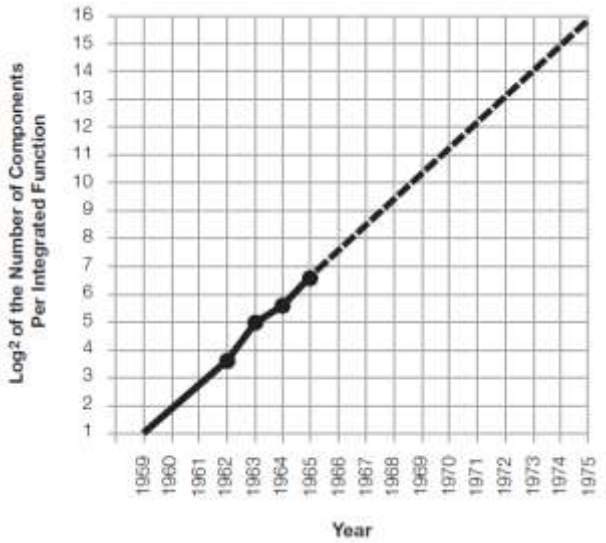


Figure 1: Prediction of Moore's Law [1].

More integration demands high density memories which in its turn demands low power, light-weight design. This type of design is essential for the huge memory requirements in the emerging market of portable devices as well.

Different classes of memories and their basic architectures are well described by Weste and Eshraghian [2]. Depending on the access type, memory can be divided into three broad categories:

I. Random Access Memory
II. Serial Access Memory
III. Content Access Memory

Random access memory at the chip level is classed as memory that has an access time independent of the physical location of the data. This is in contrasted with serial access memories which have different location dependent read-write latencies and content access memory with datum dependent read-write latencies.

Again, memories may have only read access and is called as Read Only Memory (ROM) or it can have both Read-Write Access and is commonly called as Read-Write Access Memory (RAM). Recent advances in technology provide write access to ROM and some of the new products are: Programmable Read Only Memory (PROM), Electrically Programmable Read Only Memory (EPROM), and Electrically Erasable Programmable Read Only Memory (EEPROM).

A basic building block of a memory latches the data. Fig. 2 shows how a read-write is possible in this simple circuit of a RAM. The simple 6 transistor circuit with two cross coupled N-MOSFETs provides the read-write access. When the word line is high, write process takes place. The bit say, 1 or high voltage moves through the pass transistor and switches on the left N type MOSFET on. The complement bit shown as –bit switches off the right transistor. Thus the output data remains latched and can be read when the word line goes to low state.

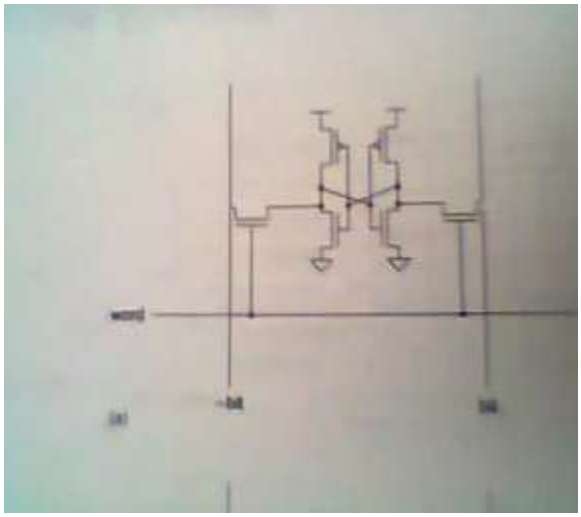


Figure 2: RAM Cell Circuit [2].

When the supply is switched off the output goes to high impedance. For this reason, the memory is called volatile memory. In early days, RAM was always volatile and ROM could be non-volatile. Recent advances club all the benefits in one product called as non-volatile memory. This has read-write access though write time is much larger than the read time. Electrical writing takes place when the control voltage or word line goes very high, 6.6 volt a typical value. Fowler-Nordheim tunneling takes place through the barrier, usually a thin film of oxide between the bit to be written and the write location.

Problem Statement

Since the commercialization of memory products lots of innovative research is going on to provide more and more attractive features to a single product as customers demand all. The complication arises as there is a trade-off for attractive features. When one feature is improved the other one degrades. The problem is minimized by using innovative strategies. There are discussed in the next section.

Proposed Solution

One of the demanding features is fast writing on the memory. To achieve low write latency the charging current through the barrier should be maximized. As a consequence the leakage current through the barrier would be increased yielding a low retention time. So data retention time lowers when a memory product is improvised for faster writing.

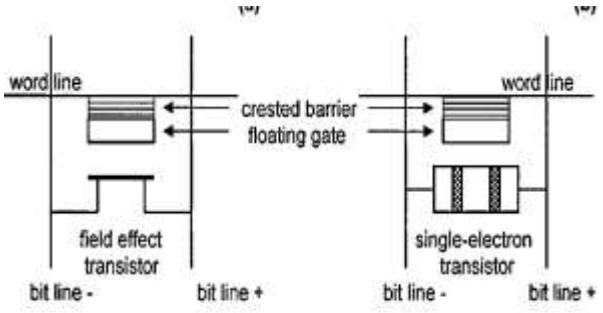


Figure 3: Two types of cells of the Non VOlatile Read-write Access Memory (NOVORAM) [3].

Likharev [3] appraised different non-uniform energy barriers to meet these two contradictory requirements. The author did mathematical modeling for 5 types of barriers and concluded that tri-layer barrier provides best write latency or maximum charging current. Figure 3 gives the cross-sections of two types of proposed devices.

Govoreanu *et al.* [4] concluded this analysis with their experimental support. Gate insulators with variable thicknesses were fabricated to create such non-uniform energy barrier. Such type of device was named as VARIOT to interpret that as VARIable Oxide Thickness.

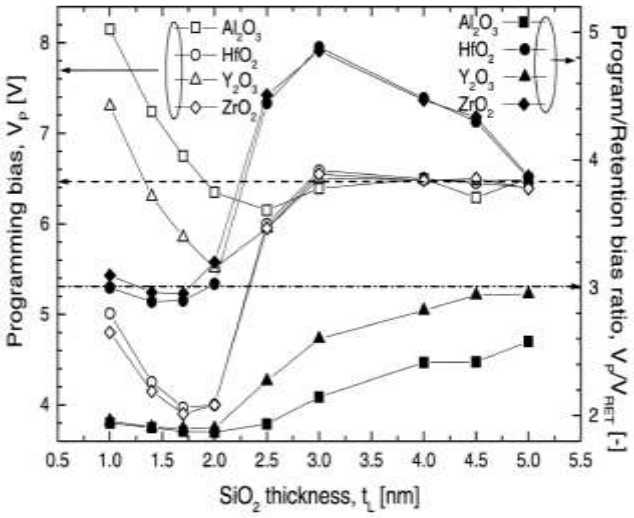


Figure 4: Change of bias with insulator thickness [4].

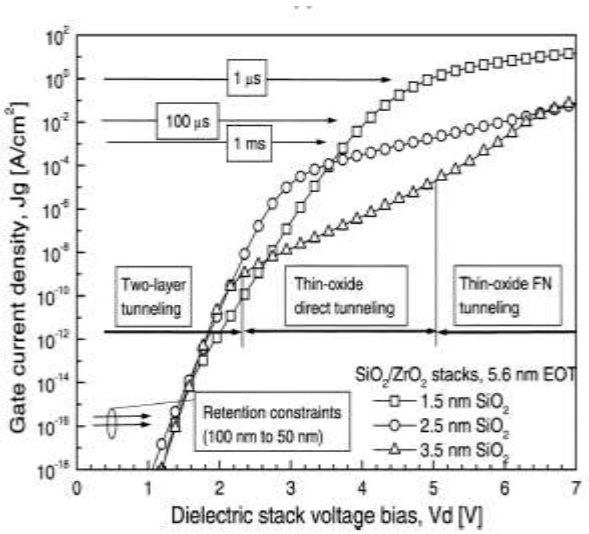


Figure 5: Different types of tunneling [4].

Another innovative strategy is proposed by Li et al. [5] to obtain high data retention time by using low leakage insulator. This reports for the first time a programmable memory using MOSFET device on 6H-SiC. 6H-SiC being a wide band gap semiconductor allows a negligible leakage current. So the memory retention time is much higher as compared to Si. The work describes a detailed device characterization that includes (i) drain current drain voltage characteristics at varying gate voltages. (ii) threshold voltage shift after programming and erasing at room temperature and at 200 degree centigrade. (iii) charge retention shown by threshold voltage shift with time after programming.

The reported channel length was 1 – 10 µm. Both the channel length and its variation are quite large as compared to matured Si technology. This is expected for any new material. The memory features were not compared with the already reported Si technology results. It was mentioned that, a more comprehensive result would be reported in future. But no such paper could be found from the same group of authors.

Variability issue is very important for even well-matured Si technology. It is recently addressed by Gupta et al. [6]. In this paper the hardware, especially memory, variability issues are addressed and the measures to handle these issues are discussed. As the devices get miniaturized a few atomic layers decide the device parameters whereas device processing tolerance is more than the atomic level. The hardware variation occurs mainly from four sources:

I. Manufacturing
II. Environment
III. Vendor
IV. Aging

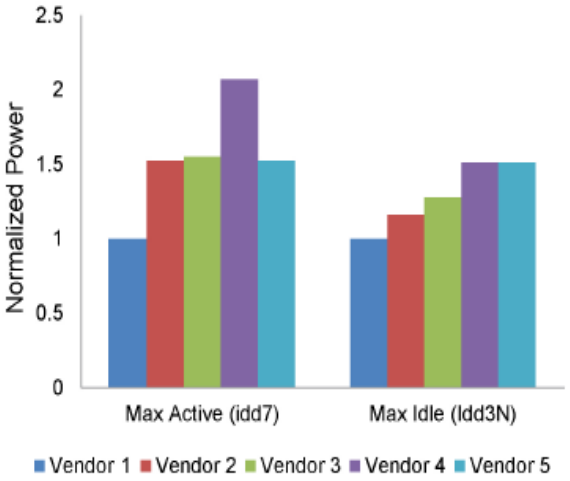


Figure 6: Power Variation for Memory Products [6].

For memories push toward "big data" applications has made the variability issue more pronounced.

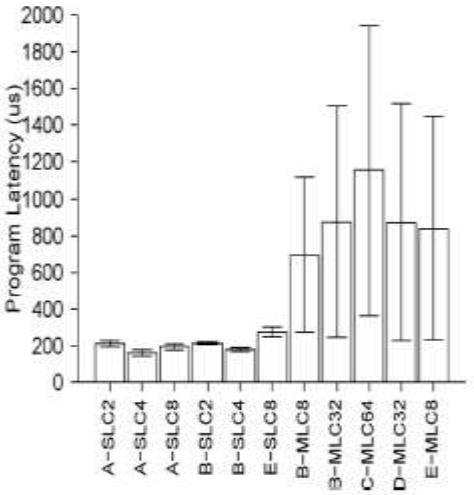


Figure 7: Program Latency variation for different memory products [6].

The solution may stem from the realization that the problem is not variability, rather how computer system designers have been treating the variability. While chip components no longer function the precisely chiseled parts of the past, the basic approach to the design and operation of computers has remained unchanged. Software assumes hardware of fixed specifications that hardware designers try hard to meet. Yet hardware designers must accept conservative estimates of hardware specifications and cannot fully leverage software's inherent flexibility and resilience.

Here such type of solution is proposed that defends the variability issue. Here an innovative idea of Under- designed and Opportunistic (UnO) computing machine is explained. Further special software stacks take care of the variability issue.

An ultra high density memory named as flash is proposed in 90-s by Masuoka *et al.* [7]. Such type of memory is very good substitute for hard disk drive. The work has a great practical importance. According to Fulford [8], this innovation along-with its counterpart NOR flash captured a $76 billion market in 2001.The high density is achieved without any scale-down. The number of contacts is reduced in this innovative design. All memory manufacturers are adding to this innovation to make the flash features more attractive. Most of these innovation details are not available in the existing open literature.

The bit addressability was demonstrated. Detailed experimental results were presented. It was claimed that reliability study was done. Under the study read retention time plot (fig. 8 in text it is mentioned as fig. 10) is referred. This measurement does not fall under reliability study.

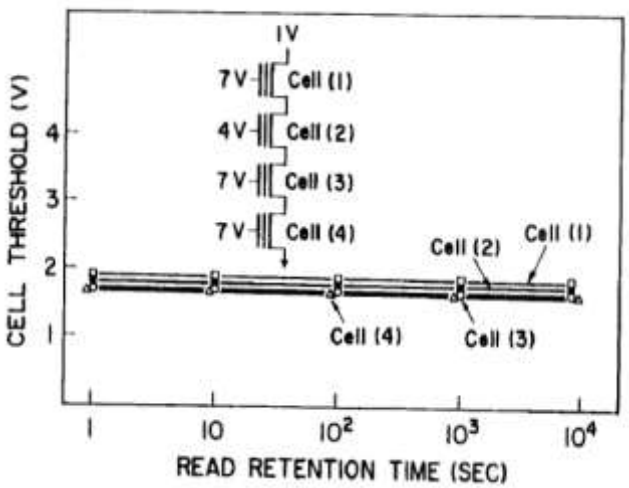


Figure 8: READ Retention Characteristics [7].

Kim et al. [9] compared the access types of different types of memories. This is given in the following table.

TABLE I
CHARACTERISTICS OF DIFFERENT STORAGE MEDIA.

Media	Access time		
	Read	Write	Erase
DRAM	60ns (2B) 2.56μs (512B)	60ns (2B) 2.56μs (512B)	-
NOR FLASH	150ns (1B) 14.4μs (512B)	211μs (1B) 3.53ms (512B)	1.2s (128KB)
NAND FLASH	10.2μs (1B) 35.9μs (512B)	201μs (1B) 226μs (512B)	2ms (16KB)
Disk	12.4ms (512B) (average)	12.4ms (512B) (average)	-

[Reference] DRAM: 2-2-2 PC100 SDRAM. NOR FLASH: Intel 28F128J3A-150. NAND FLASH: Samsung K9F5608U0M. Disk: Segate Barracuda ATA II.

From table I, it can be observed that flash has access times much less than the disk. One drawback of flash is; its effective write time is very high. Prior to writing an erase is necessary and memory to be erased is usually larger than the memory required for writing.

Authors proposed a new type of Flash Translation Layer (FTL) that effectively reduces the write time. Lots of research is going on FTL. Wu et al. [10] proposed an adaptive translational layer and Liu et al. [11] proposed a Resource Aware Sector Translation Layer (RAST). It would have been interesting to study these developments if the later results were compared with the previous ones.

Instead of innovation in FTL the total storage area network should be optimized taking into consideration both hardware and software [12, 13]. From the figure it can be seen that software has a major role on power and latency.

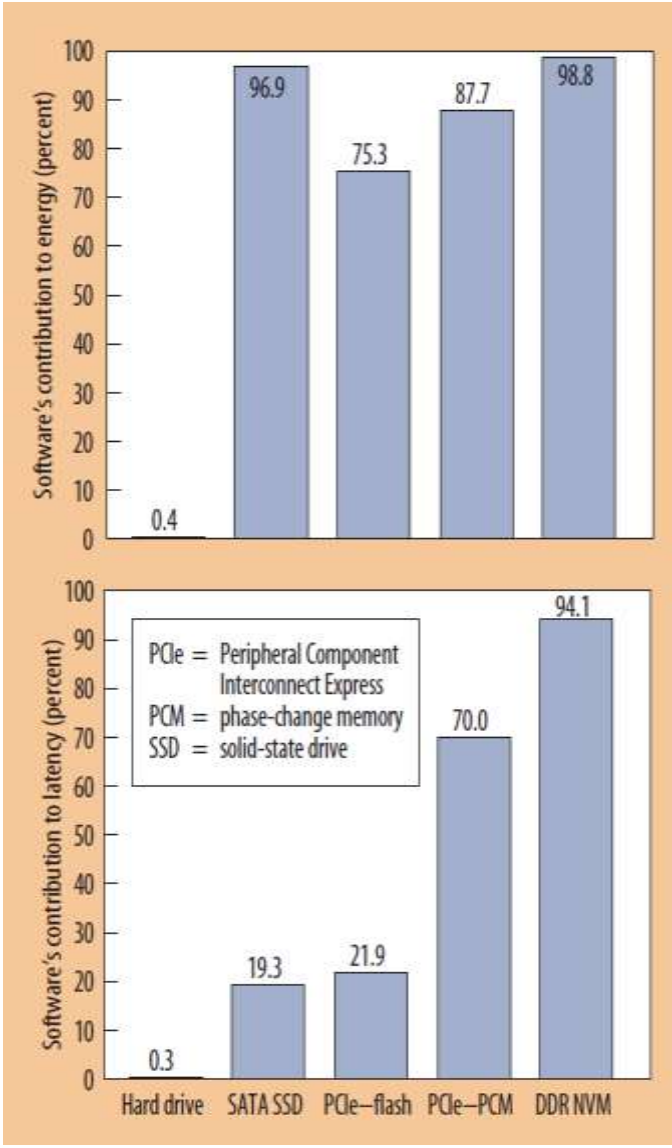


Figure 9: Contribution of software for different memory architectures [12].

One such design optimization for memory is considered in Near Data Processing (NDP). In NDP computation can be performed right at the data's home, either in caches, main memory or persistent storage. This is in contrast to the movement of data toward a CPU independent of where it resides, as is done traditionally. Examples of NDP already exist in systems that execute computations close to the disk, filtering or processing the data streaming from the disks so that a minimal number of items are transferred for processing at other parts of the system. Conceptually, the same principle can be applied at other levels of a system's memory and storage hierarchy by placing computing resources close to where data is located, and restructuring applications to exploit the resulting distributed computing architecture [14].

Another type of non-volatile memory is Phase Change Memory (PCM). Pirovano et al. [15] evaluated the performance and reliability of PCM. Here the bit is programmed depending on the phase of the material namely amorphous or crystalline. The PCM essentially consists of a chalcogenide layer usually $Ge_2Sb_2Te_5$ (GST) contacted by a bottom heater and a top metal layer. The cell resistance varies according to the GST phase, which can be suitably changed by fast electrical pulses between a low conductive state (amorphous) and a high conductive one (crystal). The devices were switched in the amorphous state (reset) by applying an electrical pulse high enough to melt the active region and rapidly quenching the device. The reset pulse width was 40 ns with a falling time shorter than 10 ns while a relatively longer pulse of 100 ns was employed to crystallize the cell or to set the bit.

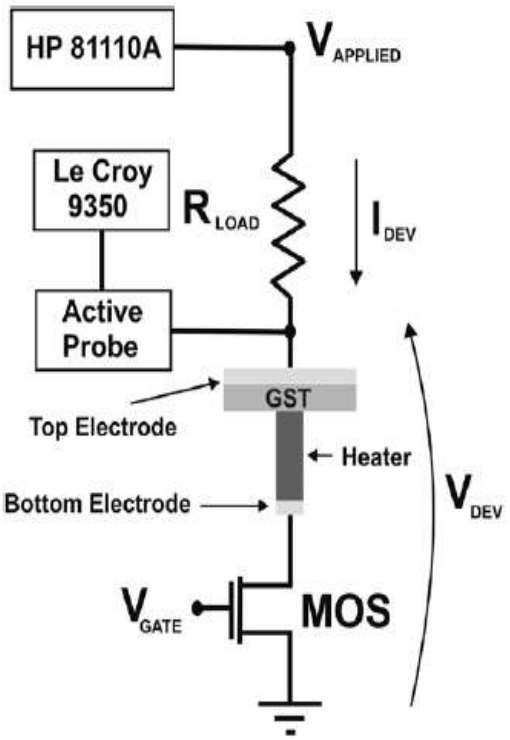


Figure 10: Experimental set-up [15].

Experimental set-up is shown in the above figure. Here the MOS device is used for current limiter. Experimental data were collected on PCM cells integrated in a standard 180 nm CMOS technology.

The programmed bit value depends on the phase of the material which in its turn depends on the temperature. It appears that, cross-talk and self-heating could be a problem. From the simulation results authors defended that cross-talk can be neglected. Since the pulses for phase changes are very narrow such interpretation is justified. But the problem of self-heating cannot be ignored.

In the reliability studies, Arrhenius plots were done at higher temperatures. From these plots data retention time was interpreted as 10 years. High temperature is a stress when the parameter degrades with

the temperature. In this case, the retention time might improve with increase in temperature as it goes deep into one phase. Such interpretation needs to be justified.

A new kind of reliability study under power-cuts was done for flash memories by Tseng et al. [16]. As the memory degrades or often fails for power cuts this type of study is extremely important. Both types of flash cells: Single Level Cell (SLC) and Multi Level Cell (MLC) were studied. Memory capacity ranges from 2 Gbits to 32 Gbits. Products from different vendors were studied.

For this work, the test platform was built with three components: the Xilinx XUP board, a custom flash testing board, and the power control circuit. The FPGA on the Xilinx XUP board contains a PowerPC 305 core running Linux. A custom flash controller on the FPGA provides us direct access to the flash device via the flash testing board. The FPGA also controls the power to the flash chips by means of a pair of high-speed power transistors, Measurements with an oscilloscope show that the system can switch the chip's power supply to 0 V within 3.7 μs.

Bit Error Rate (BER) for all the products were studied at different power cut intervals. Many interesting observations were made and some of them are really surprising. Some of the observations are:

I. The BER is much higher in flash than the Hard Disk Drive (HDD).

II. BER varies significantly from product to product. MLC flash shows larger BER than SLC flash as expected.

III. BER decreases as the power cut interval increases as expected. But what is surprising is: the decrease in BER is not monotonic. Fall in BER is a staircase function.

IV. Bit error occurs on data that was being written at the instant of power cut. Surprisingly error may also occur for already written bits. Authors named it as "Retroactive Data Corruption".

V. Data corruption may take place at the time of reading with a power cut.

Suitable back-up schemes for power cut failures were suggested by the authors.

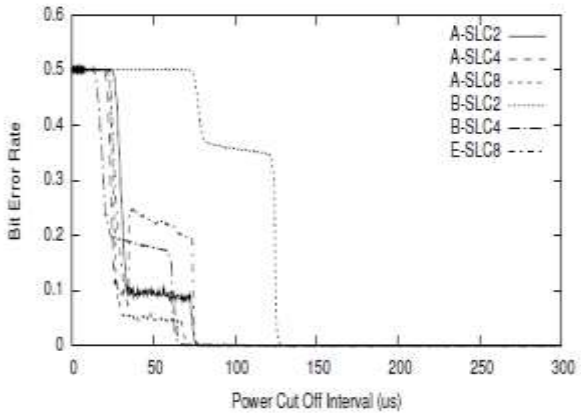


Figure 11: BER vs. power cut interval [16].

Discussion and Analysis

- Nonvolatile memory enjoys a huge market potential.
- Market demands all attractive features in the same product.

- There is a trade-off among the features.
- Innovative strategies are proposed to beat the challenge of trade-off.
- For feature optimization an overall view that includes hardware as well as software is very important.

Conclusion

Since the commercialization of memory products lots of innovative research is going on to provide more and more attractive features to a single product as customers demand all. Earlier nonvolatile memory did not have write access. Recent advances added this feature to this product. Efforts are on to add more and more attractive features. This needs optimization of hardware as well as software.

References

[1] G.E. Moore, GE 1965, "Cramming more components onto integrated circuits", *Electronics*, vol. 38, no. 8, pp. 1-4, April 1965.

[2] N.H.E. Weste, and K. Eshraghian, Principles of CMOS VLSI design: A Systems Perspective, Addison Wesley, AT&T, USA, 1993.

[3] K.K. Likharev, "Layered tunnel barriers for nonvolatile memory devices", Applied Physics Letters, vol. 73, no. 15, October, 1998, pp. 2137 – 2139.

[4] B. Govoreanu, P. Blomme, M. Rosemeulen, J.V. Houdt, and K.D. Meyer, "VARIOT: a novel multilayer tunnel barrier concept for low-voltage nonvolatile memory devices", *IEEE Electron Device Letters*, Vol. 24, no. 2, February 2003, pp. 99 – 101.

[5] C. Li, J.S. Duster, and K.T. Kornegay, "A nonvolatile semiconductor memory device in 6H-SiC for harsh environment applications", *IEEE Electron Device Letters*, Vol. 24, no. 2, February 2003, pp. 72 – 74.

[6] P. Gupta, Y. Agarwal, L. Dolecek, N. Dutt, R.K. Gupta, R. Kumar, S, Mitra, A. Nicolau, T.S. Rosing, M.B. Srivastava, S. Swanson, and D. Sylvester, "Underdesigned and opportunistic computing in presence of hardware variability", *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 1, 2013, pp. 8 – 23.

[7] F. Masuoka, M. Momodomi, Y. Iwata, and R. Shirota, "New ultra high density EPROM and flash EEPROM with NAND structure cell", 1987 International Electron Devices Meeting, vol. 33, 1987, pp. 552 – 555.

[8] B. Fulford, Unsung hero, *Forbes*. 2002. Viewed 17 December, 2014, http://www.forbes.com/glbal/2002/0624/030.html

[9] J. Kim, J.M. Kim, S.H. Noh, S.L. Min, and Y. Cho, 2002, "A space-efficient flash translation layer for Compact flash systems', IEEE Consumer Electronics, Vol.48 no. 2, May 2002, pp. 366 – 375.

[10] C.H. Wu, H.H. Lin, and T.W. Kuo, "An adaptive flash translation layer for high-performance storage systems", *IEEE* Computer-Aided Design of Integrated Circuits and Systems, Vol. 29, no. 6, June 2010, pp. 953 – 965.

[11] D. Liu, Y. Wang, Z. Qin, Z. Shao, and Y. Guan, 2012, "A Space Reuse Strategy for Flash Translation Layers in SLC NAND Flash Memory Storage Systems", IEEE Very Large Scale Integration (VLSI) Systems, vol. 20, no. 6, 2012, pp. 1094 – 1107.

[12] S. Swanson, and A.M. Caulfield, 'Refactor, Reduce, Recycle: Restructuring the I/O stack for the future of storage', IEEE Trans. Computer, vol. 46, no. 8, 2013, pp. 52 – 59.

[13] A.M. Caulfield, 2013, 'Moneta: A storage system for fast non-volatile memories', PhD Thesis, University of California, San Diego, viewed 17 December, 2014, http://cseweb.ucsd.edu/~swanson/papers/AdrianThesis.pdf

[14] Balasubramonian, R, Chang, J, Manning, T, Moreno, JH, Murphy, R, Nair, R, & Swanson, S 2014, 'Near-data processing: Insights from a micro-46 workshop', *IEEE Trans. Micro*, vol. 34, no. 4, pp. 36-42.

[15] Pirovano, A, Redaelli, A, Pellizzer, F, Ottogalli, F, Tosi, M, Ielmini, D, Lacaita, AL, Bez, R 2004, 'Reliability study of phase-change nonvolatile memories', *IEEE Trans. Device and Materials Reliability*, vol. 4, no. 3, pp. 422 – 427.

[16] Tseng, WH, Grupp, LM, & Swanson, S 2011, 'Understanding the impact of power loss on flash memory', *48 Design Automation Conference,* DAC 2011.

.