International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. 2, Special Issue 8, March 2016

OPTIMAL METHOD FOR REDUCING POWER CONSUMPTION IN EDRAM MODULES

Venkatesh.P, PG Student
Guided by
Madhesh.P,Asst.Prof
Department of Electronics and Communication Engineering
Shreenivasa Engineering College
Mail id: venkateshkpece@gmail.com

Abstract— **EDRAM cells require periodic refresh, which ends up consuming substantial energy for large last-level caches. In practice, it is well known that different eDRAM cells can exhibit very different charge-retention properties. Unfortunately, current systems pessimistically assume worst-case retention times, and end up refreshing all the cells at a conservatively-high rate. In this paper, we propose an alternative approach. We use known facts about the factors that determine the retention properties of cells to build a new model of eDRAM retention times. The model is called Mosaic. The model shows that the retention times of cells in large eDRAM modules exhibit spatial correlation. Therefore, we logically divide the eDRAM module into regions or tiles, profile the retention properties of each tile, and program their refresh requirements in small counters in the cache controller. With this architecture, also called Mosaic, we refresh each tile at a different rate. The result is a 20x reduction in the number of refreshes in large eDRAM modules practically eliminating refresh as a source of energy consumption.**

Index Terms— **Composed dynamic random access memory (DRAM) cell, data retention time, low-power DRAM, mobile DRAM, partial access mode (PAM), partial array self refresh.**

I. Introduction

An attractive approach to reduce the energy wasted to leakage in the cache hierarchy of multicores is to use embedded DRAM (eDRAM) for the lower levels of caches. EDRAM is a capacitor based RAM that is compatible with a logic process, has high density and leaks very little. While it has higher access times than SRAM, this is not a big concern for large lower-level caches. As a result, eDRAM is being adopted into mainstream products. For example, the IBM POWER7 processor includes a 32 MB on-chip eDRAM L3 cache, while the POWER8 processor will include a 96 MB on-chip eDRAM L3 cache and, potentially, an up to 128 MB off-chip eDRAM L4 cache. Similarly, Intel has announced a 128 MB off-chip eDRAM L4 cache for its Haswell processor. EDRAM cells require periodic refresh, which can also consume substantial energy for large caches. In reality, it is well known that different eDRAM cells can exhibit very different charge-retention properties and, therefore, have different refresh needs. However, current designs pessimistically assume worst-case retention times, and end up refreshing all For example, they use a refresh period of around 40 µs. This naive approach is wasteful. Since eDRAM refresh is an important problem, there is significant work trying to understand the characteristics of eDRAM charge retention. Recent experimental work from IBM has shown that the retention time of an eDRAM cell strongly depends on the threshold voltage (Vt) of its access transistor. In this paper, we note that, since the values of Vt within a die have spatial correlation, then eDRAM retention times will also necessarily exhibit spatial correlation.

Consequently, in this paper, we first develop a new model of the retention times in large on-chip eDRAM modules. The model, called Mosaic, builds on process-variation concepts. It shows that the retention properties of cells in large eDRAM modules do exhibit spatial correlation.

Generally, almost all the IC manufacturers of mobile, hand-held devices, computer is facing the power consumption issue and they have come out a lot of method to solve this problem. For example, the register file that located inside Intel Haswell processor. Register file is consuming about 27% from the overall power consumption of Haswell processor while the domino multiplexer consume the power when read process which is Read Local Bitline, Read Global Bitline, Read Local Bitline(LBL) Precharge and Read Global Bitline(GBL) Precharge. In addition, the domino multiplexer consumes highest leakage power when Read Local Bitline process is carried. Therefore, this project will study the reduction of the power of domino multiplexer. A low power methodology is proposed for the domino multiplexer used in register file which is selective power-gated domino multiplexer.However, Intel invented a method that able solve this problem which called on-demand precharge. The advantage of this method is it able to turn off the precharge circuit when the LBL is not being access in order to save the leakage power. Nevertheless, this will come out another problem when the charge-down or/and charge-up process happens too frequently as the dynamic power in that situation may surpass the leakage power.

This paper is organized as follows: in Section II, the over view of the project is given. In Section III, the existing system is described. In Section IV, the proposed system is briefly explained, where as the section v gives the module description.

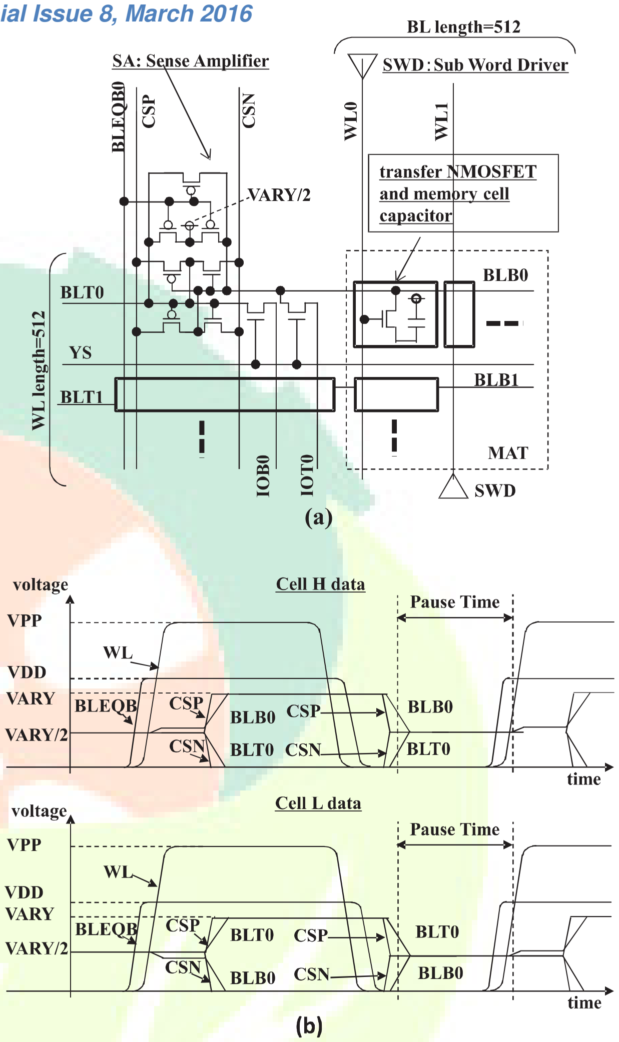
II. OVERVIEW OF THE PROPOSAL

The conversion can be realized very simply from the structure of the DRAM array circuit. This method can reduce the frequency of the disturbance and its power consumption by two orders of magnitude. The DRAM is fully compatible with a conventional DRAM. In its usual operating mode, the full memory capacity is used. In the PAM, the capacity is limited to 2-N of the total capacity; however, memory cells are fully used to share the storage charge to extend the retention time. We propose a new method to reduce the refresh current in a DRAM by extending the retention time effectively when the amount of the data to be stored is small. We call this low-power mode as the partial access mode (PAM). The retention time has been shown to exhibit both tail and main distributions. Most of the cells belong to the main distribution and have retention times significantly higher than the product specification. Only a minor portion suffers from increased leakage.

III. EXISTING SYSTEM

The DRAM memory capacity has been increasing, even though its die size has almost remained constant. In 2011, a 2-Gb DRAM was fabricated by a 30-nm [minimum feature size (F) value] process and was placed on the market. A DRAM stores a single bit in a memory cell as an amount of electrical charge on a storage capacitor. Charge is lost by the leakage current of the p-n junction, sub threshold current, and gate-induced drain leakage (GIDL). This means that a DRAM requires a rewrite operation before the memory cell loses its storage charge. This rewrite operation is called refresh. Refresh is performed by issuing an auto-refresh command (AREF). Because refresh is a type of disturbance in the system where sense amplifier (SA) activation, precharging, and read or write operations are forbidden, the frequency of the AREF command should be minimized. It is well known that there is variation in the retention time of DRAM cells. The overall distribution and the sources of variation have also been identified which shows a typical DRAM retention time distribution.

The short retention time with an extremely low probability determines the refresh interval of the memory cell. Conversion from 1 cell/bit to 2N cells/bit reduces the variation in the retention time among memory cells. Although the active power increases by a factor of 2N, the refresh time increases by more than 2N as a consequence of the fact that the majority decision does better than averaging for the tail distribution of retention time. The conversion can be realized very simply from the structure of the DRAM array circuit.



(a)

(b)

Fig. 1. (a) Array circuit and (b) its waveforms.

Since eDRAM refresh is an important problem, there is significant work trying to understand the characteristics of eDRAM charge retention. Recent experimental work from IBM has shown that the retention time of an eDRAM cell strongly depends on the threshold voltage (Vt) of its access transistor. In this paper, we note that, since the values of Vt within a die have spatial correlation, then eDRAM retention times will also necessarily exhibit spatial correlation.

The disadvantages of the existing system is as follows

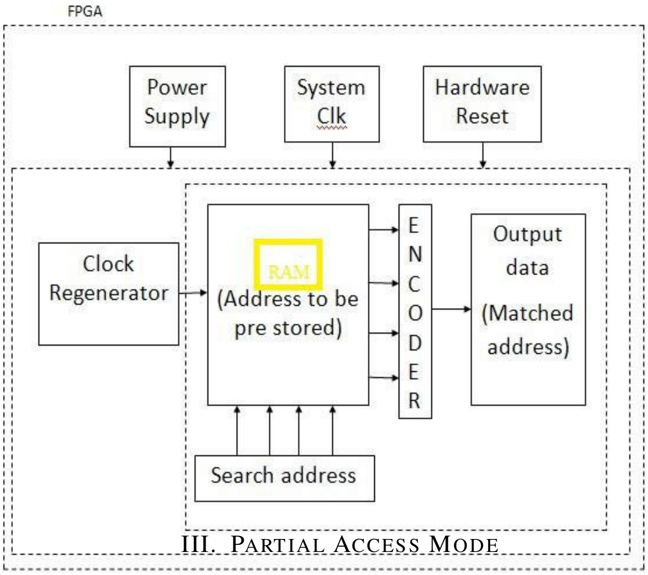
1.The rewrite voltage is saturated by the difficulty associated with the real operation.

2.The standby power increases with the memory capacity.

3.Reduction in power consumption is less in standby mode.

IV. PROPOSED SYSTEM

The proposed system consists of an access transistor and a storage capacitor. The logic state is stored as electrical charge in the capacitor. The capacitor loses charge over time through the access transistor shown as Io f in the figure. Therefore, an eDRAM cell requires periodic refresh to maintain the correct logic state. The memory has a limited number of ports, and only a number of lines equal to the number of ports can be refreshed simultaneously. Hence, some of the refreshes may need to be delayed. As a result, to ensure correctness when multiple lines need to be refreshed at the same time, and some refreshes need to be delayed, we need to provide a timing guardband. We can logically group cells into regions, profile their retention time, and set-up time counters to refresh the regions only at the frequency that each one requires. With reasonable spatial correlation, the hardware cost of the counters will be minimal.



III. Partial Access Mode

Fig. 2. Block diagram for proposed system

Mosaic induces little area and energy overhead. To see why, assume that corresponds to a 1 MB cache bank with 64-byte lines and 16-line tiles. Such an organization has 1024 tiles in the bank. It consists of an access transistor and a storage capacitor. The logic state is stored as electrical charge in the capacitor. The capacitor loses charge over time through the access transistor — shown as Io f f in the figure. Therefore, an eDRAM cell requires periodic refresh to maintain the correct logic state. This lower bound would be hard to attain. First, the memory has a limited number of ports, and only a number of lines equal to the number of ports can be refreshed simultaneously. Hence, some of the refreshes may need to be delayed.

As a result, to ensure correctness when multiple lines need to be refreshed at the same time, and some refreshes need to be delayed, we need to provide a timing guardband. The striations seen in indicate that there is spatial correlation in the Tret of adjacent lines—especially within the same way. Hence, we group sets of contiguous lines from the same way into Tiles, setting Tret for the tile to the minimum of the Tret of the lines constituting the tile. 4h and 4i show the tile-level distribution of log10 Tret for tile sizes of 16 and 64 lines, respectively. We can see that there are regions of spatial locality. The resulting distribution looks like a mosaic of tiles. With this design, the hardware cost of a counter is amortized over a whole tile. However, we have to refresh the whole tile whenever the counter rolls down to zero. We can potentially attain higher energy savings by using variable-sized tiles. For example, if there is a line or a small region of lines that have very different Tret than their neighbors, we can save refreshes by placing them in a tile of their own.

V. MODULE DISCRIPTION

Module1:

Design and analysis of cell:

The cell is sensing element in this design. The input search data address is given as an input to the cell through the search line and directly compare with every bit of the storage word using the comparison circuit in the cell. The comparison circuit has two inverter and a sensing amplifier

Module2:

Design and analysis of a matrix using cell:

The cell designed in the previous module is ordered in a form of matrices as an eight bit word. The 8×8 matrix has the binary search data as the bit stored in the cell. The compare operation of the CAM provides this module output as a matrix.

Module3:

Design and analysis of a encoder matrix:

The modules 3 have the encoder which senses the binary search data of the address from the matrix of the ML line in the evolution stage and provide the encoded output from this module. The matrix is compared in the pre-charge state and given as a output matrix.

Module4:

Design and analysis of a integration module:

The output from the previous module is 8×8 matrix which is not in a single binary search data form. Hence the integrator is used in this module to combine the data to form an 8 bit binary data output. The output is generated as an integrator output.

VI. SOFTWARE REQUIREMENTS

Programming language	: VHDL
Operating system	: Microsoft Windows 2007
Tools used for simulation	: XILINX 9.2

There are two main types of simulation: functional and timing simulation. The functional simulation tests the logical operation of a circuit without accounting for delays in the circuit. Signals are propagated through the circuit using logic and wiring delays of zero. This simulation is fast and useful for checking the fundamental correctness of the designed circuit.

The second step of the simulation process is the timing simulation. It is a more complex type of simulation, where logic components and wires take some time to respond to input stimuli. In addition to testing the logical operation of the circuit, it shows the timing of signals in the circuit. This type of simulation is more realistic than the functional simulation; however, it takes longer to perform.

VHDL is an acronym which stands for VHSIC Hardware Description Language. VHSIC is yet another acronym which stands for Very High Speed Integrated Circuits. If you can remember that, then you're off to a good start. The language has been known to be somewhat complicated. The acronym does have a purpose, though; it is supposed to capture the entire theme of the language that is to describe hardware much the same way we use schematics.

VHDL can wear many hats. It is being used for documentation, verification, and synthesis of large digital designs. This is actually one of the key features of VHDL, since the same VHDL code can theoretically achieve all three of these goals, thus saving a lot of effort. In addition to being used for each of these purposes, VHDL can be used to take three different approaches to describing hardware. These three different approaches are the structural, data flow, and behavioral methods of hardware description. Most of the time a mixture of the three methods is employed. The following sections introduce you to the language by examining its use for each of these three methodologies. There are also certain guidelines that form an approach to using VHDL for synthesis.

VHDL is a standard (VHDL-1076) developed by IEEE (Institute of Electrical and Electronics Engineers). The language has been through a few revisions, and you will come across this in the VHDL community. Currently, the most widely used version is the 1987 (STD 1076-1987) version, sometimes referred to as VHDL'87, but also just VHDL. However, there is a newer revision of the language referred to as VHDL'93. VHDL'93 (adopted in 1994 of course) is fairly new and is still in the process of replacing VHDL'87.

VHDL is an IEEE and U.S. Department of Defense standard for electronic system descriptions. It is also becoming increasingly popular in private industry as experience with the language grows and supporting tools become more widely available. Therefore, to facilitate the transfer of system description information, an understanding of VHDL will become increasingly important.

VII. SIMULATION RESULTS

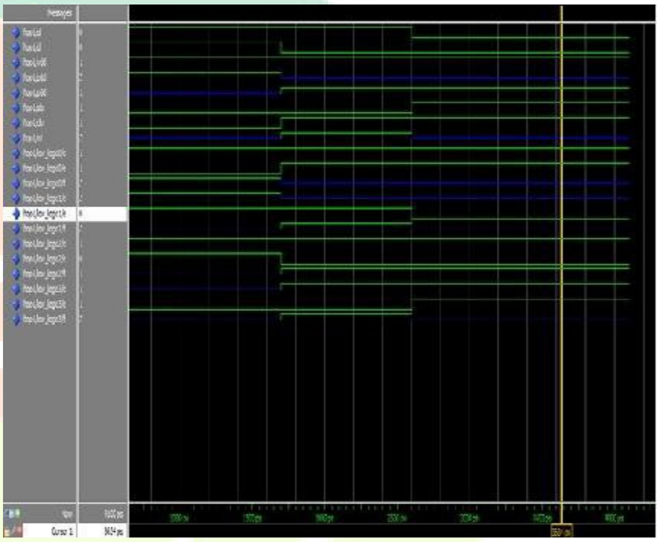


Fig 3: Waveform of Existing System

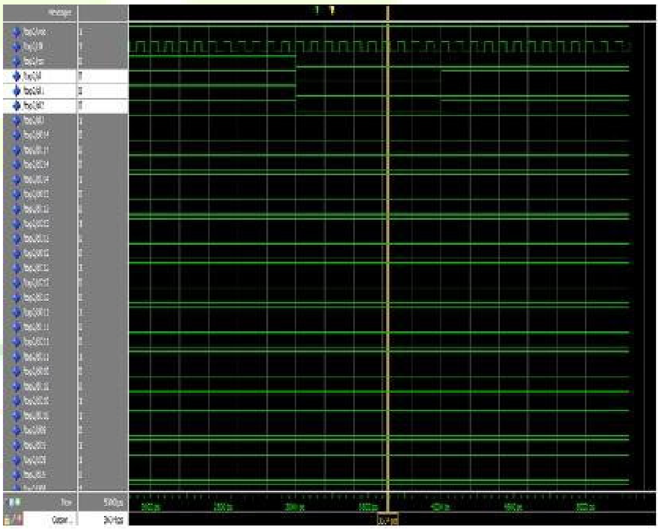


Fig 4: Waveform of Proposed System

VIII. CONCLUSION

This paper has presented a new model of the retention times in large on-chip eDRAM modules. This model, called Mosaic, showed that the retention times of cells in large eDRAM modules exhibit spatial correlation. Based on the model, we used the simple Mosaic tiled organization of eDRAM modules, which exploits this correlation to save much of the refresh energy at a low cost. We evaluated Mosaic on a 16-core multi core running 16-threaded applications. We found that Mosaic is both inexpensive and very effective. An eDRAM L3 cache augmented with Mosaic tiles increased its area by 2% and reduced the number of refreshes by 20 times. This reduction is 5 times the one obtained by taking the RAIDR scheme for main memory DRAM and applying it to cache eDRAM. With Mosaic, we saved 43% of the total energy in the L3 cache, and got very close to the lower bound in refresh energy.

REFERENCES

[1] G. A. M. Hurkx, D. B. M. Klaassen, and M. P. G. Knuvers, "A new recombination model for device simulation including tunneling," *IEEE Trans. Electron Devices*, vol. 39, no. 2, pp. 331–338, Feb. 1992.

[2] S. Amakawa and K. Nakazato, "A new approach to failure analysis and yield enhancement of very large-scale integrated systems," in *Proc. ESSDERC*, Sep. 2002, pp. 147–150.

[3] O. K. B. Lui and P. Migliorato, "A new generation-recombination model for device simulation including the Poole-Frenkel effect and phonon-assisted tunneling," *Solid-State Electron.*, vol. 41, no. 4, pp. 575–583, 1997.

[4] W. Shockley and W. T. Read, "Statistics of the recombinations of holes and electrons," *Phys. Rev.*, vol. 87, no. 5, pp. 835–842, 1952.

[5] A. Hiraiwa, M. Ogasawara, N. Natsuaki, Y. Itoh, and H. Iwai, "Field-effect trap-level-distribution model of dynamic random access memory data retention characteristics," *J. Appl. Phys.*, vol. 81, no. 10, pp. 7053–7060, 1997.

[6] S. Ueno, Y. Inoue, M. Inuishi, and N. Tsubouchi, "Leakage mechanism of local junctions forming the main or tail mode of retention characteristics for dynamic random access memories," *Jpn. J. Appl. Phys.*, vol. 39, no. 4B, pp. 1963–1968, 2000.

[7] M. Chang, J. Lin, S. N. Shih, T. C. Wu, B. Huang, J. Yang, and P. I. Lee, "Impact of gate-induced drain leakage on retention time distribution of 256 Mbit DRAM with negative wordline bias," *IEEE Trans. Electron Devices*, vol. 50, no. 4, pp. 1036–1040, Apr. 2003.

[8] K. Saino, S. Horiba, S. Uchiyama, Y. Takaishi, M. Takenaka, T. Uchida, Y. Takada, K. Koyama, H. Miyake, and C. Hu, "Impact of gate-induced drain leakage current on the tail distribution of DRAM data retention time," in *IEDM Tech. Dig.*, Dec. 2000, pp. 837–840.

[9] K. Yamaguchi, "Temperature dependence of anomalous currents in worst-bit cells in dynamic random-access memories," *J. Appl. Phys.*, vol. 87, no. 11, pp. 8064–8069, 2000.

[10] H. Kim, B. Oh, Y. Son, K. Kim, S. Y. Cha, J. G. Jeong, S. J. Hong, and H. Shin, "Study of trap models related to the variable retention time phenomenon in DRAM," *IEEE Trans. Electron Devices*, vol. 58, no. 6, pp. 1643–1648, Jun. 2011.

[11] J. P. Kim, W. Yang, and H. Y. Tan, "A low-power 256-Mb SDRAM with an on-chip thermometer and biased reference line sensing scheme," *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 329–337, Feb. 2003.

[12] C. K. Kim, J. G. Lee, Y. H. Jun, C. G. Lee, and B. S. Kong, "CMOS temperature sensor with ring oscillator for mobile DRAM self-refresh control," *Microelectron. J.*, vol. 38, pp. 1042–1049, Jan. 2007.

[13] S. S. Pyo, C. H. Lee, G. H. Kim, K. M. Choi, Y. H. Jun, and B. S. Kong, "45 nm low-power embedded pseudo-SRAM with ECC-based auto-adjusted self-refresh scheme," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2009, pp. 2517–2520.

[14] Y. Ito and H. Iwai, "Data storing method of dynamic RAM and semiconductor memory device," U.S. Patent 6 697 992, Feb. 24, 2004.

[15] S. H. Kim, W. O. Lee, J. H. Kim, S. S. Lee, S. Y. Hwang, C. I. Kim, T. W. Kwon, B. S. Han, S. K. Cho, D. H. Kim, J. K. Hong, M. Y. Lee, S. W. Yin, H. G. Kim, J. H. Ahn, Y. T. Kim, Y. H. Koh, and J. S. Kih, "A low power and highly reliable 400 Mbps mobile DDR SDRAM with on-chip distributed ECC," in *Proc. Asian Solid-State Circuits Conf.*, 2007, pp. 34–37.