Energy Reduction in SRAM using Dynamic Voltage and Frequency Management

Mohammed Shareef I, Pradeep Nair, Bharadwaj Amrutur

Abstract

This paper describes a dynamic voltage frequency control scheme for a 256X64 SRAM block for reducing the energy in active mode and stand-by mode. The DVFM control system monitors the external clock and changes the supply voltage and the body bias so as to achieve a significant reduction in energy. The behavioral model of the proposed DVFM control system algorithm is described and simulated in HDL using delay and energy parameters obtained through SPICE simulation. The frequency range dictated by an external controller is 100MHz to 1GHz. The supply voltage of the complete memory system is varied in steps of 50mV over the range of 500mV to 1V. The threshold voltage range of operation is ±100mV around the nominal value, achieving 83.4% energy reduction in the active mode and 86.7% in the stand-by mode. This paper also proposes a energy replica that is used in the energy monitor subsystem of the DVFM system.

***Index Terms*—Delay Monitor, DVFM, Energy reduction, Energy monitor, Pareto optimal curve, Replica circuits, SRAM.**

1. Introduction

Dynamic Voltage Frequency Management has been a very effective technique for optimizing the energy consumption of logic circuits [1], [2]. The same technique can be extended to memories to gain power benefits as memory accounts for a large part of the system power. In this work, we have designed a 256×64 SRAM which adjusts its operating voltage and device threshold voltage in response to a change in the operating frequency. The algorithm used to arrive at the operating point is tested using VHDL to ensure stable operation. Finally, the energy saving with our scheme is found by simulating the SRAM with the operating values obtained by VHDL simulations.

Several techniques for reducing power in SRAMs have been discussed in [3]. However at sub-micron technologies, less than 65nm, the leakage energy is a major component of the total energy, even in the active mode. The DVFM control scheme promises to optimize the total energy consumption by suitably controlling the supply voltage and the threshold voltage of the transistors in response to a change in the operating frequency.

Section II explains the SRAM design and its timing models. Section III explains the proposed DVFM algorithm and its subsystems like the delay monitor and the proposed energy replica. Section IV reports the simulation results and Section V summarizes the work.

2. SRAM design

2.1. SRAM cell sizing

The Static Noise Margin (SNM) of 6T cell depends on supply voltage, threshold voltage and trans-conductance (β) ratios $r=\beta_d/\beta_a$ and $q/r=\beta_p/\beta_d$ where β_a, β_d, β_p are the trans-conductance factors of the access, driver and pull up transistors [3]. The worst case read SNM is reached at the lowest supply voltage of 500mV and a body bias corresponding to the lowest device threshold voltage ($V_{Th\ (nominal)} - 100mV$). The cell is sized to give a reasonable SNM for the worst case. The aspect ratios of the NMOS driver, NMOS access and the PMOS pull-up transistors are $8\lambda/2\lambda$, $4\lambda/2\lambda$ and $3\lambda/2\lambda$ respectively. This sizing yields an acceptable read SNM of 85mV at 500mV and a threshold voltage of ($V_{Th\ (nominal)} - 100mV$).

The write SNM at the worst case is 200mV. To enable writing at low voltages, power-line-floating write technique is used, which also reduces power while writing [4]. The V_{DD} lines of all the cells in a row are connected through a PMOS transistor. During write the V_{DD} line of the selected row is made to float, this makes the cells in that row unstable thus easing the write operation.

2.2. Timing Signal Generation

The timing signals such as write, precharge and

1063-9667/08 $25.00 © 2008 IEEE
DOI 10.1109/VLSI.2008.47

sense clock are generated internally based on the chip select and read/write signal from the external controller. The bit-line swing should be restricted in order to avoid switching of high capacitance associated with the bit lines and large delays in reading and precharging. The bit line swing of a SRAM cell is replicated using a dummy cell, with capacitance ratio method [5], which tracks the process variations closely. The sense clock timing generation using dummy cell is shown in Fig.1. The dummy cell is designed such that its bit-line falls ten times more rapidly than the bit-line of a SRAM column. So a latch type sense amplifier is used which can amplify a differential of one-tenth of the supply voltage.

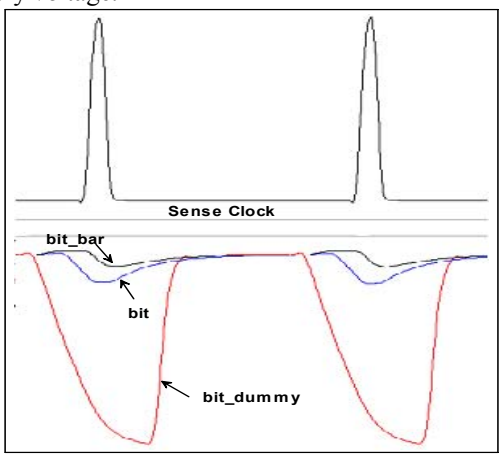

Fig. 1. Sense Clock Generation

The word line has to reset when the sense clock is activated to avoid wastage of power. This is achieved by the word-line reset circuit shown in Fig.2. So there is a dummy column in the SRAM memory array with 256 dummy cells, one for each word line.

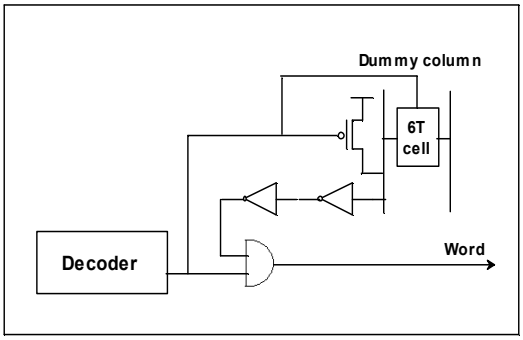

Fig 2. Word resetting circuitry

2.3. Decoder Design

Decoder design issues for Fast Low Power operation have been explained in [6]. There are a few architectures which have advantage in terms of performance and power (like DWL [7] and SCPA [8]). However, these have a greater advantage when applied to bigger SRAM blocks.

There are many different circuit styles implementing n-input AND function [6]. However, the conventional NAND implementation has the advantage of lower power as compared to the NOR implementation if the performance constraints are not tight.

The layout of a unit 6T cell, 17λ X 38λ, was considered in the modeling of the wire resistance and capacitance. The simulations were done on a 65nm PTM model files.

3. Dynamic voltage and frequency Management (DVFM)

3.1. DVFM Algorithm

The basic philosophy behind DVFM control as explained in [1] is to first find the minimum voltage at which the chip can operate, for a given frequency, and then to find the optimum body bias for which the switching and leakage currents are in a certain ratio.

The active energy consists of switching component and a leakage component.

$$E_{ACTIVE} = E_{SW} + E_{LEAK}$$

The switching and leakage energies are a strong decreasing function of V_{DD} and V_T respectively as shown in Fig. 3 and Fig. 4.

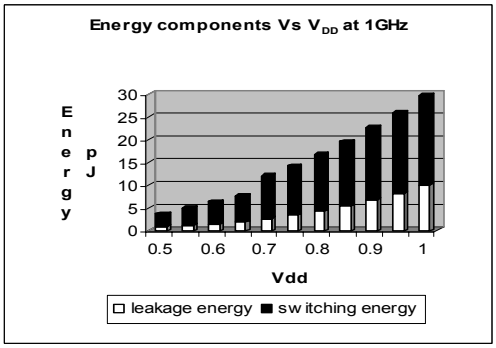

Fig.3. Energy variation w.r.t Vdd

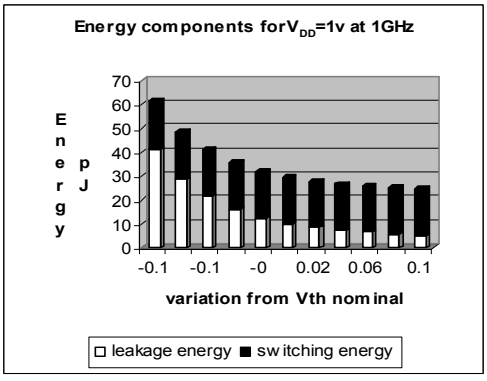

Fig.4. Energy variation Vs Vth

Our proposed DVFM algorithm for a memory system is shown in Fig. 5. The sequence of events is explained below.

1. The controller senses a change in frequency. A DVFM control cycle is initiated.

2. The supply voltage is varied in steps of 50mV until the lower bound of delay (discussed in Section 3.3) is met.

3. The body voltage is varied to change the device thresholds in steps of 100mV until the upper bound of delay (discussed in Section 3.3) is met.

4. The supply voltage is adjusted for the last time to ensure that the delay is above the upper bound by a convenient margin.

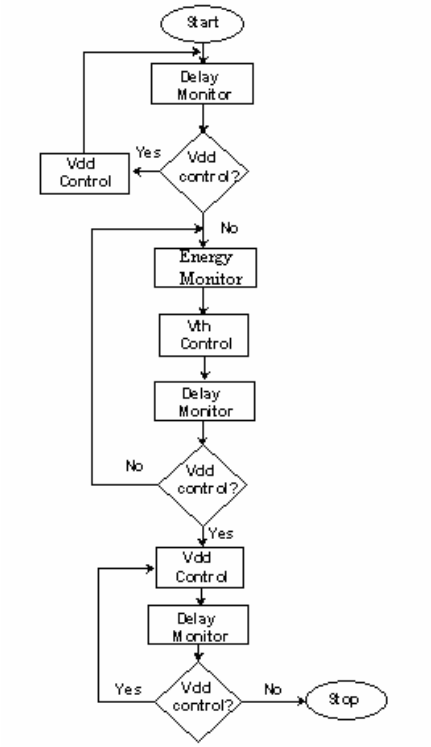

Fig.5. DVFM Algorithm for Active mode

3.2. DVFM Block diagram

The DVFM system has two main subsystems. The Delay monitor to check the delay by varying the supply voltage and the Energy monitor which monitors the energy consumed by the energy model at that operating point. Both the blocks are discussed in detail in the following section.

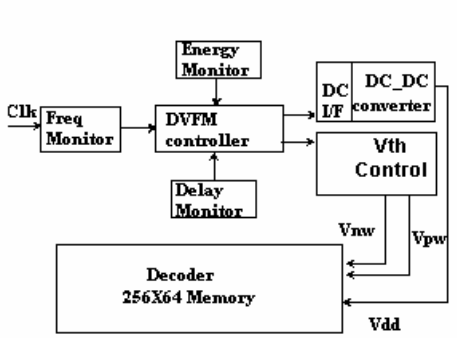

Fig.6. Block Diagram of Proposed DVFM

3.3. Delay monitor and Delay Synthesizer

The delay monitor circuit [1] [2] has a critical path delay model and buffers with 32 latches. The delay of each buffer is tailored to reflect a change in the delay monitor circuit by a Vdd step of 50mV. Each buffer in the delay line gauge has a delay of 20ps. Out of the 32 bits the first 9 bits, 180ps of delay margin, is reserved for process variations. The 21st bit and the 9th bit are, respectively, the lower and upper bounds of delay used by the DVFM algorithm.

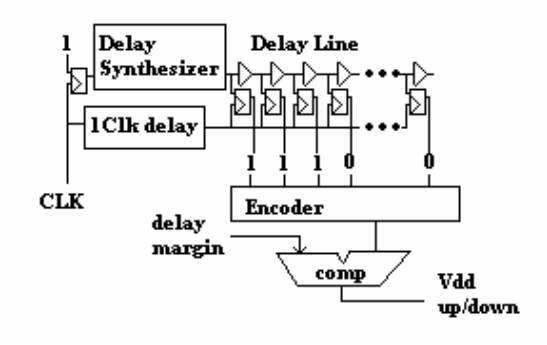

Fig.7. Delay monitor circuit

The delay synthesizer mimics the read delay from the time of arrival of address bits on to the decoder to the time data is put on to the data bus via the sense amplifiers. The circuit of delay synthesizer is shown in Fig. 8. The sense amplifier timing replica is not shown, which is very similar to the word reset replica as already discussed.

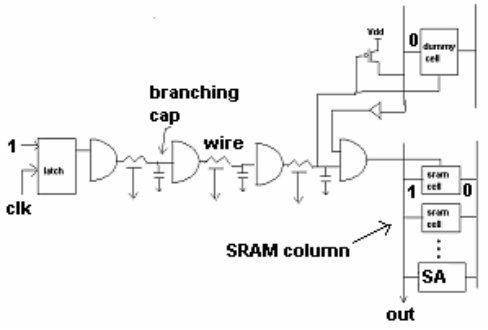

Fig.8. Delay Synthesizer.

3.4. Energy Monitor and energy replica circuit

The energy monitor circuit proposed in [1] is inaccurate because the entire circuit is mimicked by a single capacitor in the switching current monitor. Our energy monitor circuit works on the following principle. For a constant Vdd, the first step is incrementing the Vth by a step and comparing the new energy with the previous one. If the new energy is lesser than the previous then Vth is taken towards higher values. Else, it is decremented. When it reaches its nearest minima the loop settles and Vdd control is done once before exiting the DVFM loop. For every Vth step the delay is monitored and if the delay crosses the 9th bit then Vth control is stopped.

The energy monitor circuit is shown in Fig. 9. The energy monitor circuit works as follows. A capacitor C1 is charged to the supply voltage. C1 is then connected to the energy replica as its supply and the replica is operated for a particular number of read cycles. The expression $0.5*C1*(V_{INITIAL}^2 - V_{FINAL}^2)$ gives the true energy consumed by the energy replica. Now this voltage on C1 is transferred to the other capacitor C2 through a unity gain buffer [10] which is shown in Fig. 10. Now the Vth is taken to the next step and a similar read process is done on the energy model. The voltage left on the capacitor C1 is compared to the previous voltage, which is now in C2. If voltage on C1 is higher then the VTH_UP signal is generated else VTH_DOWN.

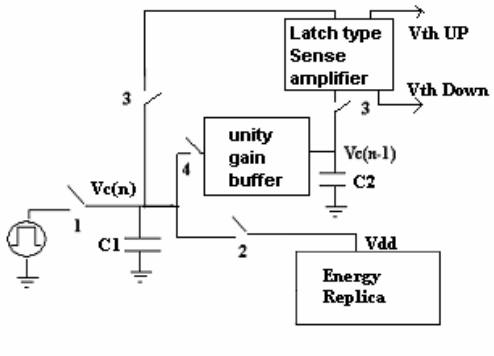

Fig.9. Proposed energy monitor circuit

A latch type sense amplifier is used to compare the voltages on the two capacitors. The capacitors are of a known value and they don't have any relation with the chip switching capacitance. The switches are closed as per the shown sequence.

TABLE I
ENERGY OVERHEAD OF DVFM BLOCKS

Entity	Energy consumed at 100MHz
Delay Monitor	51.8pJ for 11 iterations in worst case
Energy Monitor	615pJ for an average of ten steps of each ten read cycles from the energy replica.

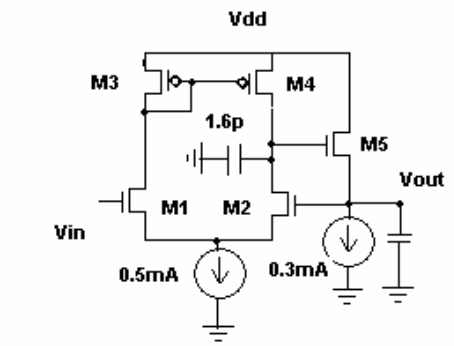

Fig.10. Unity gain buffer.

The energy replica is designed to mimic the true energy of the memory array along with the decoder faithfully. The energy replica consumes an energy 1/30th to that of the original system. Two dummy memory columns are employed in the energy replica to replicate 64 columns. Here it is to be noted that along with the energy scaling of nearly 1/30th both the switching and the leakage energies are also scaled proportionately, thus giving true measure at all operating voltages and frequencies. Similarly the decoder is also separated into switching and leaking gates appropriately and down-sized to get the same ratio as shown in Fig. 11. So the energy measured by the energy replica will be a scaled down version of the true energy. The area occupied by the energy replica is also in the same ratio which is a very small percentage (3%) of the whole memory area.

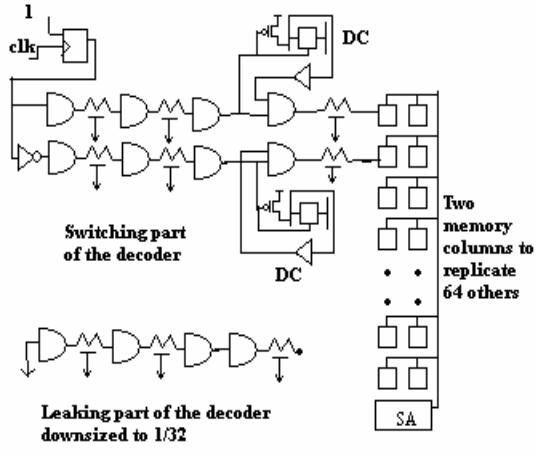

Fig.11. Proposed Energy Replica

The validation of the energy replica can be seen from Fig. 12. It shows the percentage of replica energy w.r.t the actual energy across supply voltages and across different threshold voltages. It is confined in the range 3.15% to 3.8%. This proves that the energy replica has both the switching and leakage components in the same proportion as the actual system.

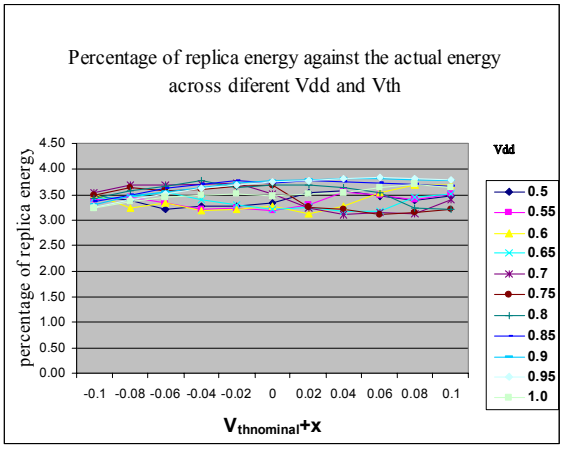

Fig.12. Plot showing the validation of replica energy across different threshold voltages and supply voltages.

3.5. Standby energy reduction

The SRAM block supply voltage is scaled down to the minimum possible voltage which is decided by the data retention voltage of the 6T cell. The SRAM cell has been sized to give acceptable noise margin at Vdd of 500mV and hold data. Then energy monitor scheme as described for the active mode is used to
vary Vth by comparing the energies consumed in the present step with the previous step. The one difference in standby mode using the energy model is to monitor the leakage energy for a particular interval of time.

3.6. DVFM system overhead

Overhead of the DVFM blocks in terms of energy is shown in Table I. The DVFM system comes into active mode of operation only when there is a frequency change indication to the memory controller. Once it arrives at the optimum operating points for that particular frequency it shuts off, till the next frequency change. The Delay monitor energy is measured assuming it works for 11 V_{DD} iterations in the worst case. The Energy Monitor similarly is assumed to work for an average of ten iterations of threshold step till the Vth control terminates. With these values the overhead of DVFM system can be calculated with an assumption that change of frequency occurs once in 100 clock cycles, which is reasonable. This energy overhead is 18% of the total energy saved through DVFM.

3.7. Integration of DVFM sub-systems with Memory

The layout of a 6T SRAM cell is shown in Fig. 13. In order to enable the implementation of power-line-floating write the V_{DD} lines of cells in the same row have to be grouped and separated from the V_{DD} lines of the other column. Fig. 14 gives an idea about the placement of DVFM sub-blocks like the delay monitor and more importantly the energy replica which has two replica memory columns. The column to the immediate left of the decoder is the dummy column used for resetting the word-line.

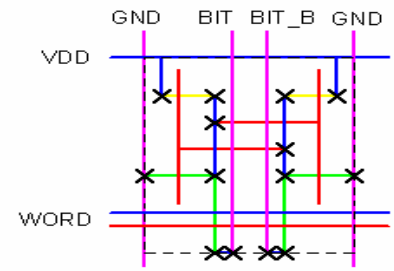

Fig.13 Layout of a 6T SRAM cell with horizontally running VDD line

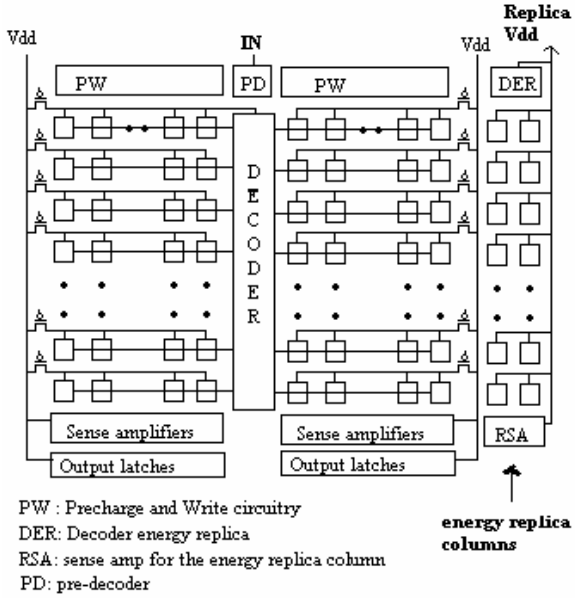

Fig.14 Placement of energy replica columns along side memory array

The supply lines of the energy replica memory columns and the decoder energy replica are connected together and given to the capacitor C1 in Fig.9. It is to be noted that the voltages of the entire memory and its peripherals are not changed until the optimum values are not arrived at and the memory is out of access for the external world. With the optimum operating values, the DVFM shuts off and these values are applied to the whole system till the next frequency change.

4. Simulation results

Fig.15 gives the energy Vs performance plot in which the pareto-optimal curve (the lower boundary) is seen, where the sensitivities of Vdd and Vth are same. The DVFM system tries to choose the optimum Vdd and Vth points on this curve. This curve was obtained by plotting energy values Vs the corresponding inverse delay values that is a representation of performance, which were obtained from exhaustive SPICE simulations of the entire SRAM cell array.

The two light gray point encircled white (lower left side among the points) on the Fig. 15 were the operating point chosen by the DVFM algorithm for 100MHz, which are found to be on the Pareto optimal curve. This shows that the algorithm converges to an optimum value.

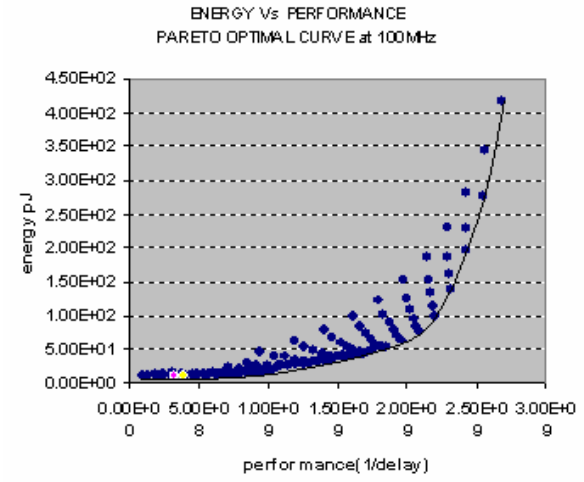

Fig.15. Pareto Optimal curve

At 100MHz operation the energy consumption at nominal Vdd and Vth is 4.35nJ for 100 cycles of read operation. With the simulated DVFM control loop in HDL with the help of SPICE values, the optimum Vdd and Vth arrived at, are 0.5V and Vthnom+40mV respectively and the energy consumed at this operating point is 0.734nJ for 100 cycles of operation. Therefore, the energy saving at 100MHz operation is 83.4% which includes the energy overhead also.

In the standby mode 86.7% energy saving is achieved with data retention at minimum operating voltage of 500mV and at a threshold voltage of Vthnom+20mV.

5. Conclusion

We have described a DVFM scheme for a 256 X 64 SRAM block for active and stand-by energy reduction. The scheme includes an iterative monitoring of delay and energy to achieve the reduction in energy. The results showed a saving of 83.4% active energy at 100 MHz and 86.7% stand-by energy. The maximum latency of the controller is found to be 46 clock cycles during the active mode control. The operating point obtained from the DVFM simulation, when plotted on a energy Vs performance scale is found to be on the Pareto optimal curve, which reiterates that the values arrived at are optimum.

6. References

[1] M. Nomura et al., "Delay and Power Monitoring Schemes for Minimizing Power Consumption by means of Supply and Threshold Voltage Control in Active and Standby Modes", IEEE Journal of Solid State Circuits, Vol. 41, No. 4, pp.805-814, April 2006.

[2] M. Nakai et al., "Dynamic Voltage and Frequency Management for Embedded Microprocessor", IEEE Journal of Solid State Circuits, Vol. 40, No. 1, pp 28-35, Jan 2005.

[3] Evert Seevinck, F. J. List and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells" IEEE Journal of Solid State Circuits, Vol. SC-22, No. 5, pp. 748-754, October 1987.

[4] Masanao Yamaoka et al., "90-nm Process-Variation Adaptive Embedded SRAM Modules With Power-Line-Floating Write Technique", IEEE Journal of Solid State Circuits, Vol. 41, No.35, pp.705-711, March 2006.

[5] Bharadwaj S. Amrutur and Mark A. Horowitz, "A Replica Technique for Word line and Sense Control in Low-Power SRAM's", IEEE Journal of Solid State Circuits, Vol. 33, No. 8, pp. 1208-1219, August 1998.

[6] Martin Margala "Low Power SRAM Circuit Design", Records of the 1999 IEEE International Workshop on Memory technology, Design and Testing, pp.115 – 122, August 1999.

[7] Bharadwaj S. Amrutur and Mark A. Horowitz "Fast Low Power Decoders for SRAMs", IEEE Journal of Solid State Circuits, Vol.36, No.10, October 2001, pp. 1506-1515.

[8] K. Itoh et al., "Trends in Low-Power RAM Circuit Technologies", Proceedings of the IEEE, Vol. 83, No. 4, pp.524-543, April 1995.

[9] M. Ukita et. al., "A Single Bitline Cross-Point Cell Activation (SCPA) Architecture for Ultra Low Power SRAMs", IEEE Journal of Solid State Circuits, Vol. 28, No. 11, pp. 1114-1118, November 1993.

[10] Behzad Razavi "Design of Analog CMOS Integrated Circuits", Tata McGraw Hill Publications.