

V.SH. MELIKYAN, N.S. EMINYAN, S.G. CHOBANYAN, N.H. BEGLARYAN

**DESIGN METHOD OF LOW-LEAKAGE HYBRID 9T-SRAM**

This paper presents a new method based on hybrid 9T cell for low-leakage SRAM design. The proposed method is based on the phenomenon that the read and write delays of a SRAM block's memory cell depend on the geometric distance of the cell from the sense amplifier and the decoder. The key idea is to use different types of 9T-SRAM cells corresponding to different threshold voltages for each transistor in the mentioned cell. Unlike other techniques for low-leakage SRAM design, the proposed method incurs no area or delay overhead. It leads only to minor change in existing SRAM design flow. The leakage current reduction achieved by using the proposed method depends on the value of the high threshold voltage, as well as the number of rows and columns in the memory cell array. Simulation results using this method show that the leakage current of an 8 Mbit SRAM block realized in 28 nm process technology is reduced by more than 30%.

**Keywords:** static random access memory (SRAM), 9T cell, low-leakage design, multiple threshold voltages.

**Introduction.** As the process technology continues to scale, leakage power has become a major concern in the design of VLSI systems. The leakage power dissipation is roughly proportional to the active area of the circuit. A significantly large segment of modern processors and SoC designs are occupied by SRAMs [1]. Therefore, the leakage power dissipation of a SRAM is one of the main components of power dissipation in today's microprocessors. Many investigations and researches have been made on SRAM's leakage issue. For example, it has been proposed to use an asymmetric SRAM cells to reduce the leakage current [2]. This method takes advantage of the fact that in regular programs most of the bits in data-cache and instruction-cache are zero. It has been also presented a forward body-biasing method for active and standby leakage power reduction in cache memories by including the device-level optimization into circuit-level techniques [3]. In [4], a dynamic threshold voltage technique to reduce the leakage power in SRAMs is proposed. By this method, the threshold voltage of the transistors of each cache line is controlled separately using body biasing. A diode-connected PMOS bias transistor to control the virtual ground [5] is also proposed to use. Although many techniques have been proposed to address the problem of low-leakage SRAM design, most of them result in hardware overhead and hence increase chip's area and reduce the manufacturing yield.

In this paper nine-transistor SRAM configuration is used. As demonstrated in [6], 9T memory cell can be an alternative design to a conventional 6T cell in future highly-integrated SRAM for sub-micron technologies because of their excellent tolerance to the process variations. To be noted, compared with the 8T [7] and 10T [8] cells the 9T-SRAM cell offers significant advantages in terms of power consumption.

Here it is presented a method for low-power SRAM design based on using different types of 9T-SRAM cells with different threshold voltage assignments. The idea is that due to the non-zero delay of interconnects different cells in a memory array have different read and

write delays. Therefore, the leakage current can be reduced by using a high threshold voltage for some transistors. This method has the following main advantages over previous techniques:

- no hardware overhead is required;
- no delay overhead is produced;
- no drastic changes in design flow are required.

In addition to the listed above advantages the proposed method improves the static noise margin (SNM) under process variation.

## SRAM Design

### Memory Architecture

The designed 8 Mbit SRAM block is built up of two 4 Mbit memory blocks along with decoding sections and control logics (Fig. 1(a)). The preferred organization for SRAM is shown in Fig. 1(b). The storage array is made up of simple cell circuits arranged to share connections in horizontal rows and vertical columns. The horizontal lines (wordlines) are driven only from outside the storage array, whereas along the vertical lines (bitlines) data flow into and out of cells. A cell which can store 0 or 1 is accessed for reading or writing by selecting its row and column. The row and column to be selected are determined by decoding binary address information. For example, consider a row decoder that has  $2^m$  output lines, moreover, each new output line is given different m-bit input code. The column decoder takes n-bit input code and produces  $2^n$  bit line access signals, of which any of them can be used at one time. The bit selection using a multiplexer (Mux) to direct the corresponding cell outputs to data registers is made.

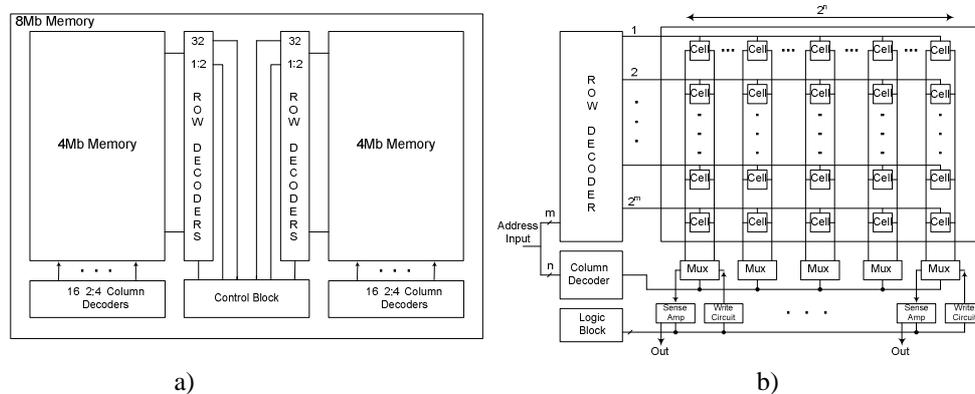


Fig. 1. Detailed view of a) 8 Mbit SRAM Structure; b) SRAM architecture

### SRAM Cell

As it has been already mentioned, the nine-transistor configuration shown in Fig. 2(a) was chosen for the cell array. The conventional 6T-SRAM cell (Fig. 2(b)) is too unstable for deep sub-micron technologies, since it fails to meet operational requirements due to a low read SNM. For 9T structure the SNM is improved by isolating the read and write operations. As seen from the 9T-SRAM structure, the configurations from M1 to M6 transistors are the same for both 9T- and 6T-SRAM cells. Since 9T structure consumes higher leakage current

compared with conventional 6T cell due to additional devices, the M7, M8, and M9 transistors have been always used with higher threshold voltages.

Generally, all SRAM cells used in a SRAM block are identical, i.e. the transistors with the same name in all SRAM cells have the same properties, that is - width, length, threshold voltage, etc. However, as it will be shown below, using non-identical cells even with the same layout footprint can realize more efficient power designs.

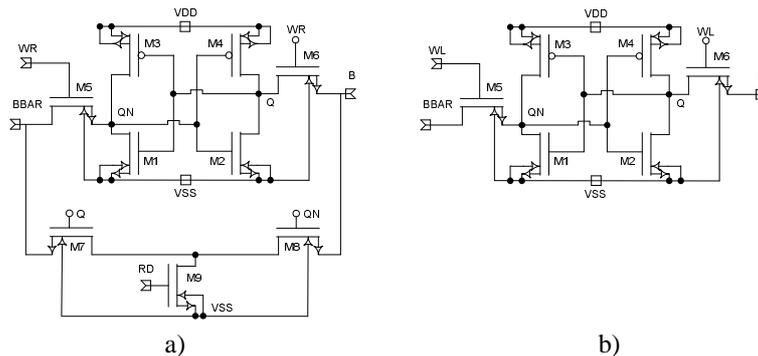


Fig. 2. SRAM cell's schematic view a) 9T; b) 6T

### *Sense Amplifier*

A sense amplifier circuit which is active only during the read operation is used to read the data from the cell. In addition, it helps reduce the delay time and minimizes power consumption in the overall SRAM chip by sensing a small difference in voltage on the bitlines.

### *Precharge Circuit*

The function of the precharge circuit is to charge the bitlines to a high voltage near supply voltage (VDD). The precharge enables the bitlines to be charged high at all times except during write and read cycles.

### *Address Decoder*

In order to significantly reduce the power consumption in SRAM block for designing decoders the following techniques have been used:

- dynamic decoder usage. Using this type of decoder will reduce the number of transistors and will increase the speed of circuit [9];
- multi-stage decoding circuit techniques usage [10]. The decoder is implemented in a tree structure by which only specific paths along the decoder will be active.
- *Control Unit*

The control unit generates internal signals of the SRAM and allows the data to either be written into the memory cell, or it passes the data from the bitlines onto the sense amplifier.

### **Hybrid Cell SRAM**

Due to parasitic parameters of the address decoder's, wordlines', bitlines', and the column multiplexer's interconnects, read/write delay time of SRAM cells in memory core are different. It is clear that read time of the closest cell to the address decoder and the column multiplexer is less than that for the farthest cell. This gives an opportunity to reduce the leakage power consumption of the memory by increasing the threshold voltage of some

transistors in SRAM cells. It is also clear that due to the delay of sense amplifiers and output buffers in read path, the read delay time of SRAM cell is higher than its write delay time. Knowing that by increasing the PMOS transistors' threshold voltage of two cross-coupled inverters in a SRAM cell will increase the write delay without much effecting on its read delay, gives an opportunity to increase the mentioned transistors' threshold voltage until the write time is below some target value for reducing the leakage power.

Since for each additional threshold voltage one more mask layer is needed in the fabrication process resulting in increasing the fabrication cost, but the benefit of having more than two threshold voltages is small [11]. In this paper only two threshold voltages are considered. However, in general, it is possible to extend the consideration for more threshold voltages.

#### SRAM Cell Configurations

To reduce the cell's leakage power consumption the threshold voltage of all or some of the transistors within the cell should be increased. Maximum reduction of leakage current can be achieved by using only transistors with high threshold voltage, but this situation has the worst effect on the access time of the cell. Thus, it is necessary to consider other configurations as well which have smaller leakage reduction and lower delay overhead.

Unlike [2], a symmetric cell configuration has been used, i.e. the symmetric transistors within a cell have the same threshold voltages. Taking into consideration, that the M7, M8, and M9 transistors in 9T-SRAM cell always will be used with higher threshold voltages, there are eight different possibilities for assigning high and low threshold voltages to the transistors within a cell (Table 1). The configurations are listed in a leakage current reduction direction; C0 is an original configuration and C7 represents a case, when the threshold voltage of all transistors is increased. With "H" symbol are marked transistors with high threshold voltage, whereas not marked transistors are with nominal threshold voltage.

Table 1

Conf.	Transistors								
	M1	M2	M3	M4	M5	M6	M7	M8	M9
C0							H	H	H
C1	H	H					H	H	H
C2			H	H			H	H	H
C3					H	H	H	H	H
C4	H	H	H	H			H	H	H
C5	H	H			H	H	H	H	H
C6			H	H	H	H	H	H	H
C7	H	H	H	H	H	H	H	H	H

All simulations have been carried on 28 nm CMOS technology with 1.0 V for the power supply voltage and 0.32V and 0.27V for the low threshold voltages for, correspondingly, NMOS and PMOS transistors at 55°C using Synopsys design environment. Fig. 3 shows the leakage current saving of mentioned configurations versus different threshold voltage values; threshold voltages of NMOS and PMOS transistors were increased until nominal values plus 0.22V.

Each of considered configuration effects read and write delays of memory cell differently. Fig.4 shows the change in read and write delays for each configuration for different

values of the high threshold voltage. It can be seen that the increase in read delay is significant for C5 and C7 configurations in case of higher threshold voltage.

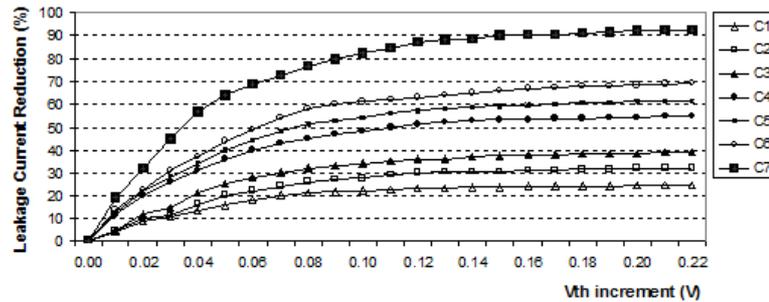
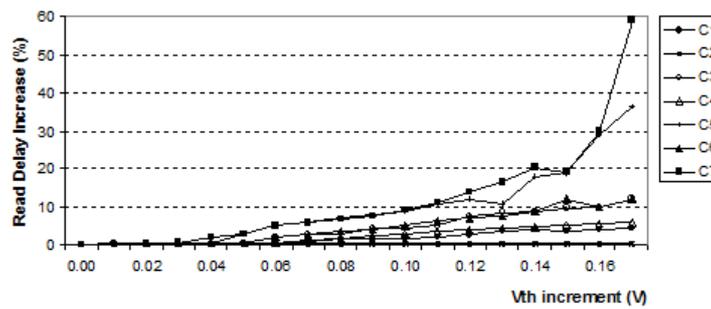
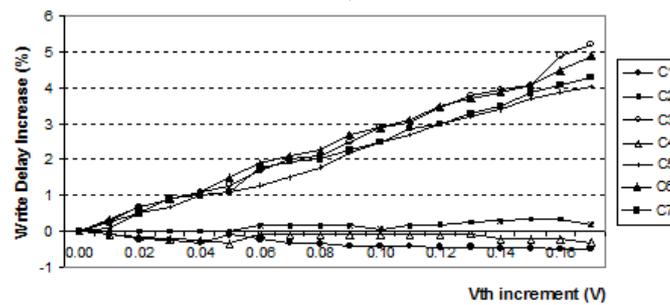


Fig. 3. Leakage current reduction for each configuration

It can also be seen that the write time increases for all configurations except C1 and C4.



a)



b)

Fig. 4. Change in a) read and b) write delay time for each configuration

#### Noise Margin

The SNM of a CMOS SRAM cell is defined as the minimum DC noise voltage necessary to flip the state of a cell [12]. The simulation results show that all configurations provide higher nominal SNM than that of original C0 configuration except C2. Moreover, in

case of all these mentioned configurations the SNM is improved with increasing the high threshold voltage. For C2 configuration, however, the SNM is slightly less than that of C0 and degrades with increasing the high threshold voltage.

Due to intrinsic parameter fluctuation the performance and yield of the device is greatly affected. To simulate this effect, the statistical analysis of the 9T-SRAM cell with all considered configurations is carried out. For this purpose Monte Carlo simulation with 1000 samples is carried out with 15% deviation in the threshold voltage. The mean and standard deviation of SNM for different configurations are shown in Table 2 for 0,42V (NMOS) and 0,37V (PMOS) high threshold voltage cases. As it can be seen, all configurations are better  $\mu-3\sigma$  than original C0 configuration.

#### Hybrid Cell Assignment

Fig. 5 shows the main part of the algorithm of the hybrid cell assignment. At first the maximum read and write delays of the memory built up with only low threshold transistors are found. The maximum read and write delay time is labeled as  $RD_{max}$  and  $WR_{max}$ , correspondingly.

*Table 2*  
Mean and standard deviation of SNM  
for different configurations

Conf.	$\mu$ (V)	$\sigma$ (mV)	$\mu-3\sigma$ (mV)
C0	0,261	21,4	196,8
C1	0,284	25,9	206,3
C2	0,252	13,9	210,3
C3	0,310	16,1	261,7
C4	0,278	16,7	227,9
C5	0,339	18,2	284,4
C6	0,302	21,9	236,3
C7	0,335	23,1	265,7

Since C7 configuration results in the highest leakage reduction among all configurations, it is replacing as many C0 cells as possible with C7 cells in such a way that the maximum read/write delay time of the memory will not become larger than  $RD_{max}/WR_{max}$ . After that the remaining C0 cells with C6, C5, C4, C3, C2, and C1 are replaced.

In the above mentioned algorithm  $n$  and  $m$  are the number of rows and columns of the SRAM, respectively. The used configuration is noted with a label.  $RD(i, j) \{c\}$  and  $WR(i, j) \{c\}$  subprograms return the read and write delay time of 9T-SRAM Cell  $(i, j) \{c\}$  cell with  $c$  configuration.

If Cell  $(i, j) \{c\}$  cell fails working at least with the last C1 configuration, then all next cells in the same  $i$  rows fail with the same configuration. The same principle is applied for the  $j$  column. The mentioned conditions are checked with  $r_j$  and  $r_i$  Boolean parameters.

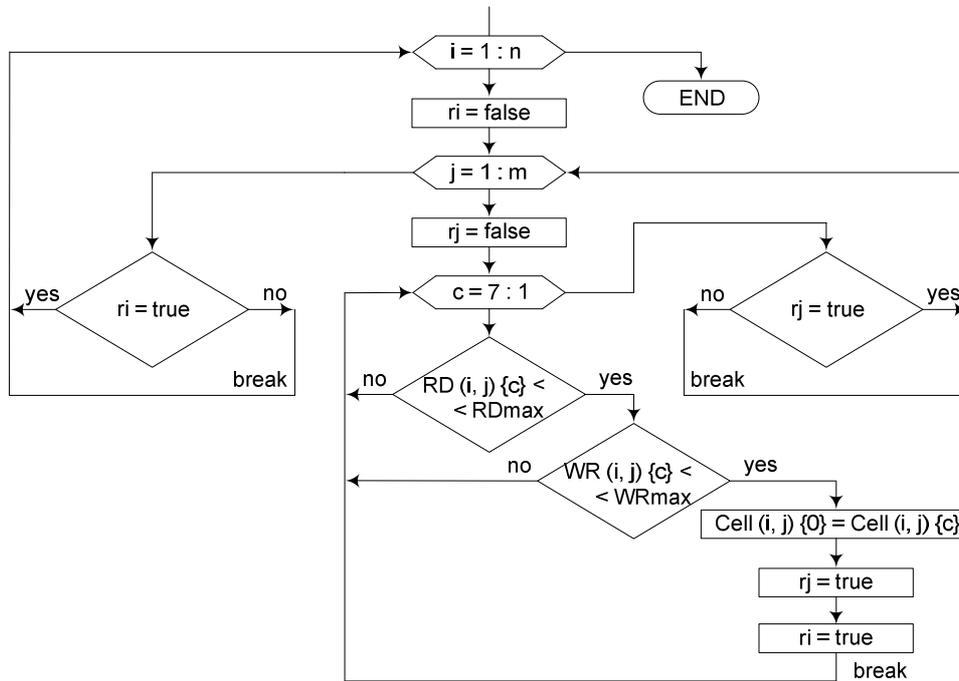


Fig. 5. The algorithm of the hybrid cell assignment

Note that for speeding up the replacement process, instead of checking for possible replacement on each single 9T-SRAM cell, one can select  $k \times k$  block and do the checking for the slowest cell in the block. If the slowest cell passes delay test, the whole block is replaced; otherwise, next configuration or block is examined. It is clear that choosing a larger number for  $k$  decreases the design time, but degrades the result.

### Simulation Results

For evaluating the efficiency of the proposed technique, a 700MHz, 8 Mbit SRAM block built with two 4 Mbit sub-blocks is designed. The design of memory is realized using 28 nm CMOS technology with 1V for the power supply voltage and 0.32V (NMOS) and 0.27V (PMOS) for the low threshold voltages. The simulation is performed with Synopsys XA [13] simulator. Table 3 shows the leakage power reduction achieved and the usage of each configuration for different values of the high threshold voltage. As it is seen from the data, the leakage saving is not very sensitive to the threshold voltage because of 0,05V increment in the threshold voltage.

Table 3

The leakage reduction and the usage of each configuration

$V_{th\ inc.}$ (V)	Leakage Reduction (%)	Usage of each Configuration (%)							
		C0	C1	C2	C3	C4	C5	C6	C7
0,01	16,18	11,3	0	3,9	0	2,1	1,2	3,6	77,9
0,02	28,07	11,4	0	7,1	0	6,7	0	7,6	67,2
0,03	35,79	11,4	0	7,6	0	15,2	0	11,3	54,5
0,04	40,55	11,3	0	13,1	0	17,2	0	10,0	48,4
0,05	43,13	11,8	0	13,6	0	22,2	0	16,7	35,7
0,06	39,68	11,8	0	19,1	0	24,7	0	34,7	9,7
0,07	37,89	11,8	0	26,5	0	37,1	0,2	23,3	1,1
0,08	37,04	11,9	0	27,1	0	53,4	0,3	7,1	0,2
0,09	36,13	11,9	0	31,4	0	54,0	0,3	2,4	0
0,10	37,25	12,2	0	32,5	0	55,3	0	0	0
0,11	37,11	12,3	0	36,5	0	51,2	0	0	0
0,12	36,02	13,0	0	41,9	0	45,1	0	0	0
0,13	34,94	13,9	0	44,7	0	41,4	0	0	0

From the table above it can be seen, that C5 configuration is rarely used in the low-leakage SRAM. The reason is that the delay of C5 is almost equal to the delay of C7, but its leakage is higher (Fig. 3 and Fig. 4). Therefore, in most cases, C7 is used instead of C5. Similarly, C6 and C4 are used instead of C3 and C1, respectively. Since C1, C3, and C5 are dominated over other configurations, they may be deleted from the set of configurations. The number of iterations will be reduced as a result. But if a smaller number of configurations are used, saving in the leakage current will decrease. Hence, there is a trade-off between memory design time and leakage saving. It is clear also that the selection of suitable configurations depends on high threshold voltage value.

Fig. 6 shows the leakage current reduction versus different values of the high threshold voltage when only four configurations (C2, C4, C6, and C7) are used along with original C0. As it is seen, in most cases the leakage saving is more than 30%.

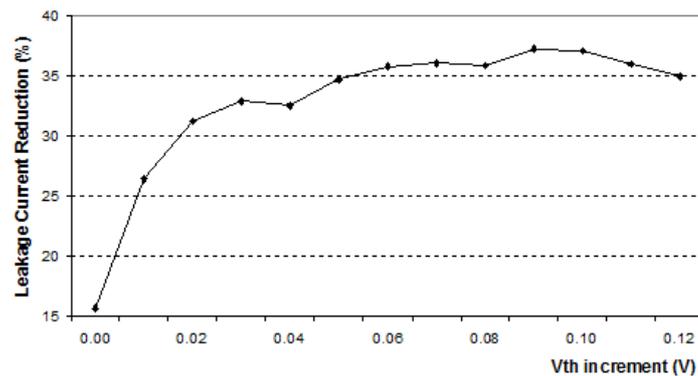


Fig. 6. The leakage reduction of SRAM with five configurations usage

**Conclusion.** In this paper a new technique for 9T-cell based on low-leakage SRAM design is presented. The proposed method is based on the phenomenon that the read and write delays of a SRAM block's memory cell are not identical and depend on the geometric distance of the cell from the sense amplifier and the decoder due to the non-zero delay of interconnects. Thus, the threshold voltage of some transistors in the 9T-cell can be increased without degrading the memory performance. By using different SRAM cells' configurations, it was achieved a low-leakage SRAM without area or delay overhead only making a minor change in existing SRAM design flow. By applying proposed technique to an 8 *Mbit* SRAM block built with two 4 *Mbit* sub-blocks more than 30% leakage current reduction is achieved. Moreover, the proposed technique improves the static noise margin under process variation.

### References

1. **Molina C.** Non redundant data cache // Int. Symp. on Low Power Electronics and Design. - 2003. - P. 274-277.
2. **Azizi N.** Low-leakage asymmetric-cell SRAM // IEEE Trans. On VLSI Systems. - 2003. - P. 701-715.
3. **Kim C.** A forward body-biased low-leakage SRAM cache: device, circuit and architecture considerations // IEEE Trans. On VLSI Systems. - 2005. - P. 349-357.
4. **Kim C., Roy K.** Dynamic Vt SRAM: a leakage tolerant cache memory for voltage microprocessor // Proc. ISLPED. - 2002. - P. 251-254.
5. **Bhavnagarwala A.** A pico-joule class, 1GHz, 32 kB 64b DSP SRAM with self reversed bias // Proc. Symp. VLSI Circuit. - 2003. -P. 251-252.
6. **Athe P., Dasgupta S.** A comparative Study of 6T, 8T and 9T Decanano SRAM cell // IEEE Symp. on Industrial Electronics and Applications. - 2009. -P. 889-894.
7. **Sil A., Ghosh S., Bayoumi M.** A novel 8T SRAM cell with improved read-SNM // IEEE Northeast workshop on circuit and system. - 2007. -P. 1289-1292.
8. **Calhoun B., Chandrakasan A.** A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation // IEEE Solid-State Circuits. - 2007. -P. 680-688.
9. **Margala M.** Low-Power SRAM Circuit Design // IEEE Int. Workshop Memory Technology, Design, and Testing. - 1999. - P. 115-122.
10. **Hirose T.** A 20ns 4Mb CMOS SRAM with Hierarchical Word Decoding Architecture // IEEE Int. Solid-State Circuits Conference. - 1996. - P. 132-133.
11. **Sirvastava A.** Simultaneous Vt selection and assignment for leakage optimization // Proc. ISLPED. - 2003. - P. 146-151.
12. **Seevinck E.** Static-Noise Margin Analysis of MOS SRAM Cells // Journal of Solid-State Circuits. - 1987. - P. 748-754.
13. XA - User Guide. Synopsys, Inc., 2009.

SEUA. The material is received on 25.04.2011.

**Վ.Շ. ՄԵԼԻՔՅԱՆ, Ն.Ս. ԷՄԻՆՅԱՆ, Ս.Գ. ԶՈՐԱՆՅԱՆ, Ն.Հ. ԲԵԳԼԱՐՅԱՆ**  
**ՓՈՔՐ ԿՈՐՍՏԱՅԻՆ ՀՈՍԱՆՔՈՎ 9Տ ՀԻՐԲԻԴԱՅԻՆ ԿԱՌՈՒՑՎԱԾՔՈՎ ՍՏՏՏԻԿ**  
**ՕՊԵՐԱՏԻՎ ՀԻՇՈՂ ՍԱՐՔԻ ՆԱԽԱԳԾՄԱՆ ՄԵԹՈՂ**

Ներկայացվում է փոքր հոսակորստով ստատիկ օպերատիվ հիշող սարքի (ՍՕՀՍ) նախագծման՝ հիբրիդային 9Տ տարրի վրա հիմնված նոր մեթոդ: Առաջարկված մեթոդը հիմնված է այն երևույթի վրա, որ ՍՕՀՍ տարրի ընթերցման և գրառման հապաղումները կախված են զգայունության ուժեղարարից ու վերձանիչից ունեցած երկրաչափական հեռավորություններից: Մեթոդը հիմնված է տարբեր տեսակի 9Տ-ՍՕՀՍ տարրերի օգտագործման վրա՝ տարրի յուրաքանչյուր տրանզիստորի համար՝ տարբեր շեմային լարումներ: Ի տարբերություն փոքր հոսակորստով ՍՕՀՍ-ի նախագծման այլ մեթոդների՝ առաջարկված մեթոդը չի հանգեցնում մակերեսի և հապաղման մեծացման: Այն միայն պահանջում է ՍՕՀՍ-ի առկա նախագծման ընթացակարգի աննշան փոփոխություն: Առաջարկված մեթոդով կորստային հոսանքի նվազեցման չափը կախված է բարձր շեմային լարման արժեքից, ինչպես նաև հիշողության զանգվածի տողերի ու սյուների քանակից: Մոդելավորման արդյունքները ցույց են տվել, որ այս մեթոդով 28 նմ տեխնոլոգիայով իրականացված 8 Մբիթանոց ՍՕՀՍ-ի կորստային հոսանքը նվազեցվում է ավելի քան 30%-ով:

***Առանցքային բառեր.*** ստատիկ օպերատիվ հիշող սարք, 9Տ բջիջ, փոքր կորստային հոսանքով նախագիծ, բազմակի շեմային լարումներ:

**В.Ш. МЕЛИКЯН, Н.С. ЭМИНЯН, С.Г. ЧОБАНЯН, Н.О. БЕГЛАРЯН**  
**МЕТОД ПРОЕКТИРОВАНИЯ ГИБРИДНОГО СТАТИЧЕСКОГО**  
**ОПЕРАТИВНОГО ЗАПОМИНАЮЩЕГО УСТРОЙСТВА С МАЛЫМ ТОКОМ**  
**УТЕЧКИ**

Представлен новый метод проектирования статического оперативного запоминающего устройства (СОЗУ) с малым током утечки, основанный на гибридной ячейке. Метод основан на том, что времена считывания и записи ячейки СОЗУ зависят от геометрического расстояния ячейки от усилителя считывания и декодера. Ключевая идея этой статьи состоит в использовании различных типов ячеек 9Т-СОЗУ, соответствующих различным пороговым напряжениям для каждого транзистора в упомянутой ячейке. В отличие от других, предложенный метод не увеличивает занимаемую площадь и время задержки, а приводит только к незначительному изменению в существующем процессе проектирования СОЗУ. Сокращение тока утечки, достигнутое при использовании предлагаемого метода, зависит от величины высокого порогового напряжения, так же, как и от числа рядов и колонок массива памяти. Результаты моделирования с использованием предлагаемого метода для 8 Мбитного СОЗУ, реализованного по технологии 28 нм, показали, что ток утечки уменьшается более чем на 30%.

***Ключевые слова:*** статическое оперативное запоминающее устройство, 9Т ячейка, проектирование с малым током утечки, многократные пороговые напряжения.