Comparison of hardware and timing penalties for eliminating SRAM failures

IOP Conf. Series: Materials Science and Engineering **498** (2019) 012017 doi:10.1088/1757-899X/498/1/012017

Comparison of hardware and timing penalties for eliminating SRAM failures

L A Shchigorev[1,2,*] **and I I Shagurin**[1,2]

[1] National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, 115409 Moscow, Russian Federation
[2] JSC "Module", 3 Eight March 4th Street, 125190 Moscow, Russian Federation

[*]Corresponding author e-mail address: l.shchigorev@module.ru

Abstract. This article provides a quantitative estimation of timing and hardware penalties, which are the implication of different SRAM methods for failure elimination. Different ways of building fault-tolerant blocks of SRAM are discussed using error-correcting codes and self-testing – self-repair units. As criteria for evaluating hardware costs, the additional chip area required for the placement of fault tolerance is considered, the timing penalties are determined by the growth of memory access time. Comparative analysis of obtained results is made. It allows estimating the effectiveness of the considered methods of fault tolerance improvement.

1. Introduction

The reliability of modern digital systems is largely determined by the accuracy of the information stored in their storage devices. Therefore, in highly reliable systems, different methods are used to eliminate the effects of failed elements of the memory blocks. Most often, the following methods are used:

- Using error correction codes (ECC) units.
- Using built-in self-test and self-repair units in memory blocks.

The most commonly used technique of ECC is the Hamming code [1], which provides a correction value of one stored data bit of the information word. Words, which are submitted for storing in memory, are encoded with the introduction of additional bits that provide the detection and correction of erroneous bits. While reading, the decoding of the selected memory word is made with error correction, and the correct value is provided. This way you can parry a single failure in the stored word. The additional hardware penalties, in this case, are due to the increase in the bit size of the stored coded words and the inclusion of ECC encoders and decoders. An estimation of hardware costs for the implementation of Hamming encoder-decoders is introduced here [2].

The BISR implementation requires the addition of self-test units (STU) and self-repair units (SRU). If failed elements are detected during the self-test of the memory, the BISR unit connects the redundant units instead, which are placed on the crystal in the form of additional rows or columns of the memory matrix. This method is used at the stage of production control of chips in order to increase the yield of integrated circuits. To eliminate failures during the operation process, it is necessary to initiate the self-test and self-repair sessions of SRAM blocks.

The easiest way of replacing failed memory elements is to shift the data bits by one position. Figure 1 shows an example of performing such a shift in the reading channel of the word by using N multiplexers "2 in 1", where N is the bit width of the corrected word, FE – damaged element, RE –

redundant element. Thus, recording channel has the elements selectors to "1 in 2", which upon detection of a failure element shift the corresponding bits to the adjacent element. The control signals for the multiplexers and selectors are formed by the block of SRU according to the results of previously executed self-test.

BISR method is effectively used in the case of a single failure (Single Event Upset – SEU), when the corrected word (or words) has to replace only one bit. A frequently used variant of the double failure correction, when for each word two redundant elements are supplemented, one of which serves as the correction of the low part and the other for correction of the high part. In this case, it is possible to replace two failed bits, if they are located in different halves of the word.

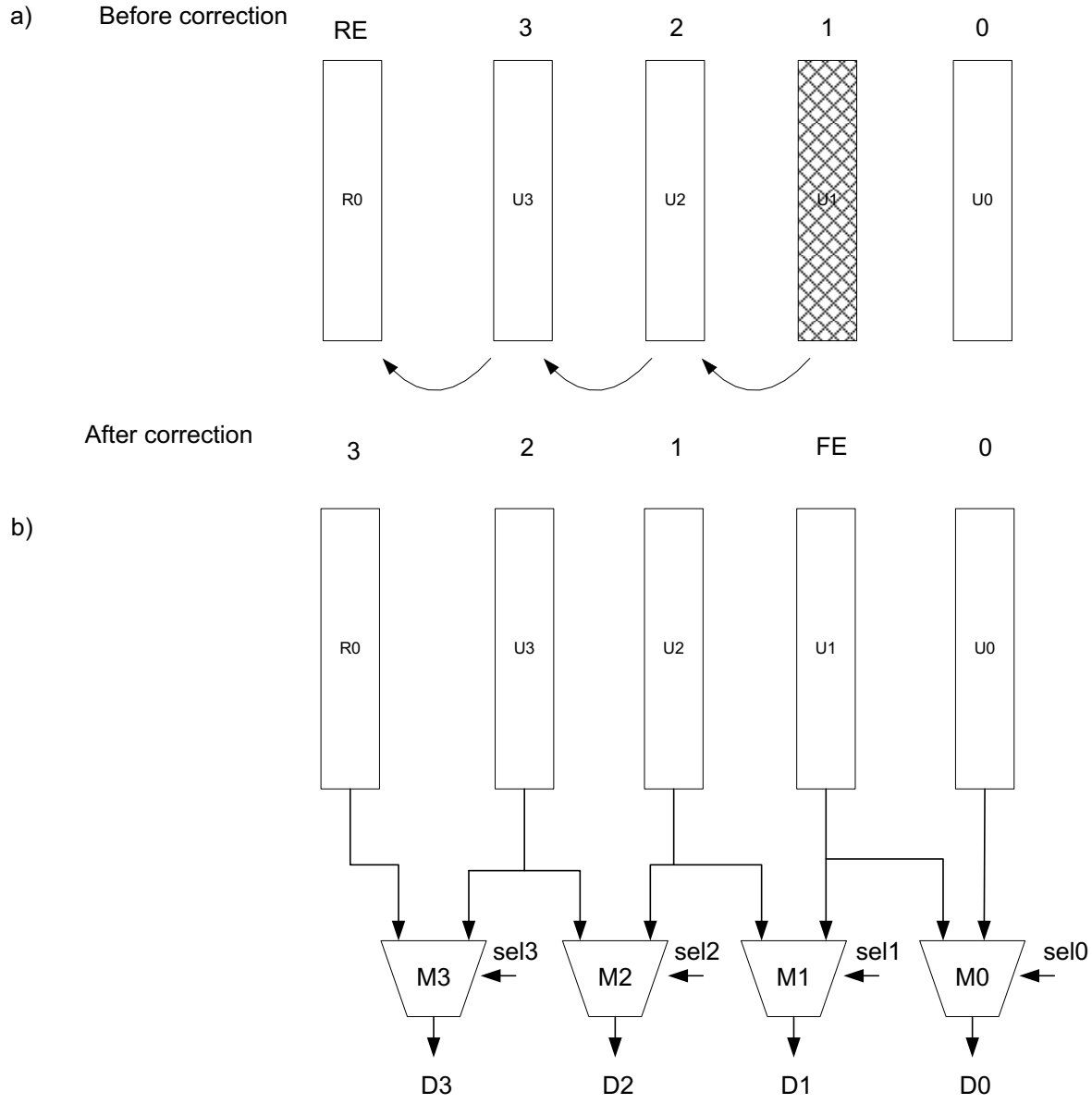

Figure 1. Replacing a failed memory element by shifting the bit numbers (a) and the scheme implementation of the shift in the read channel using multiplexers (b)

For correction using the self-repair method of multiple failures (Multiple Cell Upset – MCU), the use of multiple redundant elements for each word stored in the memory is required. The required number

of redundant elements m is determined by the number of correcting bits – maximum number bits of the word to be corrected. In this case, for correction of N-bit words when the number of failures is m, N multiplexers "(m+1)-to-1" [3] are required. Figure 2 shows an example of enabling the multiplexers in the read channel memory for correction of the failure with multiplicity m=3 in the word having N=4 bits. The channel of writing includes the selectors "1-to-(m+1)", which provides turning off of the failed element while maintaining an appropriate bit in one of the redundant elements.

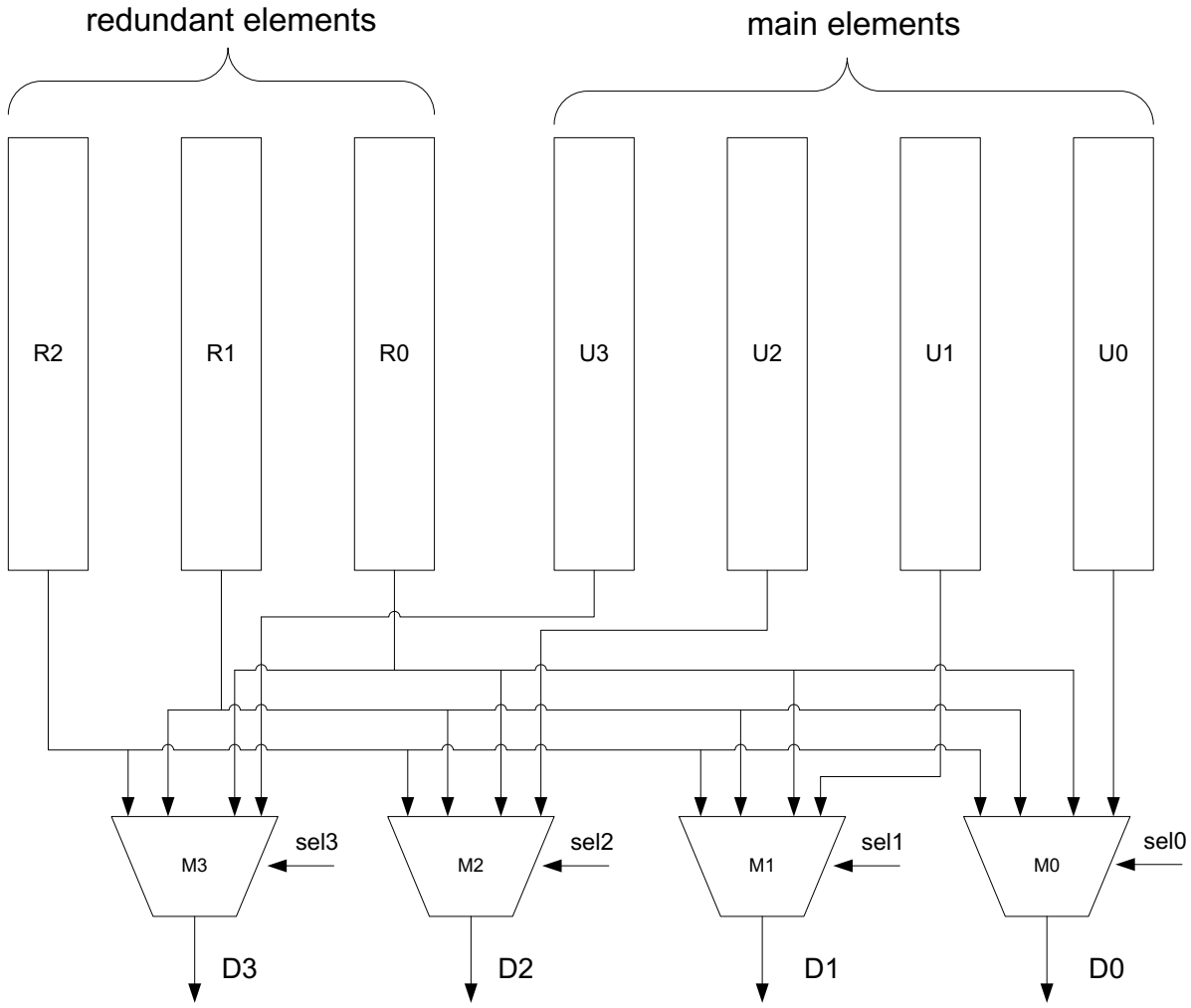

Figure 2. Implementation scheme of the multiple replacement of faulty memory elements by using multiplexers (read channel)

The additional hardware penalties during the BISR implementation caused by the introduction of the necessary number of redundant memory elements include multiplexers, which provide a replacement of the failed elements. These hardware costs lead to an increase in the area occupied on chip block memory, as well as to increased energy consumption.

In addition to hardware costs, the use of ECC methods and BISR induces an increase in data access time that is due to both to the increase of the geometrical sizes of the memory blocks and the associated parasitic capacitance, and the added combinational logic to the data communication scheme– multiplexers and encoder-decoders.

The purpose of this study was to obtain and compare quantitative estimates of time and hardware costs arising from the use of different methods of parrying of elements failures in blocks of static RAM.

2. The estimation of area and timing penalties

For the investigation, the most popular data widths were selected: 8 bits, 16 bits, 32 bits and 64 bits. The capacity of each block is 4096 (4K) words.

For eliminating single failures using ECC method the encoders-decoders of Hamming correcting single failures (ECC-SEC). Eliminating failure elements by the BISR method can be performed by replacing failed elements by a shift scheme (BISR-SHIFT) or by multiplexing bits from the redundant and the main elements in the read channel (BISR-MUXm). In the embodiment of BISR-SHIFT, each word had 2 redundant elements to parry failures in low and high parts of the word. In implemented embodiments BISR-MUXm the selected number of spare elements is m = 1, 2, and 4. Therefore, the option BISR-MUX1 provides eliminating of single failures, options BISR-MUX2, BISR-MUX4 allow to fend off the 2-fold and 4-fold failures. Table 1 shows the number of additional bits in the stored input word to implement different variants of failure elimination.

Table 1. Number of additional bits to implement different methods of failure eliminating

	Word width, bit			
Method of eliminating	8	16	32	64
ECC-SEC	4	5	6	7
BISR-SHIFT	2	2	2	2
BISR-MUX1	1	1	1	1
BISR-MUX2	2	2	2	2
BISR-MUX4	4	4	4	4

To evaluate the performance of different variants of fault-tolerant RAM blocks, their synthesis was carried out using CAD Cadence Encounter RTL Compiler during use of CMOS technology with library with the 28 nm process. For the comparative assessment of RAM blocks parameters of a similar organization, which do not have any features for parrying of failures (NC – not corrected), the above-mentioned figures were synthesized. Table 2 shows the data on the chip area required to implement the control options of blocks of SRAM with a capacity of 4K words at different bits of stored words. Figure 3 is a diagram showing the relative area increase of SRAM blocks for the cases of applying the above options to improve resiliency.

Table 2. Memory area without any features for parrying of failures

	Word width, bit			
	8	16	32	64
Area, um^2	10077	15688	26911	49356

Note that the estimations for the BISR method do not take into account the hardware cost for the implementation of the units of STU and SRU, as there are various algorithms and means of functional diagnostics of memory, the analysis of which is beyond the scope of this paper. The capacity of implementing self-test procedures can be found in [4-6], different ways of self-repair are described in [7-12].

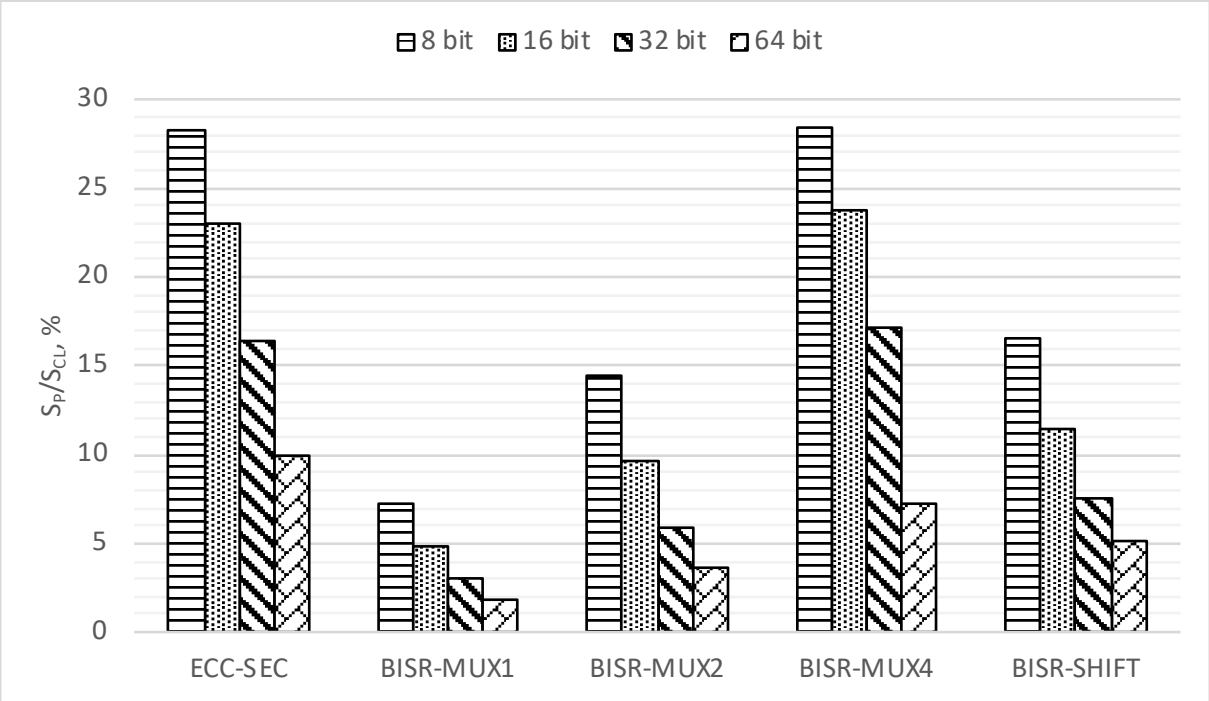

Figure 3. The relative increase of the chip area of SRAM blocks using different variants of failure eliminating

In the simulation of the synthesized variants of the SRAM blocks, the access time of the sample data with different values of the supply voltage U_{dd} and the temperature T °C is determined. The simulation was performed for the operating conditions specified in Table 3.

Table 3. Memory area without any features for failure elimination

Shorthand	U_{dd}, V	T, °C
LPLT (Low Power, Low Temperature)	0,9	-40
LPHT (Low Power, High Temperature)	0,9	125
TT (Typical Power, Typical Temperature)	1,0	25
HPLT (High Power, Low Temperature)	1,1	-40
HPHT (High Power, High Temperature)	1,1	125

Figure 4 shows the obtained estimations of the maximum access time values of sampling (data reading) with implementation of different methods for failure elimination.

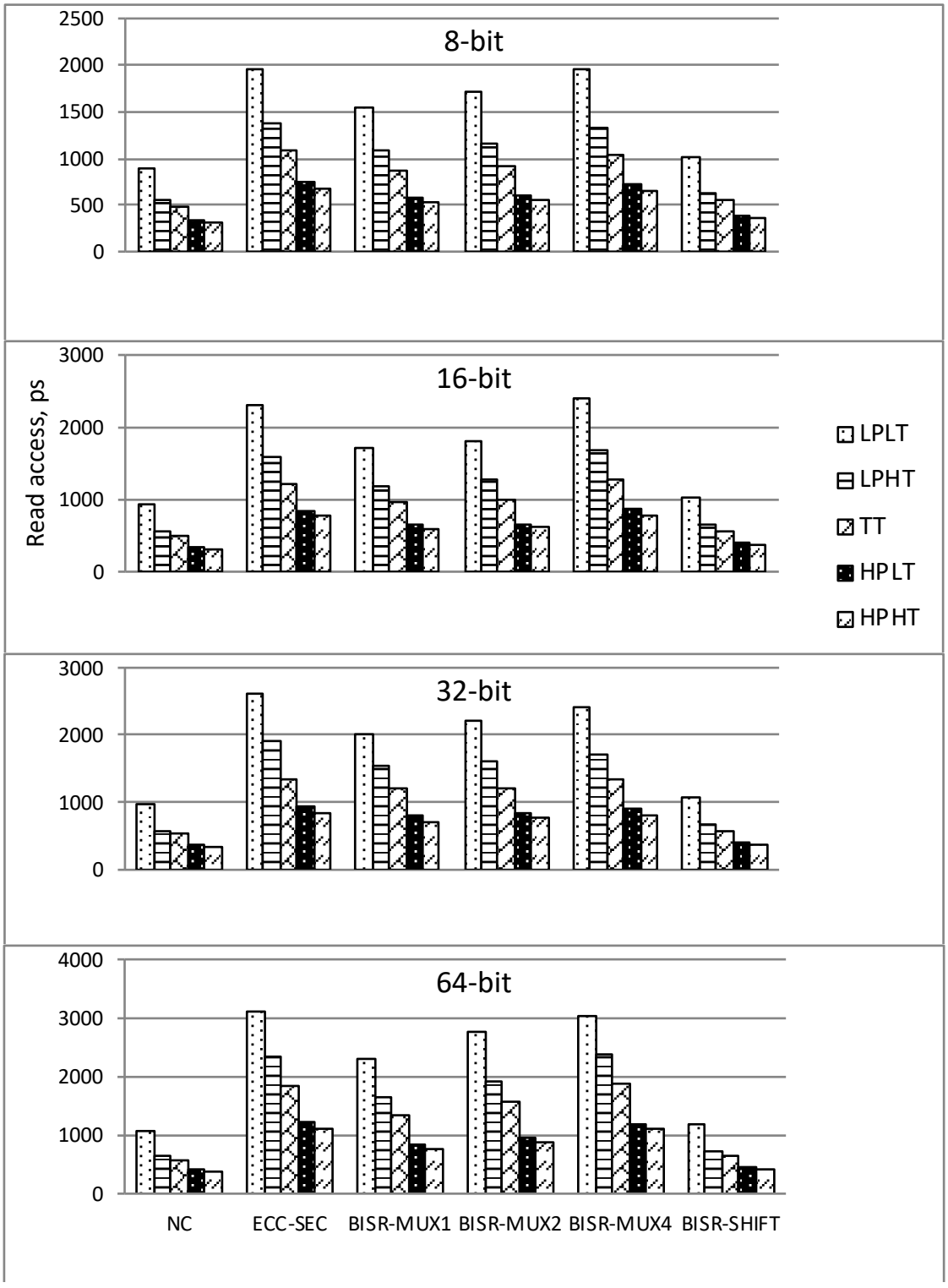

Figure 4. Access time for SRAM blocks with different word width

3. Analysis of obtained results

The analysis of the results of the synthesis and simulation of different variants of SRAM blocks implementation allows to draw the following conclusions:

- The relative cost of chip area with implementation of the methods for failure eliminating depends on the word width N and the number of additional bits. With the increase of N, the additional cost square is substantially reduced. With the value N = 32 or more bits, the increase in chip area does not exceed 17%.
- Access time depends significantly on the supply voltage and operation temperature. The maximum value of access time is observed at a lower supply voltage and temperature (worst case). The maximum and minimum access time in the range of the supply voltage U_{dd} = 0,9...1,1 V and temperature T = -40 to 125 °C can differ in 2,5...3,5 times.
- Access time increases with the word width N. When N changes from 8 to 64 bit access time increase for the cases of NC and BISR-SHIFT is about 20%, and for the other SRAM – 1.5 ...1.9 times.
- With the implementation of the discussed methods of failure eliminating access time increases by 10...20% for cases of BISR-SHIFT, 2.5...3.0 times for ECC-SEC and BISR-MUX4 2.1...2.5 times for BISR-MUX2 and 1.7...2.2 times for BISR-MUX1.
- For single error elimination, BISR-SHIFT provides the access time of 1.5 times less, yet it takes 2 times more area compared to BISR-MUX1. The ECC-SEC method has 2...3 times larger access time and 1.8...2.1 times larger area compared to the BISR-SHIFT but does not require the addition of STU and SRU blocks.
- The BISR-MUX4 and ECC-SEC methods have close values of access time and additional occupied area. BISR-MUX4 provides failure elimination with multiplicity m = 4 but requires the addition of STU and SRU blocks.

4. Conclusion

The investigated methods of SRAM failure elimination have each their own features that determine the capabilities and effectiveness of their application. When choosing a method of increasing fault tolerance, it is necessary to account for such factors as the mode of SRAM operation, the requirements for its characteristics, chip area, access time, energy consumption and the nature of the occurring faults. BISR can be used if the mode of SRAM operation allows for the implementation of testing and repairing sessions. If the SRAM operation conditions require an on-speed correction of failures in the process of continuous operation of the system, the failures can be eliminated with the application of the ECC method. However, the use of ECC leads to higher hardware and time penalties, especially when responding to multiple failures. These penalties degrade the performance of SRAM and systems in which they operate. Therefore, the BISR technique is more efficient in cases where its use is allowed in the operation mode of the system.

The quantitative estimations of the parameters for the fault tolerant SRAM blocks implementing the methods of the ECC, BISR, obtained in this paper, allow you to choose the most effective elimination method with consideration of the required characteristics and type of the failures occurred.

References

[1] Hamming R 1950 Error detecting and error correcting codes *Bell Syst. Tech. J.* **29** 147-60
[2] Elchin I, Shagurin I, Osipenko P and Vasilegin B 2006 *Izvestiya vysshikh uchebnykh zavedenii. Electronika*. **4** 65-70
[3] Nicolaidis M, Achouri N and Boutobza S 2003 *Design, Automation and Test in Europe Conf. and Exh. Proc*. **1** 590-5
[4] Vardanian V and Zorian Y *IEEE Int. Workshop on Mem. Tech. Design and Testing. Proc*. **1** 62-7
[5] Hamdioui S, Van de Goor A and Rodgers M 2002 *IEEE Int. Workshop on Mem. Tech. Design and Testing. Proc*. **1** 95-100
[6] Papavramidou P and Nicolaidis M 2016 *IEEE Trans. on Comp*. **65** 2284-98

[7] Shchigorev L and Shagurin I *IOP Conf. Series: Materials Science and Engineering* **151**
[8] Lu S, Tsai C and Hashizume M 2016 *IEEE Trans. VLSI Sys.* **24** 2726-34
[9] Lu S, Yang C, Hsiao Y and Wu C 2010 *IEEE Trans. VLSI Sys.* **18** 184-93
[10] Shchigorev L 2016 *Problems of Advanced Micro- and Nanoelectronics Sys. Dev.* **3** 178-85
[11] Tseng T, Li J and Chang D 2006 *Design, Automation and Test in Europe Conf. and Exh. Proc.* **1** 51-6
[12] Kim I, Zorian Y, Komoriya G, Pham H, Higgins F and Lewandowski J 1998 *Int. Test Conf. Proc.* **1** 1112-1119