Designing a Fault Tolerant Neural Network Computing System Based On Nanoscale Electronic Elements

M.V. Makarov[a], N.S. Trantina[a]

[a] *Murom Institute of Vladimir State University, 602264,Murom, Russia*

Abstract

This article observes the potential for building neural network computing systems designed via use of nanoscale electronic elements. The theory of interrelation between the fault-tolerance index of such systems and its predetermining factors has been systematized. We have also developed an approach to analysis of properties of parallel computing systems including nanoscale electronic elements at the stage of designing computing systems for the purpose of providing the maximum fault-tolerance index. By means of computer-generated simulation, we have experimentally tested this approach and it has proved to be superior to the available methods for solving this particular problem.

Keywords: nanoscale electronic elements; high-performance computing systems; artificial neural networks; fault tolerance

1. Introduction

Creating the components of technical computing systems using the neural network architecture is the most promising but a rather difficult way of their development. The theory and practice of engineering of artificial neural networks are at the stage of formation. There is no common terminology for the phenomena arising in the process of an artificial neural network functioning as a technical object. Moreover, there is no common methodology of the numerical determining and providing the required level of reliability and an artificial neural networks performance quality [1].

Nanoscale electronics is a new and the most promising direction for the development of element base of high-performance computing systems with the parallel architecture.

In particular, the approach to building neurocomputers, using nanoscale electronic elements (memristors) in the capacity of synapses, meets the important requirements for making high-performance new-gen facilities [2]:

- Switching from von Neumann architecture to the parallel one.
- Removing the physical separation between the storage and information processing areas.
- Reducing energy consumption due to the possibility of applying a pulse frequency modulation.
- Considerably increasing the scale of integration circuits via use of the nanoscale level.
- Integrating the potential of digital and analogue methods for information processing.

The theory and practice of engineering hardware implementations of parallel computing systems with the specified reliability parameters regardless of the applied element base are being formulated and developed. One of the most crucial and unsolved problems is the development of methods for providing the required fault-tolerance level which is not achieved automatically within the specified tolerance, and in some cases it cannot be achieved in principle due to some physical peculiarities that show up within the nanometer range.

However, pursuant to the standards applicable both in Russia and abroad, fault tolerance is one of the properties of technical objects which are subject to regulation by the corresponding criteria on a mandatory basis both during the development and operation.

The first step towards solving the problem in question is to determine the interrelation between the fault tolerance index of parallel information processing systems, implemented on the basis of nanoscale electronic elements, and the type of their architecture as well as the internal and external factors which predetermine its variation.

The solution of this problem will make a great contribution to building fault-tolerant parallel architectures of computing systems on any element base, which have maximum reliability indexes along with the required computation capacity and power consumption necessary in the modern context.

Thus, this research is mainly aimed at developing a general approach to engineering fault-tolerant neural network and neuromorphic components of information processing systems. The approach takes into account the integral-to-operation distortion of information processed and the variation of parameters' values of the system's elements when affected by external and internal destabilizing factors of any nature.

2. The object of the study

In technical and scientific literature the concept of fault tolerance is defined as "the ability of a technical object to retain the required quality (accuracy) of operation within the specified tolerance under any variations of the parameters of elements or structures when affected by internal or external factors". Work [3] develops the concept of fault tolerance applied to parallel computing systems (in particular, to artificial neural networks) and it separates the factors, which predetermine variation of the

fault tolerance index of the object under examination, into internal and external. The internal factors are those that are specified by the system's structure: a qualitative and quantitative set of elements as well as interrelation between them. The external factors are those that are related to the teaching process (insufficient generalizing performance) and destabilizing factors. Various distortions of an information signal, deliberate physical impacts on the system and others can serve as such factors. For the purpose of making a complete classification all these factors can also be separated into the ones which have informational (caused when some informational processes or a person interfere in the system's information status) nature and those with material origins (physical properties of the material used for making the system's elements).

The major conclusion, which has to be drawn on the basis of the proposed classification, is that despite the reason causing distortion or full failure of a system's element, it results in a variation of the parameter's number value of this element. For this reason, it is advisable to consider only the variation value and not to pay attention to the factor that has caused this variation. This approach will make it possible to develop multi-purpose methods for providing the specified fault tolerance indexes by engineering facilities. Such methods will introduce a common practice of engineering information processing facilities and make it possible to reduce the existing dependence on the development of new nanoscale materials designed to make electronic elements with the best reliability performance.

3. Methods

3.1. Applying the structural and functional approach to analyzing fault tolerance of parallel computing systems at the design engineering stage

Earlier was defined a set of factors which determine the fault tolerance index of parallel computing systems. One of them is a structure considered as a bunch of elements and their interrelations. The scientific analysis of a technical object's structure is based on the structural approach.

The rest of the factors (teaching process, distortion of input information, variation of elements' physical properties) can be described as a variation of the qualitative values of the system's elements' parameters when these or other elements fulfill their functions. However, some functions, which have a negative impact on the indexes we are interested in, run along with the functions, which facilitate the quality growth of the system's operation (the fault tolerance index in particular). The moment of the system's overfitting when further teaching (fine adjustment of the neurons' weighting factors and biases) leads to the reduction in the accuracy of the system's information processing, may serve as an example that proves the presence of such a phenomenon.

We propose a new approach to solving the problem of analyzing fault tolerance of parallel computing systems, according to which the possibilities of the structural and functional analysis are combined. Implementing the structural and functional approach in the form of specific and general methodologies will enable us to synthesize an algorithm for fault-tolerant operation of computing systems with the parallel architecture made via use of nanoscale elements. Such an approach meets all the requirements, formulated in this work, for design engineering of such systems:

- It considers the full set of factors forming the system's fault tolerance index.
- It considers any failure as a variation of the qualitative value of an element's parameter of the system.
- It considers any physical location of failures and the time factor when they are being determined.

3.2. Developing a methodology of the structural and functional analysis of fault tolerance of parallel computing systems made via use of nanoscale electronic elements

On the basis of the proposed structural and functional approach to analyzing fault tolerance of a parallel computing system made via use of nanoscale electronic elements, we have developed a specific methodology of analyzing the system aimed at further ensuring its fault-tolerant operation. The methodology comes down to the following successive actions:

- Decomposing the system according to the functional features and defining the structural characteristics at the element level inside the system and at the level of the system's interrelation with the ambient environment.
- Modeling the subsystem's operation that ensures the teaching of the system with a successive analysis of the variations of qualitative values of the subsystem's parameters and an analysis of the impact of these variations on the system's operation as a whole.
- Modeling the subsystem's operation that ensures the input information processing with a successive analysis of the variations of qualitative values of the subsystem's parameters and an analysis of the impact of these variations on the system's operation as a whole.
- Modeling the subsystem's operation that characterizes the physical processes related to the use of nanoscale elements with a successive analysis of the variations of qualitative values of the subsystem's parameters and an analysis of the impact of these variations on the system's operation as a whole.
- Building the structural and functional model of the system that helps to reveal cause and effect relations while forming the fault tolerance index of the system.
- Forming the process of fault-tolerant operation of the computing system in question at the design engineering stage.

Analytical review of scientific materials [4-10] hasn't revealed any existing multi-purpose, complete and of practical use methods for analyzing, calculating, increasing or providing the fault tolerance index of information processing systems with the parallel architecture and made via use of nanoscale electronic element base.

At present there are methodologies which cannot fully solve the problem of analyzing, calculating and providing the specified index of fault tolerance. These methodologies have a lot of weak points; they are inconsistent with each other and hardly meet the applicable Russian and international standards regulating the reliability of information processing engineering systems.

The imperfection of the available methods is caused by a number of operational features of parallel computing systems:

- The parameters of some elements have a complex influence on the system's work capacity as a whole.
- There is no aprior information on the parameters of each neuron (weighing factors, activation functions, biases) until the teaching process is finished.
- The reliability level and the formation principle of this indicator are individual for each problem solved.
- Parallel computing systems can be inaccurate even though they function correctly.
- In most cases it is impossible to formalize the mechanism of solving a problem by the system.

Based on the theory introduced above it can be inferred that solving the fault-tolerance analysis problem with regard to parallel computing systems must be based on the interpretation of information as an object for the conversion in information systems. It helps take into account all the factors predetermining this index's values and not to delve into their origins. While the majority of available approaches suggest considering only physical phenomena and processes which carry information and cannot fully solve the problem of providing the specified reliability indexes of a computing facility.

There is a scientific novelty value in the proposed methodology due to the following peculiarities:

- The methodology based on the structural and functional approach, invariants to the structure and type of the task completed.
- The methodology based on the structural and functional approach considers all the factors predetermining the fault tolerance index.
- The methodology based on the structural and functional approach is able to provide the specified fault tolerance index at the design engineering stage without complicating the computing system that is being created.

4. Results and Discussion

A crossbar array neural network computing architecture (Fig. 1) with the use of nanoscale elements – memristors acting as connection (synapses) among the neuron layers was simulated as an experimental study of the methodology proposed in this research [11].

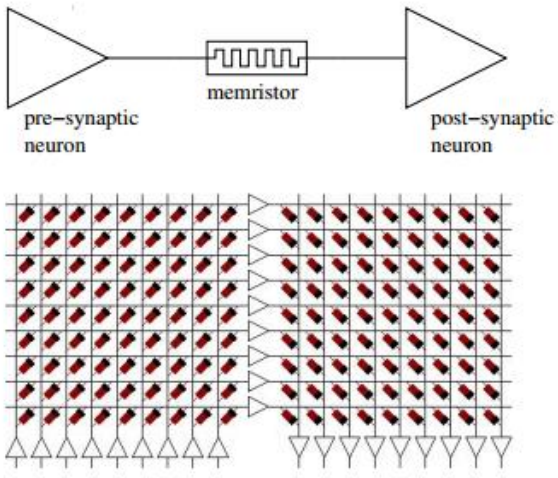

Fig. 1. A fragment of the computing memristor crossbar array controlled by CMOS transistors.

A synapse is an element that performs a weighed signal transfer (1) from the neuron of one layer to the neuron of the next layer within an artificial neural network.

$$f(x)=w\cdot x, \quad (1)$$

where $f(x)$ – input of the neuron, w – weights of the synapse, x – output of the previous neuron.

A memristor is a passive element in microelectronics; its resistance depends on a charge that has passed through it. When a voltage in the circuit is turned off, a memristor doesn't change its state and it registers the last resistance value [12, 13]. The following equations describe the voltage-current characteristic (2) and resistance (3) of a memristor:

$$v(t) = \left(\frac{R_{on} \cdot w(t)}{D} + R_{off} \left(1 - \frac{w(t)}{D} \right) \right) \cdot I(t); \quad (2)$$

$$R = \rho \cdot \frac{d}{S}; \quad (3)$$

$$w(t) = \mu \frac{R_{on}}{D} \cdot I(t); \quad (4)$$

$$M(q) = R_{off} \left(1 - \frac{\mu \cdot R_{on})}{D^2} \cdot q(t) \right), \quad (5)$$

where $v(t)$ – voltage, $I(t)$ – current, $w(t)$ – thickness of memristor's doped region, D – total thickness of a memristor, R_{on} – the minimum resistance value of a memristor, R_{off} – the maximum resistance value of a memristor, R – resistance, ρ – resistivity of the material, d – thickness of the active layer, S – contact surface, $M(q)$ – the ionic mobility.

As for an environment for the experiment, we chose a tool for modeling and simulating physical objects – the «Simscape» application software package that is controlled by the «Matlab». We synthesized simscape memristor models (figure 2) which constituted a two-layer feedforward neural network in the shape of a crossbar array; its schematic visualization is shown in figure 1. Its function is to approximate the differential equation (6):

$$y'(x) = ac + b \cdot \ln(x). \quad (6)$$

The artificial neural network was trained with the help of Neural Network Toolboxes, after that they obtained weighing coefficients were recorded in the synapse model, so that the crossbar array could be used as a computing system. The system's maximum accuracy of operation was achieved when there were 23 neurons in the first layer and one output neuron in the second layer. The neuron activation functions – the hyperbolic tangent sigmoid transfer function in the first layer and the linear transfer function in the second one. The training algorithm – Levenberg-Marquardt backpropagation. Approximate error (sum squared error performance function) amounted to $4.605 \cdot 10^{-18}$.

The generated computer model of the computing system has been divided into 3 subsystems: inputs, synapses as well as a subsystem integrating the synapses and biases. These subsystems meet the main factors which predetermine the fault tolerance index. A separate subsystem integrating all the 3 subsystems is a structure of the synthesized model of a crossbar array neural network computing architecture.

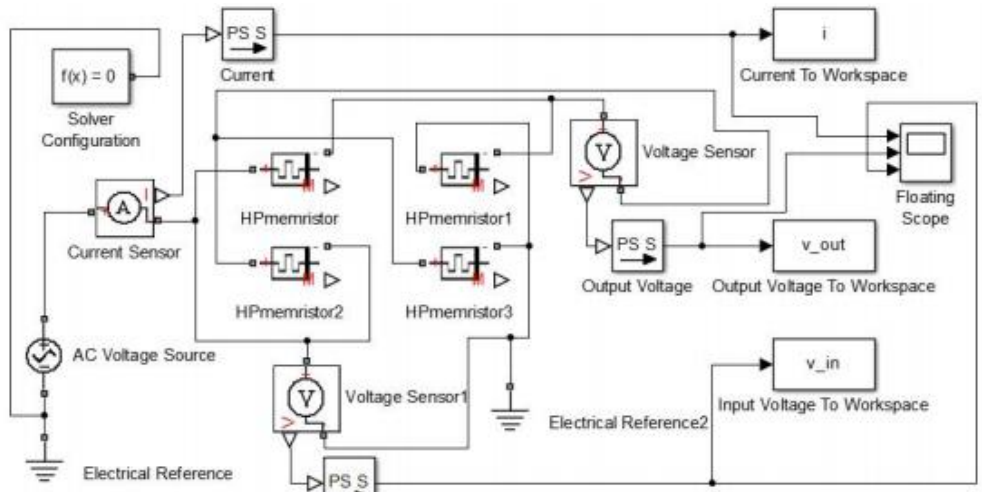

Fig. 2. Neuron synapse based on a simscape memristor model.

At the first stage we modeled the system's learning. Learning (the training system) is a process of adapting the weighting factors of the synaptic connections and the biases of every neuron. After each iteration process we analyzed both variation of the subsystem's state and impact of those variations on the structural subsystem. Then, we programmatically modeled the operation process of the taught system by submitting input information (both with and without distorted information) for its processing.

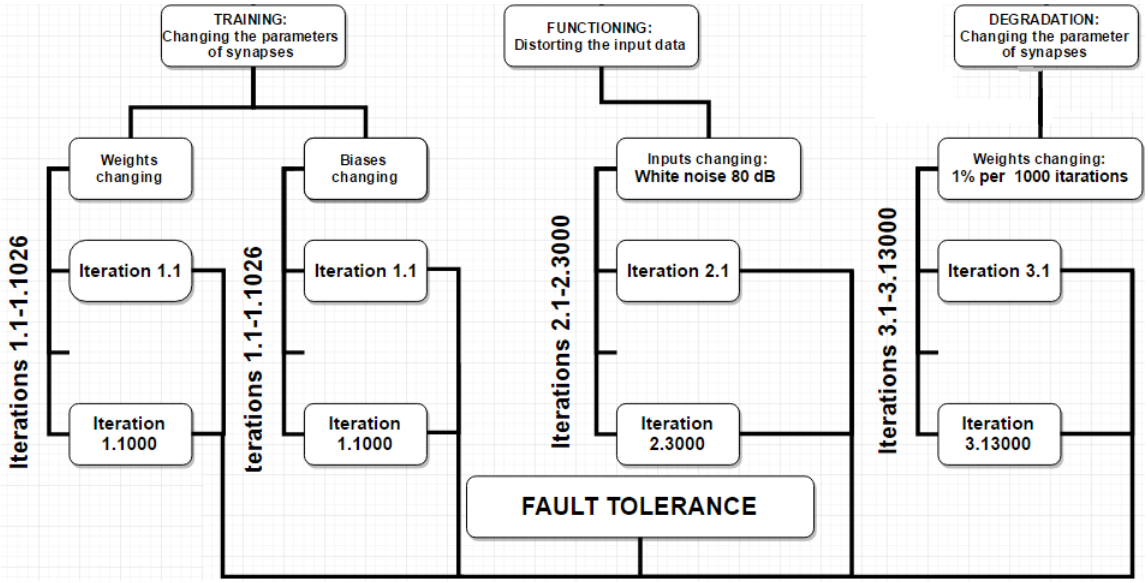

Fig. 3. The structural and functional model of the variations of the subsystems' parameters and the system's reaction to the variations in process.

This process acted as an analysis of the next subsystem – inputs. Just as in the first case we took the degree of involvement of those variations in the operation of the entire system for the experiment results. At the final stage we only considered the variations of weighting factors of the synapses connected to the physical properties of the objects in question. Such properties included the resistance variations of the memristor (3) used as a synapse related to the degradation of nanoscale material caused by multiple current passage through it due to frequent addressing to the data stored in the element. Taking into account the results of these modeling experiments we built a general structural and functional model which demonstrates the history of qualitative and quantitative variations of the subsystems' parameters and the system's reaction to the variations in process. The achieved structural and functional model is demonstrated in figure 3.

Recommendations received during the development of this structural and functional model allowed optimizing the fault tolerance of the computer model of a neural network system. Adjustments in the learning process, testing and operation of the system reflected in changes in indexes of the functioning accuracy (Table 1).

Table 1. The results of optimization of fault tolerance after studying the structural and functional model

Training the neural network				Functioning the neural network			
Type of failure	The range of iterations	Functioning accuracy	Optimized accuracy	Type of failure	The quantitative value of failure	Functioning accuracy	Optimized accuracy
Training iterations (Weights changing)	0-250	$1.521 \cdot 10^{-8}$	$8.687 \cdot 10^{-11}$	The distortion of inputs (White noise)	95 dB	$1.531 \cdot 10^{-17}$	$3.991 \cdot 10^{-18}$
Training iterations (Weights changing)	250-500	$5.779 \cdot 10^{-13}$	$7.073 \cdot 10^{-14}$	The distortion of inputs (White noise)	90 dB	$2.133 \cdot 10^{-15}$	$7.551 \cdot 10^{-16}$
Training iterations (Weights changing)	500-750	$2.801 \cdot 10^{-16}$	$2.743 \cdot 10^{-16}$	The distortion of inputs (White noise)	85 dB	$5.551 \cdot 10^{-13}$	$1.546 \cdot 10^{-13}$
Training iterations (Weights changing)	750-1000	$4.605 \cdot 10^{-18}$	$6.221 \cdot 10^{-19}$	The distortion of inputs (White noise)	80 dB	$4.524 \cdot 10^{-11}$	$2.680 \cdot 10^{-11}$
Training iterations (Biases changing)	0-350	$5.635 \cdot 10^{-9}$	$7.495 \cdot 10^{-14}$	Degradation of memristors (Weights changing)	1%	$2.461 \cdot 10^{-17}$	$3.218 \cdot 10^{-18}$
Training iterations (Biases changing)	350-700	$6.631 \cdot 10^{-14}$	$5.442 \cdot 10^{-16}$	Degradation of memristors (Weights changing)	2%	$1.222 \cdot 10^{-15}$	$8.685 \cdot 10^{-16}$
Training iterations (Biases changing)	700-1000	$4.605 \cdot 10^{-18}$	$6.221 \cdot 10^{-19}$	Degradation of memristors (Weights changing)	3%	$5.808 \cdot 10^{-12}$	$2.555 \cdot 10^{-13}$

Table 2 shows the comparison of the proposed methodology for optimization of functioning accuracy with known ones. The results of the comparison testify to the superiority of the proposed methodology on most criteria.

Table 2. The comparison of the proposed methodology of analysis and optimization of fault tolerance with known ones

	Criteria for comparison of methodologies				
Analysis of the fault tolerance index by	Optimization of the functioning accuracy	Detection of gradual failures	Detection of distortion of input data	Detection of gradual failures	The increasing complexity of the system
The proposed methodology	Yes	Yes	Yes	Yes	No
Redundancy of elements	Yes	No	No	No	Yes
Reservations of elements	Yes	No	No	No	Yes
Optimization of the neural network topology	Yes	Yes	No	No	Yes
Optimization of the generalizing ability	Yes	No	Yes	No	No
Correcting codes in modular arithmetic	Yes	No	No	No	Yes
Using the complex positional codes	Yes	No	No	No	Yes

5. Conclusion

In the course of this research we have achieved its primary objective and solved all the problems set herein. The main research result is the structural and functional approach to analyzing the fault tolerance index of computing systems with the parallel architecture at their design engineering stage. Analyzing such a model might facilitate the development of efficient methods of design engineering of parallel fault-tolerant computing systems relying on objective recommendations on the process of information processing by such a system.

Acknowledgements

The reported study was funded by RFBR, according to the research project No. 16-37-60061 mol_a_dk.

References

[1] Danilin, D.S. Design of artificial neural networks with a specified quality of functioning / D.S. Danilin, M.V. Makarov, and S.A. Shchanikov // International Conference on Engineering and Telecommunication. – 2014. – Vol.1. – P. 67-71.

[2] Danilin, D.S. The research of memristor-based neural network components operation accuracy in control and communication systems / D.S. Danilin, S.A. Shchanikov, and A.I. Galushkin// International Siberian Conference on control and communications. – 2015. – Vol.1. – P. 1-6.

[3] Makarov, M.V. Fault-tolerant operation of high-performance computing systems with the parallel architecture based on nanoscale electronic elements / M.V. Makarov // International Conference on Russian Supercomputing Days. – 2016. – Vol.1. – P. 792-801.

[4] Moritz, C.A. Fault-Tolerant Nanoscale Processors on Semiconductor Nanowire Grids / C.A. Moritz, T. Wang, P. Narayanan, M. Leuchtenburg, Y. Guo, C. Dezan, and M. Bennaser // IEEE Transactions on Circuits and Systems I: Regular Papers. – 2007. – Vol.54(11). – P. 2422-2437.

[5] Yakopcic, C. Tolerance to Defective Memristors in a Neuromorphic Learning Circuit / C. Yakopcic, R. Hasan, and T.M. Taha // Proceedings of IEEE National Aerospace and Electronics Conference. – 2014. – P. 243-249.

[6] Chabi, D. Hight Fault Tolerance in Neural Crossbar / D. Chabi, and J-O. Klein // Proceedings of 5th International Conference Design and Technology of Integrated Systems in Nanoscale Era. – 2010. – P. 1-6.

[7] Velasquez, A. Fault-Tolerant In-Memory Crossbar Computing Using Quantified Constraint Solving / A. Velasquez, and S.K. Jha // Proceedings of 33rd IEEE International Conference Computer Design. – 2015. – P. 101-108.

[8] Melouki, A. Fault-tolerance techniques for hybrid CMOS/nanoarchitecture / A. Melouki, S. Srivastava, and B.M. Al-Hashimi // IET Computers & Digital Techniques. – 2010. – Vol.4(3). – P. 240-250.

[9] Simsir, M.O. Fault-Tolerant Computing Using a Hybrid Nano-CMOS Architecture / M.O. Simsir, S. Cadambi, F. Ivancic, M. Roetteler, and N.K. Jha // 21st International Conference on VLSI Design. – 2008. – P. 435-440.

[10] Lehtonen, T. Fault Tolerance Analysis of NoC Architectures / T. Lehtonen, P. Liljeberg, and J. Plosila // IEEE International Symposium on Circuits and Systems. – 2007. – Vol.1. – P. 361-364.

[11] Kim, K.H. A functional hybrid memristor crossbar-array. CMOS system for data storage and neuromorphic applications / K.H. Kim, S. Gaba, D. Wheeler, J.M. Cruz-Albrecht, T. Hussain, N. Srinivasa, and W. Lu // Nano Letters. – 2012. – Vol.12. – P. 389-395.

[12] Chua, L.O. Memristor - the missing circuit element / L.O. Chua // IEEE Trans. Circuit Theory. – 1987. – Vol.18. – P. 507-519.

[13] Strukov, D.B. The missing memristor found / D.B. Strukov, G.S. Snider, D.R. Stewart, and R.S. Williams // Nature. – 2008. – Vol.453. – P. 80-83.