A Role-Base Approach and a Genetic Algorithm for VLAN Design in Large Critical Infrastructures

Igor Saenko
ITMO University
49, Kronverkskiy prospekt, St.Petersburg, Russia
ibsaen@comsec.spb.ru

Igor Kotenko[1,2]
[1] Saint-Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
14-th Liniya, 39, St.Petersburg, 199178, Russia
[2] ITMO University
49, Kronverkskiy prospekt, St.Petersburg, Russia
ivkote@comsec.spb.ru

ABSTRACT[1]

The use of the virtual local area networks (VLAN) technology is a well-known method of restricting access to network elements in heterogeneous network infrastructures, including critical infrastructures. The essence of the VLAN usage is to create a set of virtual subnets and in the distribution of computers on these subnets. This distribution is described by a "computer-subnet" matrix, called the VLAN-based access control scheme. The formation of an access control scheme is an NP-complete problem, and for its solution several methods are known, including genetic algorithms. However, for large critical infrastructures, the solution of this problem by known methods is very laborious due to its large dimension. The paper proposes to introduce the concept of "roles" for VLANs and use the role-based approach to optimize VLAN-based access control schemes. To solve the optimization problem, an improved genetic algorithm is proposed. Improvements are associated with the multi-chromosome representation of individuals, a new type of fitness function, a two-dimensional type of crossover operation, and a number of other aspects. The results of the experiments show the high efficiency of the proposed genetic algorithm.

CCS CONCEPTS

• **Security and privacy → Access control;** *Computing methodologies*; Heuristic function construction

KEYWORDS

VLAN, access control, genetic algorithm, critical infrastructure

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3326853

ACM Reference format:
I. Saenko and I. Kotenko, 2019. A Role-Base Approach and a Genetic Algorithm for VLAN Design in Large Critical Infrastructures. In *Proceedings of ACM GECCO conference, Prague, Czech Republic, July 2019 (GECCO'19)*, 8 pages.
DOI: 10.1145/3319619.3326853

1 INTRODUCTION

The use of VLAN technology is a well-known method of enhancing computer security in heterogeneous network infrastructures [1]. It allows one to differentiate access to network elements by forming virtual subnets in local area networks (LAN). As a result, the exchange between computers becomes possible only when they belong to the same virtual subnet [2]. Virtually all routers and switches currently have the VLAN technology capabilities. This technology does not require large computational costs and allows one to create an additional line of protection for network resources without affecting the capabilities of other security tools available on the network.

Critical information infrastructures require a high level of protection. At present, the number of security threats that are affecting critical infrastructures is dramatically increasing, and the importance of successfully protection of network resources in such systems is increasing. Since interaction with external networks (the Internet) in critical infrastructures is usually prohibited or under reliable control, internal users (insiders) are the main violators of security in such systems. An insider may accidentally or intentionally try to make an unauthorized access to another computers to which this access is denied. VLAN is designed to guarantee protection from such an attempt. Therefore, the use of VLAN technology in critical infrastructures is one of the most important means of protection.

The essence of the VLAN use is to create a set of virtual subnets and in the proper distribution of computers on these subnets. This distribution is described by the "computers - subnet" matrix, which will be called the *VLAN access control scheme*. The formation of this scheme that meets the specified safety criteria is an NP-complete task. To effectively solve this problem, heuristic methods are needed. One of these methods that has found wide application for solving similar optimization problems is genetic algorithms. In our previous works, we have shown how genetic algorithms can be successfully applied for solving optimization

problems for various access control schemes, including VLANs [3–6]. However, for large critical infrastructures, the solution of this problem by known methods [5, 6] is very laborious due to its large dimension.

The paper proposes to introduce the concept of "roles" for VLANs and use the role-based approach to optimize VLAN-based access control schemes. This is explained by the fact that in practice in the matrix "computers - subnets" of a large dimension there are always groups of computers that have the same connections with virtual subnets. Here is an analogy with the Role-Based Access Control (RBAC) model, which is used for access control in databases. As a result, when searching for a solution to the optimization problem of the VLAN access control scheme, each individual in the genetic algorithm will have not one big chromosome, but two shorter ones. This approach, by analogy with the optimization problem of the RBAC scheme [3,4], should lead to a higher speed of solving the problem.

Comparative evaluation of the results of solving the optimization problem of the Role-Based VLAN access control scheme is one of the main goals of this paper. In addition, the objectives of the paper are: (1) the mathematical formulation of the task of forming the Role-Based VLAN access control scheme and the development of methods for its solution; (2) the development and improvement of the genetic algorithm for solving the problem in the proposed formulation; (3) the implementation of a test bed to evaluate experimentally the proposed genetic algorithm and conducting this evaluation.

The main *theoretical contribution* of this paper is as follows. Firstly, the mathematical foundations to form the Role-Based VLAN access control scheme, which describe the formulation of the problem are proposed. Secondly, an improved genetic algorithm for solving the problem is suggested. Improvements include a multi-chromosomal representation of individuals, a new kind of fitness function, a two-dimensional crossover operation and some other aspects. Thirdly, the developed test bed takes into account the condition of guaranteed availability of a solution to the problem, which allows to increase the reliability of the estimates obtained during simulation.

Further structure of the paper is as follows. Section 2 provides an overview of the related work. Section 3 presents the mathematical foundations of the problem. The development of the genetic algorithm is outlined in section 4. Section 5 discusses the test bed and the results of experiments. Conclusions and directions for further research are presented in section 6.

2 RELATED WORK

The problem of optimizing the VLAN access control scheme belongs to the class of Boolean Matrix Factorization (BMF) problems. In [7, 8], it was proved that all problems of the BMF type are NP-complete. For some types of these problems, heuristic methods for solving them were proposed. Thus, in [9] the possibility to solve the BMF problems using bio-inspired population-based algorithms was analyzed. Such methods as genetic algorithms, Particle Swarm Optimization, Differential Evolution, Fish School Search and Fireworks Algorithms were considered. The analysis showed that among the considered methods the genetic algorithms are the most efficient in solving the BMF problems. In [10,11], a genetic algorithm was proposed for solving the BMF problem, in which the fitness function is based on the Euclidean distance between the initial and resulting matrices. In [12], a genetic algorithm with a specific mutation operator was proposed for solving the nonnegative matrix factorization (NMF) problem. However, despite the fact that these algorithms showed high efficiency in solving various BMF and NMF problems, they cannot be used to optimize the VLAN access control scheme. The reason is that they do not ensure the coincidence of the resulting matrices into which the original matrix is decomposed.

Analysis of the works on VLAN design shows that the formation of a set of virtual subnets is carried out, as a rule, in an empirical way without setting and solving optimization problems (as shown, for example, in [13]). This is largely due to the high complexity of such tasks. However, in some papers, the use of formal methods for solving such problems is proposed. For example, in [14] it was proposed to apply cluster analysis to form the VLAN access control scheme. However, this approach is mainly focused on use in mobile ad hoc networks. In [15], it was proposed to use a genetic algorithm to optimize the access scheme in a LAN, which uses the VLAN technology. However, this algorithm finds the matrix of computer connectivity, and not the distribution of computers on virtual subnets. In [16], the formulation of the problem of optimizing the VLAN access control scheme by the criterion of minimum energy costs was proposed. An heuristic algorithm based on k-means clustering is suggested, and it is shown that this method is more efficient than the nonlinear integer programming method.

However, this approach is not acceptable for VLAN design, as the criterion used is not safety oriented. In [17], the problem of VLAN design with multiple traffic conditions is considered. However, this task and the algorithms developed for its solution also do not take into account the security criteria. In [18], the VLAN design task is security oriented. However, its solution is related to the implementation of an application-based secure VLAN architecture based on some security mechanisms, for example, IPSec.

In our papers [5, 6], the formulation of the optimization problem of the VLAN access control scheme, based on security criteria, was developed, and quite effective genetic algorithms were proposed for solving them. However, with a large dimension of the problem, typical for large critical infrastructures, the speed of these algorithms decreases sharply.

For this reason, we suggested to apply a role-based approach for the VLAN design task, which is used in the RBAC model. In [19], it was shown that the optimization problem of the RBAC scheme, as well as the VLAN design, is related to the BMF tasks. In [3,4], it was demonstrated how genetic algorithms can be used to optimize the RBAC scheme. Thus, combining a role-based approach and a genetic algorithm should lead to development of a more efficient method for solving a VLAN design problem.

3 MATHEMATICAL BACKGROUND

3.1 Traditional Problem of VLAN Access Control Scheme Optimization

Consider first the traditional formulation of the problem of the VLAN access control scheme optimization. Suppose that in a computer network there is a set of computers $\mathbf{C} = \{C_i\}$, $i = 1, \dots, n$, and between these computers a scheme of allowed information flows, defined by the Boolean matrix $\mathbf{A}[n, n]$, is specified. If $a_{ij} = 1$ ($i, j = 1, \dots, n$), then the exchange between computers i and j is allowed. Otherwise, this exchange is impossible.

Further, we suppose that we have formed in the computer network a set of virtual subnets $\mathbf{V} = \{V_j\}$, $j = 1, \dots, k$. Each of these subnets connects two or more computers. We assign the distribution of computers across subnets using the matrix $\mathbf{Z}[n, k]$. If $z_{ij} = 1$, then computer i belongs to subnet j. Otherwise, subnet j does not cover computer i.

The matrix $\mathbf{Z}$ is associated with matrix $\mathbf{A}$ by the following expression [5]:

$$\mathbf{A} = \mathbf{Z} \otimes \mathbf{Z}^{\mathrm{T}}, \quad (1)$$

where $\mathbf{Z}^{\mathrm{T}}$ is the transposed matrix $\mathbf{Z}$, the symbol $\otimes$ denotes a Boolean matrix multiplication, which is a form of matrix multiplication based on the rules of Boolean algebra. The Boolean matrix multiplication allows to obtain the elements of the matrix $\mathbf{A}$ according to the following expression: $a_{ij} = \vee_{j=1}^{n} (z \wedge z_{ji})$.

Formula (1) is illustrated in Fig. 1. The subnet V_1 consists of computers C_1, C_2, and C_4,, and the subnet V_2 consists of computers C_2, C_4, and C_5. Between the set of computers and the set of subnets there is a map $\mathbf{Z}: \mathbf{C} \to \mathbf{V}$, between a set of subnets and a set of computers – the map $\mathbf{Z}^{\mathrm{T}}: \mathbf{V} \to \mathbf{C}$ (Fig. 1-*a*). The mapping $\mathbf{A}: \mathbf{C} \to \mathbf{C}$ is shown in Fig. 1-*b*. It is easy to see that the mapping $\mathbf{A}$ is determined by the product of the mappings $\mathbf{Z}$ and $\mathbf{Z}^{\mathrm{T}}$.

If we consider the matrix $\mathbf{Z}$ as a variable in (1), and the matrix $\mathbf{A}$ as a given one, then (1) is a formulation of the BMF problem. Indeed, the matrix $\mathbf{A}$ must be divided into two matrices $\mathbf{Z}$ and $\mathbf{Z}^{\mathrm{T}}$, the boolean product of which is equal to the matrix $\mathbf{A}$. It is easy to note that problem (1) has a large set of possible solutions. However, in the BMF problems it is required to choose among possible solutions such a matrix $\mathbf{Z}[n, k]$, for which the dimensionality k is minimal.

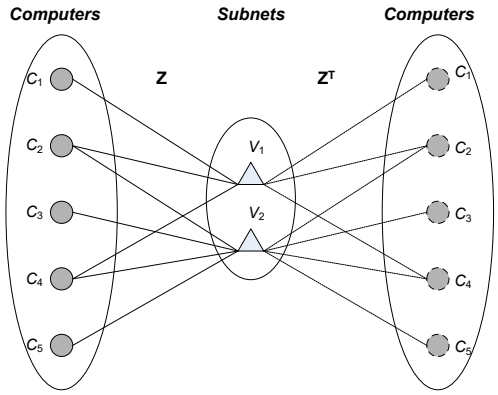
Thus, the formulation of the traditional problem of the VLAN access control scheme optimization is as follows. The initial data in this problem is the given matrix $\mathbf{A}[n, n]$.

It is required to find the matrix $\mathbf{Z}[n, k]$ such that (1) holds, and the dimension k is minimal. Formally, this statement can be written as follows:

$$\begin{cases} \mathbf{A} = \mathbf{Z} \otimes \mathbf{Z}^{\mathrm{T}} \\ Dim(\mathbf{Z}, 2) \to \min \end{cases}, \quad (2)$$

where $Dim(\mathbf{Z}, q) \to \min$ is a function that determines the q-th dimension in the matrix $\mathbf{Z}$.

a) mapping between **C** and **V**



b) mapping between **C** and **C**

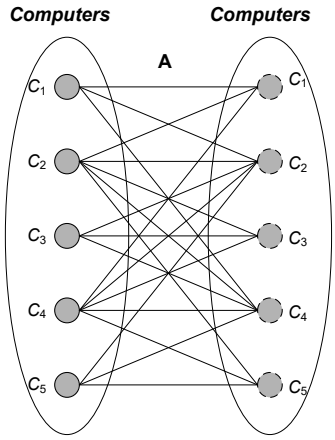


Figure 1. Example of mappings between sets of computers and subnets

3.2 Optimization of the Role-Based VLAN Access Control Scheme

If the task (1) has a higher dimension, then the elements of the set $\mathbf{C}$, as a rule, can be combined into groups. These groups we will call roles, in which each element has an equal belonging to some subset of virtual subnets $\mathbf{V}$. Denote the set of roles as $\mathbf{R} = \{R_s\}$, $s = 1, \dots, r$.

As in the RBAC model, there is a many-to-many relationship between the elements of the sets $\mathbf{C}$ and $\mathbf{R}$. We define it using the matrix $\mathbf{X}[n, r] = \{x_{is}\}$. If $x_{is} = 1$, then computer i belongs to the role s. Otherwise, the role s does not contain computer i. The "many-to-many" relationship between the sets $\mathbf{R}$ and $\mathbf{V}$ is defined using the matrix $\mathbf{Y}[r, k] = \{y_{sj}\}$. If $y_{sj} = 1$, then the role of s is associated with the subnet j. Otherwise, the role of s is not associated with the subnet j.

Matrices $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{Y}$ are related to each other using the following expression:

$$\mathbf{A} = \mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Y}^{\mathrm{T}} \otimes \mathbf{X}^{\mathrm{T}}, \quad (3)$$

where $\mathbf{Y}^{\mathrm{T}}$ is a transposed matrix $\mathbf{Y}$.

Analyzing the relations between computers and virtual subnets in Fig 1-*a*, one can see that three roles can be assigned to all computers. The role R_1 includes C_1 and C_4, the role $R_2 - C_2$ and C_4, the role R3 – C_3 and C_5. On the other hand, the roles R_1 and R_2 form the subnet V_1, and the roles R_2 and R_3 – the subnet V2. These relationships between the sets **C**, **R**, and **V** are shown in Fig. 2.

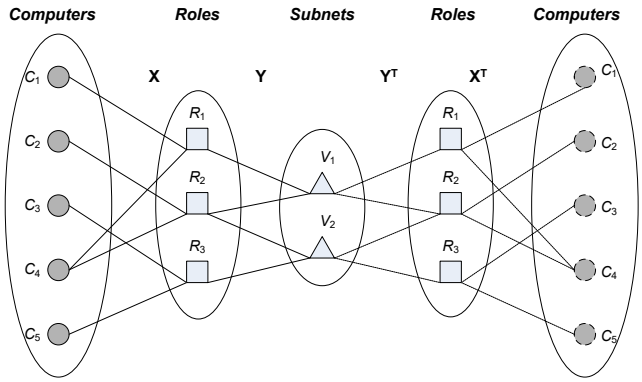


Figure 2. Example of mappings between sets of computers, roles, and subnets

We will consider the matrices **X** and **Y** as variables in (3), and the matrix **A** as the given one. Then (3) is the statement of the BMF problem, in which the initial matrix is divided into four factors. Like (1), the problem (3) has a large set of possible solutions. We will seek such solutions that the dimensions *r* and *k* are minimal.

Then the statement for the problem of optimization of the Role-Based VLAN access control scheme is as follows. The initial data in this problem is the given matrix **A**[*n*, *n*]. It is required to find the matrices **X**[*n*, *r*] and **Y**[*r*, *k*] such that (3) is true, and the dimensions *r* and *k* are minimal. Formally, this statement can be written as follows:

$$\begin{cases} \mathbf{A} = \mathbf{X} \otimes \mathbf{Y} \otimes \mathbf{Y}^{\mathrm{T}} \otimes \mathbf{X}^{\mathrm{T}} \\ Dim(\mathbf{X}, 2) \to \min \\ Dim(\mathbf{Y}, 2) \to \min \end{cases}. \quad (4)$$

Two important points should be made. First, there is a simple relationship between (2) and (4). If we assume that $\mathbf{Z} = \mathbf{X} \otimes \mathbf{Y}$, then (4) follows from (2). Secondly, the problem (4) seems to be even more complicated than (2). The validity of this statement is confirmed by the fact that only one problem of finding the optimal matrix **Y**, on the solution of which the design of RBAC schemes is based [19], is already NP-complete.

At the same time, it should be noted that if genetic algorithms are used for solving the problem (4), then this complexity will mainly consist in a more complex form of fitness function. At the same time, the presence in (4) of not one, but two variable matrices allows to implement in the genetic algorithm a multi-chromosomal approach to coding solutions. And this approach, as expected, should provide a higher efficiency of the genetic algorithm for large dimensions of the problem.

4 DEVELOPMENT OF A GENETIC ALGORITHM

To solve the problem (4), we propose an improved genetic algorithm. Its improvements are designed to provide a higher speed of solving the problem. The main features of the proposed algorithm, that distinguish it from the well-known ones, are in the following solutions: the formation of chromosomes and the initial population of solutions; fitness function; cross operations and mutations. Consider these features in more detail.

4.1 Chromosomes and Initial Population

The solution of the problem (4) is completely determined by two matrices **X** and **Y**. These matrices are independent of each other. For this reason, it is logical to use two chromosomes to encode the solutions in the genetic algorithm: the chromosome *Chr* (**X**) will encode the matrix **X**, and the chromosome *Chr* (**Y**) – the matrix **Y**. The matrix **X** has the dimension *n* x *r* (*n* is the number of rows, *r* is the number of columns). It can be represented as $\mathbf{X} = \{X_s\}$, where $X_s = (x_{s1}, x_{s2}, \ldots, x_{sn})$ is the *s*-th column of the matrix ($s = 1, \ldots, r$). Therefore, we define the chromosome *Chr* (**X**) as a string representing a set of genes in which the column X_s of the matrix **X** is used as a gene. Formally, this representation can be written as follows:

$$Chr(\mathbf{X}) = \{gene_s(\mathbf{X})\},\ gene_s(\mathbf{X}) = X_s,\ s = 1, \ldots, r. \quad (5)$$

Similarly, we form the second chromosome *Chr* (**Y**) to encode the matrix **Y**. The dimension of the matrix **Y** is *r* x *k*. The column *j* of the matrix **Y** has the form $Y_j = (y_{j1}, x_{j2}, \ldots, x_{jr})$, $j = 1, \ldots, k$. The formal representation of this chromosome is as follows:

$$Chr(\mathbf{Y}) = \{gene_j(\mathbf{Y})\},\ gene_j(\mathbf{Y}) = Y_j,\ j = 1, \ldots, k. \quad (6)$$

It can be seen from (5) and (6) that the length of the chromosome *Chr* (**X**) is equal to *r*, and the length of the chromosome *Chr* (**Y**) is equal to *k*. However, from the statement of problem (4) it is clear that these values are subject to minimization (*Dim* (**X**, 2) = *r*, *Dim* (**Y**, 2) = *k*). Consequently, both chromosomes have a variable length, which gradually decreases during the course of the algorithm, reaching a minimum by the end of its operation. This aspect is taken into account when implementing the mechanism for performing a crossover operator.

The number of individuals in the initial and in all subsequent populations of the genetic algorithm remains constant and is determined by the coefficient N_{pop}, which is a parameter of the genetic algorithm. However, a very important question arises: what should be the length of these chromosomes in the initial population? To do this, note that between the values *n*, *r*, and *k* there is the following relationship:

$$n \geq r \geq k. \quad (7)$$

In fact, the maximum number of roles can be equal to $r = n$ only when the matrix **X** is symmetric, and in it the '1'-elements are contained only on the main diagonal. But in this case there is no need to apply a role-based approach. In other cases, the condition $n > r$ is true. The validity of $r \geq k$ is proved similarly. Therefore, in the initial population, the values of *r* and *k* are

limited to the value of n. On the other hand, with a large n, the use of large values of r and k in the initial population also leads to a significant increase in the chromosome processing time. Therefore, we introduce the coefficients η and ρ, which establish a connection between n, r, and k as follows:

$$r = \eta \cdot n;\ k = \rho \cdot r; 0 \le \eta, \rho \le 1. \quad (8)$$

The coefficients η and ρ determine the size of chromosomes in the initial population. Their value is chosen empirically for each specific case of solving the problem (4). We will consider them as parameters of the genetic algorithm.

4.2 Fitness Function

Analyzing the formulation of the problem (4), we can conclude that the intermediate solutions in the genetic algorithm must first ensure the equality of the matrix product and the matrix **A**, then ensure the minimum of the value of k, then ensure the minimum of the value of r. Consequently, the fitness function can be represented as follows:

$$F = \alpha F_1 + \beta F_2 + \gamma F_3\,, \quad (9)$$

where F_1 is a function that reflects the complete coincidence of the matrix product and the matrix **A**; F_2 is the function responsible for minimizing k; F_3 is the function responsible for minimizing r; α, β, and γ are weighting coefficients that control the direction of the search for a solution. We assume that the best solutions are those whose value of the fitness function is minimal. Then between the weight coefficients, the following relationship holds: $\alpha \gg \beta \gg \gamma$. This ratio ensures that first of all there is a search for solutions for which $F_1 = 0$, then a search for solutions with a minimum value F_2 and, finally, a search for solutions with a minimum value F_3.

The F_1 function will be as follows:

$$F_1 = \sum_{i=2}^{n} \sum_{j=1}^{n-1} \left(a_{ij} - \sum_{s=1}^{k} z_{is} z_{sj}\right), \quad (10)$$

$$z_{ij} = \sum_{s=1}^{r} x_{is} x_{sj}\,. \quad (11)$$

From (10) and (11) it can be seen that the calculation of F_1 occurs through a preliminary calculation of the matrix $\mathbf{Z} = \mathbf{X} \otimes \mathbf{Y}$, which determines the relationship between the sets **C** and **V**. In addition, since matrix **A** is symmetric, the calculations in (10) are performed only on the elements if the matrix **A** lying above the main diagonal. Note also that in (10) and (11) all operations of summation and multiplication are operations 'AND' and 'OR'.

Functions F_2 and F_3 have a simpler look. They determine, respectively, the values k and r. Therefore, they are as follows:

$$F_2 = k;\ F_3 = r\,. \quad (12)$$

4.3 Crossover

Crossover allows one to get from a pair of individuals-parents, which are selected from the current population with a probability W_{cross}, new individuals - descendants. It is fulfilled through the exchange of parts of the parent chromosomes. In the traditional algorithm one chromosome is used to encode solutions, and two descendants are formed as a result of crossover. In our case, when the solution is encoded by two chromosomes, as a result of crossover, four descendant individuals are formed (Fig. 3).

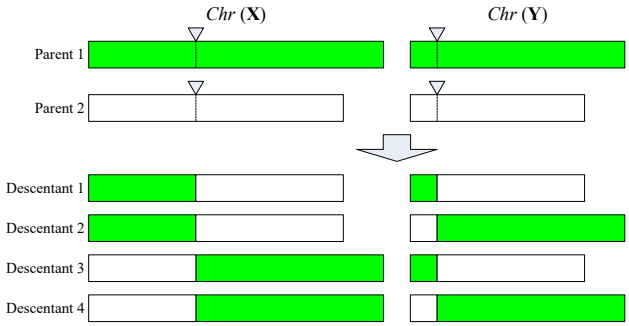


Figure 3. Formation of descendants in the operation of crossover with two chromosomes

Another improvement implemented in the crossover operation is the *two-dimensional crossover mode*. The two-dimensional crossover is aimed at obtaining descendants with zero columns, which should speed up the process of minimizing the values r and k. To achieve this goal, a crossover operation must additionally exchange the parts of genes.

In the two-dimensional mode of crossover the matrices of the parent chromosomes before crossing are divided into two parts diagonally [5]. The essence of two-dimensional crossover is clearly illustrated in Fig. 4 for two parental chromosomes displaying the matrix **Y**.

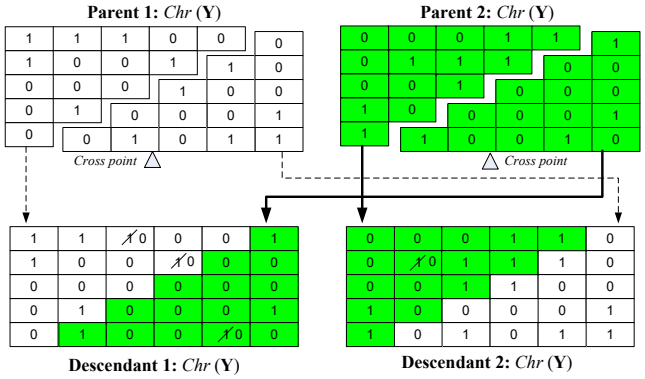


Figure 4. Example of two-dimensional crossover

As can be seen from Fig. 4, the division of the parent chromosomes is performed diagonally, the middle of the diagonal passes through the cross-point. Cross-point is the same for both parental chromosomes and is randomly selected. Crossover results in two descendant chromosomes that have only one '1'-element in some columns of their chromosomes.

This means that there is only one computer in the corresponding virtual subnet. Such subnets are not covered in the VLAN access control scheme. Therefore, these elements are reset to zero (as shown in Fig. 4). As a result, a new solution that

corresponds to a descendant individual will have a smaller number of subnets than it was in the decisions corresponding to the individual parents. In particular, the chromosome *Chr* (**Y**) in the first descendant corresponds to the scheme with three subnets, and in the second descendant - with five subnets.

4.4 Mutation

For mutation, an individual is selected from the current population with a probability W_{mut}. We propose to perform the mutation operation in two stages. At the first stage, in accordance with the traditional approach, the genes of chromosomes - the columns of matrices **X** and **Y** - are selected for mutation with a given probability W_{gen}. At the second stage, the elements of the selected columns are inverted with a probability W_{el}. W_{mut}, W_{gen} and W_{el} are the parameters of the genetic algorithm.

5 TEST BED AND EXPERIMENTS

5.1 Test Bed

To assess the developed genetic algorithm, a test bed was developed. The programming language was C #.

The test bed provides a solution to the VLAN design problem in two ways: by solving the problem in the traditional formulation (2), by solving the problem in the formulation (4), which corresponds to the proposed role-based approach. The structure of the test bed is shown in Fig. 5.

Initiator enters the initial data and parameters of genetic algorithms. The source data includes: (1) the parameters of the genetic algorithm N_{pop}, W_{cross}, W_{mut}, W_{gen}, W_{el}, η, and ρ; (2) the matrix dimensions: *n*, *r*, and *k*.

Generators, based on the values (*n*, *k*) and (*n*, *r*, *k*), obtained from the Initiator, generate the traditional reference VLAN access control scheme (the matrix $\mathbf{X_0}$) and the reference Role-Based VLAN access control scheme (the matrices $\mathbf{X_0}$ and $\mathbf{Y_0}$), as well as the required matrix $\mathbf{A_0}$ of logical connectivity of computers.

Genetic Algorithm Engine implements two types of genetic algorithms - the traditional algorithm, solving the problem (2), and the Role-Based algorithm, solving the problem (4). The input of each algorithm is the matrix $\mathbf{A_0}$. The output of the algorithms are either the resultant matrix **X** for the problem (2), or the resultant matrices **X** and **Y** for the role-based problem (4).

Evaluator produces a comparative evaluation of the developed algorithms by the operational speed and accuracy.

The test bed implements the following functions: generation of reference VLAN access control schemes; finding a solution to the VLAN design problem in two ways: traditional, solving the problem (2), and based on the role-based approach, solving the problem (4); evaluation of the effectiveness of genetic algorithms that provide two ways to solve the VLAN design problem.

The effectiveness of genetic algorithms is determined by two indicators: operational speed and accuracy. The first indicator is determined by the number of iterations of the algorithm. The second indicator shows how the resulting solutions of the VLAN task differ from the reference ones.

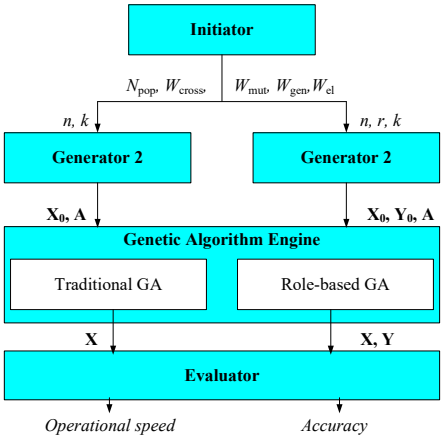


Figure 5: The structure of the test bed

As the solution factorizations' product coincides with the target matrix **A**, this indicator is calculated according to the following expression:

$$\theta = 1 - \frac{|k_0 - k|}{k_0}, \quad (13)$$

where k_0 is the number of virtual subnets in the reference VLAN access control scheme; k is the number of virtual subnets in the resulting VLAN access control scheme.

5.2 Experimental results

Experimental evaluation of the proposed genetic algorithm for solving the VLAN design problem using the Role-based approach was carried out in two modes:

(1) for a specific fragment of a LAN critical infrastructure;

(2) for examples generated by the test bed in order to determine the dependencies of the efficiency of the genetic algorithm on the dimension of the problem.

In all modes, the following values of the main parameters of the genetic algorithm were used: $N_{pop} = 200$, $W_{cross} = 0.1$, $W_{mut} = 0.01$, $W_{gen} = W_{el} = 0.5$. These values were chosen in such a way that they were equal to similar values of GA parameters used in [5, 6]. It provides compatibility of the current experimental results on GA assessment with previous ones.

5.2.1 Evaluation on a specific fragment. The structure of the specific fragment of the critical infrastructure used to evaluate the genetic algorithm is shown in Fig. 6.

In this fragment, all workstations (their number is 44) are grouped into three groups. In the center of each group there is a router. Workstations routers are connected to a server cluster router consisting of 4 servers providing various information services (Video, mail, FTP, VoIP, GIS, applications).

The required permissions for logical connectivity are divided into three groups: (1) ensuring the availability of individual workstations of each group to the servers providing various information services; (2) ensuring the availability of workstations to each other within their own group; (3) providing access to each other for some computers belonging to different groups.

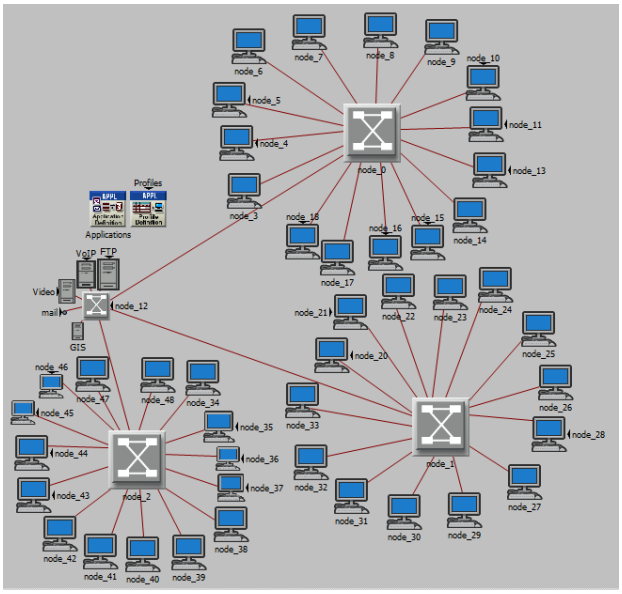


Figure 6: LAN structure for experimental evaluation

The solution of the VLAN design problem for this fragment led to the following results. By means of the traditional GA algorithm the solution of the problem, containing 22 virtual subnets, was found. By means of the role-based approach the more efficient solution, containing 20 subnets, was found. At the same time the 29 roles were created. Therefore, it is clear that the role-based approach allows one to get solutions that better meet the criterion of minimal virtual subnets. The traditional algorithm took 650 iterations to find a solution, and the role-based algorithm solved the problem in 445 iterations, or 1.46 times faster. Thus, the experiments showed that the proposed role-based genetic algorithm has significantly higher accuracy and speed of work.

5.2.2 Evaluation on generated examples. In this mode, the dependences of the efficiency indicators of genetic algorithms on the dimension of the problem were determined. The dimension n took the following values: 100, 200, 500, 1000. The dimension k was in the range from 20 to 50 percent of the value n.

The experiments were carried out according to the following scheme. First, the network size n was chosen. Then a reference Role-Based VLAN access control scheme, defined by the reference matrices $\mathbf{X_0}$ and $\mathbf{Y_0}$, was formed. These matrices allowed to form the required matrix $\mathbf{A}$ of logical connectivity of computers. Next, using the proposed genetic algorithm, the resulting matrices $\mathbf{X}$ and $\mathbf{Y}$ were found. Then, by comparing the resulting matrices with the reference ones, the accuracy index was calculated according to (13). The number of iterations T spent on the search for matrices $\mathbf{X}$ and $\mathbf{Y}$ was considered as an indicator of the speed of the algorithm because all iterations in the traditional and proposed GAs had approximately the same run time. At the same time, the traditional optimization problem of the VLAN access control scheme was solved, for which the matrix $\mathbf{A}$ was also used. The solutions found, determined by the matrix $\mathbf{X}$, were also evaluated using the accuracy and speed of the algorithm.

Table 1 depicts the results of the experiments. The following symbols are used: T_0 – an operational speed for the traditional genetic algorithm; T – an operational speed for role-based genetic algorithm; δ – a gain in the speed of the role-based genetic algorithm, $\delta = T_0 / T$; θ – an accuracy of the role-based genetic algorithm, it was determined according to (13). The values of the estimated indicators are obtained as averages on a random sample of 10 tests. At the same time the dispersion of values in statistical selection did not exceed 10 percent.

Table 1: Experimental results

n	k	T_0	T	δ	θ
100	20	1495	1053	1.42	0.89
100	30	1252	934	1.34	0.90
100	50	1094	848	1.29	0.92
200	40	1797	1102	1.63	0.85
200	60	1501	1028	1.46	0.88
200	100	1348	970	1.39	0.90
500	100	3504	1353	2.59	0.84
500	200	2501	1232	2.03	0.87
500	250	1997	1148	1.74	0.89
1000	200	9387	1618	5.80	0.82
1000	300	6478	1546	4.19	0.84
1000	500	4977	1447	3.44	0.85

Analyzing the data presented in Table 1, we can draw the following conclusions.

First of all, it should be noted that in all experiments the role-based algorithm gains in speed of operation. At the same time, this gain becomes greater if the dimension n increases and if k decreases. This is explained quite logically. For a larger dimension n, the traditional algorithm operates with one very large matrix, and the role-based algorithm with two shorter matrices. Therefore, the role-based algorithm is faster, and the magnitude of this gain is the greater, the larger the dimension n. On the other hand, if k decreases, then the genetic algorithm takes more time to reduce the length of its chromosomes to a given value. Therefore, the smaller the value of k, the greater the number of iterations required for the operation of the algorithm.

Analyzing the results of evaluating the accuracy of the role-based genetic algorithm, we can conclude that in all cases this algorithm had a sufficiently high value of this indicator lying in the range from 0.82 (with the largest dimension of the problem) to 0.92 (with the smallest dimension of the problem). At the same time, the accuracy, like the operational speed, decreases slightly with increasing dimension n and decreasing k. It may be explained by the fact that the optimization problem became more hard when n increases or k decreases. Thus, from the obtained experimental data it follows that the proposed role-based genetic optimization algorithm for the VLAN access control scheme is more efficient than other known algorithms.

The obtained experimental results allow us to make a number of recommendations for the network administrator of a large

critical infrastructure on the VLAN design. If the dimension of the task is not very high (no more than 100 workstations in the network), then the difference between using the traditional genetic algorithm and the role-based genetic algorithm is small. It is possible to use the traditional genetic algorithm. Our proposed role-based genetic algorithm has higher efficiency with higher dimensions of the problem, therefore, its use should be considered more efficient. When using the role-based genetic algorithm, the values of the parameters η and ρ, which establish the relationship between the dimensions *n*, *r*, and *k* in the matrices **X** and **Y** of the initial population, should be specified.

Experiments have shown that the values of these parameters lie in the range from 0.45 to 0.8. One can take them equal to each other. The value η, ρ = 0.45 should be chosen when we assume that the value of *k* is 20 percent of *n*. The value η, ρ = 0.8 is chosen when we assume that *k* is 50 percent of *n*.

6 CONCLUSIONS

The paper proposed a role-based approach to solving the VLAN design problem, which consists in optimizing the VLAN access control scheme. The formulated statement of the role-based VLAN design problem is a special kind of Boolean Matrix Factorization, in which the initial Boolean matrix is decomposed into two pairs of direct and transposed shorter Boolean matrices. This problem is more complex than the well-known problems of Boolean Matrix Factorization. For this reason, the use of known mathematical methods to solve it is impossible. To solve the problem, it is proposed to use an improved genetic algorithm.

The main improvements proposed in the paper are: (1) coding solutions in the form of two independent variable-length chromosomes; (2) consideration of three criteria in fitness functions that are responsible for the coincidence of the initial and resulting matrix of logical connectivity, as well as minimizing the number of virtual subnets and roles; (3) implementation of multi-chromosomal crossover of individuals in a two-dimensional mode. Experimental evaluation of the proposed genetic algorithm showed its rather high efficiency. According to the operational speed for different dimensions of the problem, it gives a gain of 1.5 to 5 times, and the greater the dimensionality of the problem, the greater the gain. It ensures a sufficiently high accuracy of the solution. Further work is the application of the proposed approach to other access control areas and investigation of the effect of further decompositions based on a hierarchy of roles.

ACKNOWLEDGMENTS

This work was partially supported by grants of RFBR (projects No. 16-29-09482, 18-07-01369, 18-07-01488, 18-37-20047, 18-29-22034, 18-37-20047, 19-07-00953, 19-07-01246), by the budget (the project No. 0073-2019-0002), and by Government of Russian Federation (Grant 08-08).

REFERENCES

[1] Mario E. Gomez-Romero, Mario Reyes-Ayala, Edgar A. Andrade-González, and Jose A. Tirado-Mendez. 2010. Design and implementation of a VLAN. In *Proceedings of the 2010 international conference on Applied computing conference* (ACC'10). Wisconsin, USA, 87-90.

[2] Xin Sun and Geoffrey G. Xie. 2016. An Integrated Systematic Approach to Designing Enterprise Access Control. *IEEE/ACM Transactions on Networking*, 24, 6(2016), 3508–3522. DOI: http://dx.doi.org/10.1109/TNET.2016. 2535468.

[3] Igor Saenko and Igor Kotenko. 2012. Design and Performance Evaluation of Improved Genetic Algorithm for Role Mining Problem. In *Proceedings of the 2012 20th Euromicro International Conference on Parallel, Distributed and Network-based Processing* (PDP '12). IEEE Computer Society, Washington, DC, 269–274. DOI: http://dx.doi.org/10.1109/PDP.2012.31.

[4] Igor Kotenko and Igor Saenko. 2015. Improved genetic algorithms for solving the optimisation tasks for design of access control schemes in computer networks. *Int. J. Bio-Inspired Comput.* 7, 2 (2015), 98–110. DOI: http://dx.doi.org/10.1504/IJBIC.2015.069291.

[5] Igor Saenko and Igor Kotenko. 2015. A genetic approach for virtual computer network design. In *Proceedings of the 8th International Symposium on Intelligent Distributed Computing* (IDC'14). Intelligent Distributed Computing VIII. Studies in Computational Intelligence, 570. Springer-Verlag, Berlin, Germany, 95–105. DOI: http://dx.doi.org/10.1007/978-3-319-10422-5_11.

[6] Igor Saenko and Igor Kotenko. 2014. Design of virtual local area network scheme based on genetic optimization and visual analysis. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications* 5, 4 (2014), 86–102. DOI: http://dx.doi.org/10.22667/JOWUA.2014.12.31.086.

[7] Ervina Çergani, Pauli Miettinen. 2013. Discovering Relations using Matrix Factorization Methods. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. ACM, New York, NY, USA, 1549–1552. DOI: http://dx.doi.org/10.1145/2505515. 2507841.

[8] Pauli Miettinen and Jilles Vreeken. 2014. MDL4BMF: Minimum Description Length for Boolean Matrix Factorization. *ACM Trans. Knowl. Discov. Data* 8, 4, Article 18 (2014), 31 pages. DOI: http://dx.doi.org/10.1145/2601437.

[9] Andreas Janecek and Ying Tan. 2011. Using population based algorithms for initializing nonnegative matrix factorization. In *Proceedings of the Second international conference on Advances in swarm intelligence - Volume Part II* (ICSI'11), Ying Tan, Yuhui Shi, Yi Chai, and Guoyin Wang (Eds.), Lecture Notes in Computer Science, Vol. 6729. Springer, Berlin, Heidelberg, 307–316. DOI: http://dx.doi.org/10.1007/978-3-642-21524-7_37.

[10] Vaclav Snasel, Jan Platos, Pavel Krömer, Dusan Husek, Roman Neruda, and Alexander Frolov. 2008. Investigating Boolean Matrix Factorization. In *Proceedings of the Workshop on Data Mining using Matrices and Tensors* (DMMT'08). Las Vegas, USA.

[11] Vaclav Snasel, Jan Platos, Pavel Krömer. 2008. On Genetic Algorithms for Boolean Matrix Factorization. In *Eighth Intern. Conference on Intelligent Systems Design and Applications* (ISDA 2008), Vol. 2. IEEE Computer Society, 170–175. DOI: http://dx.doi.org/10.1109/ISDA.2008.317.

[12] Masoumeh Rezaei and Reza Boostani. 2014. Using the genetic algorithm to enhance nonnegative matrix factorization initialization. *Expert Sys: J. Knowl. Eng.* 31, 3 (2014), 213-219. DOI: http://dx.doi.org/10.1111/exsy.12031.

[13] Sasalak Tongkaw and Aumnat Tongkaw. 2018. Multi-Vlan Design Over IPSec VPN for Campus Network. In *Proceedings of the 2018 IEEE Conference on Wireless Sensors* (ICWiSe). IEEE, Malaysia, 66–71. DOI: http://doi.org/10.1109/ICWISE.2018.8633293.

[14] Cheng-Feng Tai, Tzu-Chiang Chiang, and Ting-Wei Hou. 2011. A virtual subnet scheme on clustering algorithms for mobile ad hoc networks. *Expert Syst. Appl.* 38, 3 (2011), 2099–2109. DOI: http://dx.doi.org/10.1016/j.eswa. 2010. 07.148.

[15] Igor Saenko and Igor Kotenko. 2010. Genetic optimization of access control schemes in virtual local area networks. In *Proceedings of the 5th international conference on Mathematical methods, models and architectures for computer network security* (MMM-ACNS'10), Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, Berlin, Heidelberg, 209-216.

[16] Keqiang He, Yi Wang, Xiaofei Wang, Wei Meng, and Bin Liu. 2012. GreenVLAN: An energy-efficient approach for VLAN design. In *Proceedings of the 2012 International Conference on Computing, Networking and Communications* (ICNC). IEEE, New York, NY, 522–526. DOI: http://dx.doi.org/10.1109/ICCNC.2012.6167478.

[17] Xin Sun, Yu-Wei E. Sung, Sunil D. Krothapalli, and Sanjay G. Rao. 2010. A systematic approach for evolving VLAN designs. In *Proceedings of the 29th conference on Information communications* (INFOCOM'10). IEEE Press, Piscataway, NJ, USA, 1451–1459.

[18] Minli Zhu, Mart Molle, and Bala Brahmam. 2004. Design and Implementation of Application-Based Secure VLAN. In *Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks*. IEEE, Washington, DC, USA, 407–408. DOI: https://doi.org/ 10.1109/LCN.2004.42.

[19] Haibing Lu, Jaideep Vaidya, and Vijayalakshmi Atluri. 2008. Optimal Boolean Matrix Decomposition: Application to Role Engineering. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering* (ICDE '08). IEEE Computer Society, Washington, DC, USA, 297-306. DOI: https://doi.org/10.1109/ICDE.2008.4497438.