Multicriteria Optimization of Virtual Machine Placement in Cloud Data Centers

Andrew Toutov, Natalia Toutova
Moscow Technical University of Communication and Informatics
Moscow, Russia
andrew_vidnoe@mail.ru, e-natasha@mail.ru

Anatoly Vorozhtsov, Ilya Andreev
Moscow Technical University of Communication and Informatics
Moscow, Russia
as.vorojcov@mail.ru, andi@1c.ru

Abstract—**The problem of virtual machine placement on physical servers in cloud data centers is considered. The resource management system has a two-level architecture consisting of global and local controllers. Local controllers analyze the state of physical servers on which they are located and determine possible underloading, overloading, and overheating states based on the forecast for the next observation window. The global controller gathers the information from local controllers and start the process of destination server selecting and virtual machines migrating. In this paper we propose to place virtual machines based on the criteria of minimum resource wastage and SLA-violation. The mathematical formulation of the optimization problem is given, which is equivalent to the known main assignment problem in terms of structure, necessary conditions, and the nature of variables. Reducing the assignment problem to a closed transport problem allowed us to effectively solve the problem of multicriteria virtual machine placement in real time. We could significantly increase its dimension compared to heuristic algorithms, which makes it possible to maintain the quality of cloud services in conditions of rapid resource demand growth of data centers. The developed mathematical formulation of the problem and the results of computational experiments can be included in the mathematical software of virtual machine live migration.**

I. Introduction

There are increasing demands on the data center resources due to the growth of Internet traffic, the emergence of "big data", the development and spread of cloud services and artificial intelligence systems. Data centers must provide sufficient resources to hosted applications, which workload can vary significantly over time.

In order to avoid performance degradation, dynamic reallocation of resources is used. In cloud data centers, resource allocation is performed by moving virtual machines (VM) between physical servers. This process is called virtual machine migration. If the migration occurs without interruption of the VM, then this migration is called "live". Such migration allows data centers to guarantee service level agreements (SLAs), balance the load between physical machines (PM), and host VMs on fewer PMs to improve overall resource utilization and reduce resource wastage. Servers released by this process can be turned to lower power states (such as suspended or turned off) with the goal of minimizing the overall power consumption.

In addition, information resources can be redistributed between different data centers in accordance with the resource demand, for example, due to time zones. This makes us look for efficient and fast algorithms for resource allocation, taking into account the growing dimensions of problems.

The process of dynamic resource allocation includes three stages: monitoring servers for detecting critical situations, VM selection for migration and destination server selection [1-3].

This paper focuses on the third stage: destination server selection for hosting virtual machines. The problem of multi-criteria optimization is set and the method to find the solution is chosen.

The main contributions of this paper are the following.

1) Formal definition of multi-criteria optimization of virtual machine placement in the form of assignment problem.

2) Reducing the assignment problem to a closed transport problem allowed to find exact solution of the VM placement problem in real time and significantly increase its dimension.

3) Competitive analysis of proposed solution with the First Fit Decreasing (FFD), Best Fit Decreasing (BFD) heuristics based on simulation evaluation.

The remainder of the paper is organized as follows. In Section 2, we highlight several relevant research works. In Section 3, the typical cloud data center resource management system is discussed. Section 4 provides the system overview, assumptions, and problem description as well as mathematical statement of the VM placement problem. The proposed method to solve the VM placement problem is presented in Section 5. Section 6 reflects the simulation environment and the performance evaluation. Finally, Section 7 concludes the article.

II. Related works

Nowadays, IaaS providers mostly rely on either static VM provisioning policies, which allocate a fixed set of physical resources to VMs using bin-packing algorithms, or dynamic policies, capable of handling load variations through live VM migrations and other load balancing techniques. These policies can either be reactive or proactive, and typically rely on knowledge of VM resource requirements, either user-supplied or estimated using monitoring data and forecasting [4]. These approaches can be applied together and included in the main work cycle on resource management in cloud data center [5].

Most of the works that considered static VM placement use bin-packing or knapsack statements of VM placement problem [6-9]. These problems belong to the class of NP-hard problems. Therefore, in practice greedy heuristics such as FFD, BFD and their modifications are widely used for destination server selection to place migrating VMs [2, 10, 11]. However, as VM consolidation is a NP-hard problem, greedy approaches are not guaranteed to generate near optimal solutions.

Recently, Ant Colony Optimization (ACO) metaheuristics and genetic algorithms have been used to address bin packing problem and VM consolidation [1, 9, 12, 13]. However, such approaches are not guaranteed to give optimal solutions. Furthermore, some works [12, 13] take into account only one-dimensional resource.

The time required to solve the optimization problem is one of the major factors that affect the quality of real-time decision-making. One cycle of operation of the controller lasts a few minutes. In work [2], a 5-minute cycle was used for simulation. In [1], a 2-minute cycle was used to monitor processor load, and a 6-minute cycle was used to detect low energy efficiency. In work [14], the problem of choosing the optimal window size to ensure the stability of the migration process was considered. It is based on the estimation of migration duration given in the works [15,16]. During this time, the controllers must detect a problem on the servers (overload, underload, or overheating), select the VMs for migration and servers for hosting the VMs, and migrate them. Delays in decision-making can lead to significant penalties for violating SLA agreements and additional operating costs. An unregulated increase in delays will make it impossible to implement innovative high-yield cloud data center services.

Some dynamic VM consolidation approaches [17,18] try to pack VMs into a minimal number of servers while reducing the number of migrations. In this paper, we do not discuss the issue of initiating migration, but suggest that it is necessary. The question is which physical servers for hosting the VMs to choose.

Multicriteria placement problems were considered in [1, 3, 6, 9, 13, 19]. To solve such multi-criteria problems, methods of forming a generalized criterion are most often used [1,2,6,9]. In [19], the method of hierarchy analysis was used to select the server, which is not appropriate in real time. A detailed overview of existing VM placement algorithms can be found in [20].

However, most of the above works do not guarantee an exact solution and have not fully considered the issues of dimension and solving time. Therefore, it is advisable to formulate the problem that would allow to scale resources over a wide range and make it possible to get an exact solution in real time.

III. SYSTEM ARCHITECTURE

Dynamic allocation of computing resources with live migration of virtual machines is the main stage of the cycle of the cloud data center management system, the block diagram of which is shown in Fig. 1.

A typical cloud data center resource management system has a two-level architecture consisting of global and local controllers.

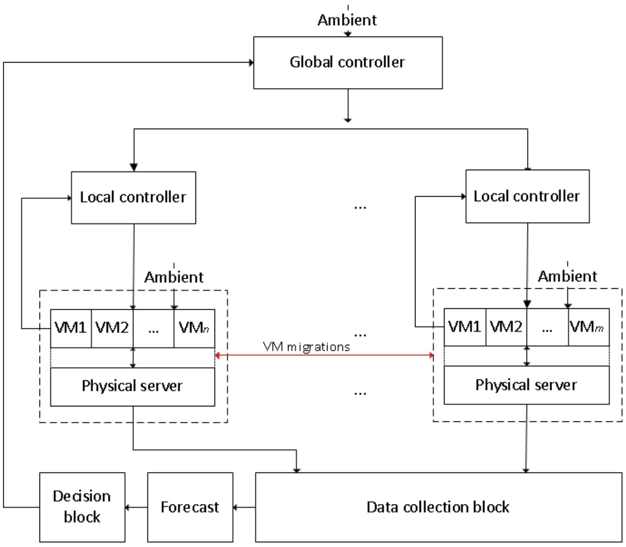

Fig. 1. System architecture

Local controllers constantly analyze data derived from the monitoring system about the state of the physical servers on which they are located. Possible states of underload, overload, and overheating are determined based on the forecast for the next observation window. System indicators are checked sequentially in accordance with the importance of the criteria, while the monitoring process is carried out continuously, including the VM migrations (Fig. 2).

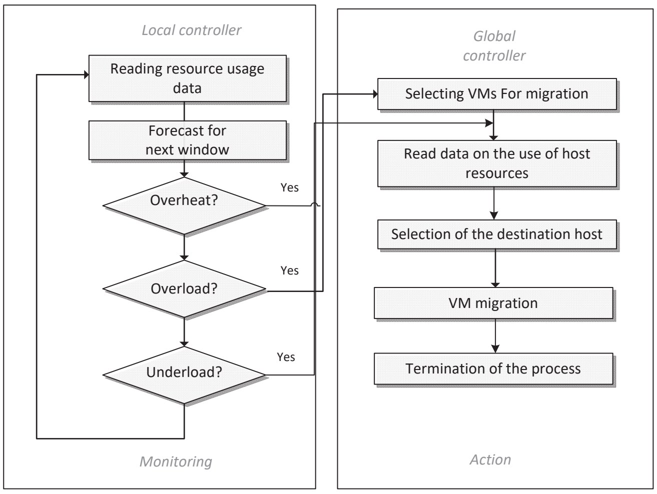
Local controller
Reading resource usage data
Forecast for next window
Overheat?
Yes
Overload?
Yes
Underload?
Yes
Monitoring
Global controller
Selecting VMs For migration
Read data on the use of host resources
Selection of the destination host
VM migration
Termination of the process
Action

Fig. 2. Controllers' algorithms

If one of these states is detected, the local controller sends a message to the global controller that initiates the VM migration process: the VMs to migrate and the destination servers are selected. Destination server selection is a critical step, since an unsuccessful server selection can lead to new undesirable migrations, since the migrations themselves additionally load to the system which leads to performance degradation and downtime of virtual machines.

IV. THE PROBLEM STATEMENT

The process of virtual machine hosting on physical servers is described by a number of indicators: resource efficiency [1, 3, 9], power consumption [1, 3, 9, 11, 21, 22, 17, 23], uniform temperature distribution [1, 3], SLA-violations [1,3,5], load balancing [10], traffic minimization [24, 25], and others. In [3] a combined metric is proposed that captures both the level of SLA violations and energy consumption.

In this paper, we consider only one of the stages of the data center resource management cycle, namely, the selection of destination servers for hosting virtual machines. It is assumed that the VMs have already been selected for migration. Therefore, such a criterion as energy consumption is not suitable for this task. The most appropriate criteria are uniform resource utilization and the level of SLA violations.

Suppose that there are *N* active PMs and the same number of VMs for migration. These PMs are running and can serve other VMs.

For each VM, the VM_i^{CPU} processor performance and VM_i^{RAM}, *i*=1,…, *N* RAM size are set. Each server has its own characteristics: the $PM_j^{CPU\_0}$ processor performance and RAM ($PM_j^{RAM\_0}$), and some of the resources are already occupied by other VMs. The occupied part of the processor is denoted as $PM_j^{CPU\_1}$, and the occupied part of the memory is denoted as $PM_j^{RAM\_1}$. Let's assume that the PM has enough physical resources to host any VM.

The global controller must define the PMs to which the VMs will be moved. Since placement occurs dynamically over a single controller cycle, only one VM can be placed on each individual server, just as each VM can be placed on only one server (Fig. 3).

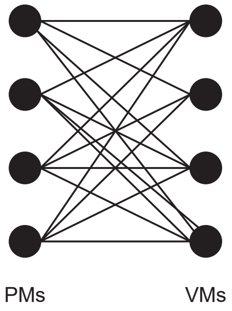

Fig. 3. Illustration of VM placement problem

The problem is to find the optimal placement of virtual machines, taking into account the minimum decision-making time for the placement process, which does not exceed the controller's operating cycle of 2-6 minutes.

Let's assume that a VM hosted on a server takes up all of its allocated memory and CPU time. Then we denote u_{ij}^{CPU} the CPU load of server *j* after placing the VM *i*, and u_{ij}^{RAM} the memory load of server *j* after placing the VM *i*.

$$u_{ij}^{CPU} = \frac{PM_j^{CPU\_1} + VM_i^{CPU}}{PM_j^{CPU\_0}} \tag{1}$$

$$u_{ij}^{RAM} = \frac{PM_j^{RAM\_1} + VM_i^{RAM}}{PM_j^{RAM\_0}} \tag{2}$$

The resource wastage criterion reflects how unbalanced resources are used on each server. The remaining resources on each server must be balanced across several types of resources. For example, an unsuccessful placement of virtual machines is a placement where all the server memory is used and the processor is low-loaded.

The **resource wastage criterion** can be formulated as follows:

$$f_{res}^{j}(u_{ij}^{CPU}, u_{ij}^{RAM}) = 1 - u_{ij}^{CPU} \cdot u_{ij}^{RAM} \tag{3}$$

This criterion reflects how fully the resources of various types of servers are loaded. The values of this criterion range from 0 to 1. The closer the criterion value is to zero, the better the server resources are loaded.

Another criterion considered is **SLA-violation.** The quality of service requirement for a cloud service is set as the average response time, which depends mostly on the CPU utilization. If a certain utilization threshold is exceeded, the application performance degrades and the response time increases. In [23], the influence of the utilization threshold on SLA violations was investigated. Most works uses utilization threshold in the range of 80-90% [1, 10].

The following logistic function is selected as **SLA-violation criterion**:

$$f_{SLA}^{j}(u_{ij}^{CPU}) = 1 - \frac{1}{1 + e^{-(u_{ij}^{CPU} - 0,8)}} \tag{4}$$

The value of this function also belongs to the range from 0 to 1. At the threshold point $u_{ij}^{CPU} = 0.8$, the function value is 0.5 and increases rapidly when the threshold value is exceeded. This criterion should be minimized.

A. Mathematical statement of the problem

Let x_{ij} are task variables that will correspond to the VM's assignment to the PM. In this case, $x_{ij} = 1$ if the *i*-th VM is assigned to the *j*-th PM for execution, and $x_{ij} = 0$ if the *i*-th VM is not assigned to the *j*-th PM for execution.

The mathematical formulation of the VM placement problem is as follows:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} f_{res}^{j}(u_{ij}^{CPU}, u_{ij}^{RAM}) \cdot x_{ij} \to \min_{x \in \Delta_\beta},$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} f_{SLA}^{j}(u_{ij}^{CPU}) \cdot x_{ij} \to \min_{x\in\Delta_\beta},$$

where the set of acceptable alternatives Δ_β is formed by the following constraint system:

$$\begin{cases} \sum_{j=1}^{N} x_{ij} = 1, \forall i \in \{1,2,\dots,N\} \\ \sum_{i=1}^{N} x_{ij} = 1, \forall j \in \{1,2,\dots,N\} \\ x_{ij} \in \{0,1\}, \forall i,j \in \{1,2,\dots,N\} \end{cases}.$$

The formulated problem in structure, necessary conditions, and character of variables is equivalent to the well-known main assignment problem.

The classical statement of the assignment problem is formulated as follows. There are a finite number of types of work that can be performed by potential candidates. However, each candidate can be assigned to only one job, and each job, in turn, must be performed by only one candidate. The effectiveness of each job performed by any of the potential candidates is known. It is necessary to distribute all candidates by job so that the overall performance of all work is the highest. The evaluation function in this task is the overall performance of all work, and the restrictions are additional conditions for each job to be performed by only one candidate and for each candidate to participate in only one job.

V. Method of solution

The problem presented in this paper belongs to the class of multi-criteria problems, for which several approaches and methods of solving have been developed, the most common of which is the linear convolution method. It consists in assigning coefficients in a linear convolution of the initial criteria in one way or another and then finding its extremum on the set of acceptable alternatives. According to this method, the solution found in this way is considered the best [26]. In addition, this method is most suitable for solving problems with binary variables. Thus, the transition from a multi-criteria task to a single-criteria one is actually carried out.

A special Hungarian method was developed for solving the single-criteria assignment problem [27]. The originality of this method is based on the following property of the cost matrix. If an arbitrary constant number u_i is added to all elements of the c_{ij} of a certain i-th row, and an arbitrary constant number v_j is also added to all elements of the j-th column, then a new cost matrix with elements will be obtained: $d_{ij} = c_{ij} + u_i + v_j$. Thus, the idea of finding the optimal solution to the assignment problem using the Hungarian method is to move to an equivalent problem by modifying the objective function and the system of constraints: the smallest elements are sequentially subtracted from the elements of each row and each column of the original cost matrix. After that, obtaining a feasible solution is analyzed. If a feasible solution is obtained that corresponds to zeros in the modified cost matrix, then it is the optimal assignment. If the solution is invalid, then further modification of the cost matrix is performed. This approach made it possible to obtain solution in polynomial time: the complexity $O(n^4)$ [27] and $O(n^3)$ in [28]. However, this method is not intended for solving multicriteria and high-dimensional problems.

It is known [28, 29] that the assignment problem can be reduced to a closed transport problem by replacing the constraint with $x_{ij} \geq 0$. This makes it possible to use well-known and available software to get the optimal solution. Let's choose this method for solving the problem of placing virtual machines on physical servers.

As a result, the problem statement will look like this:

$$\alpha_1\sum_{i=1}^{N}\sum_{j=1}^{N} f_{res}^{j}(u_{ij}^{CPU}, u_{ij}^{RAM}) \cdot x_{ij} + \alpha_2\sum_{i=1}^{N}\sum_{j=1}^{N} f_{SLA}^{j}(u_{ij}^{CPU}) \cdot x_{ij} \to \min_{x\in\Delta_\beta}$$

under constraints

$$\begin{cases} \sum_{j=1}^{N} x_{ij} = 1, \forall i \in \{1,2,\dots,N\} \\ \sum_{i=1}^{N} x_{ij} = 1, \forall j \in \{1,2,\dots,N\} \\ x_{ij} \geq 0, \forall i,j \in \{1,2,\dots,N\} \end{cases},$$

where α_1, α_2 – criteria weights; $\alpha_1 + \alpha_2 = 1, \alpha_1, \alpha_2 \geq 0$.

This task is solved if the number of servers is equal to the number of virtual machines. In this case, it can be reduced to a closed transport problem. However, if the number of servers and VMs is not equal, we can use the technique that is used to reduce an open transport task to the closed one—by introducing dummy nodes. The disadvantage of this method is the possibility of an increase in computational costs for bringing the problem to a balanced form. However, reduction to a closed transport problem allows us to solve the problem of multicriteria and significantly expand the scalability of the problem.

VI. Performance evaluation

As the targeted system is an IaaS, a Cloud computing environment that is supposed to create a view of infinite computing resources to users, it is essential to evaluate the proposed resource allocational algorithms on a large-scale virtualized data center infrastructure. However, it is extremely difficult to conduct repeatable large-scale experiments on a real infrastructure, which is required to evaluate and compare the proposed algorithms. Therefore, to ensure the repeatability of experiments, simulations have been chosen as a way to evaluate the performance of the proposed method.

Two series of experiments were conducted to evaluate the proposed multi-criteria approach to VM placement in terms of placement efficiency and scalability. The resource settings of the virtual machines were generated randomly. The CPU performance of virtual machines in GHz is evenly distributed from the following set of values {0.25, 0.5, 1, 1.5, 2, 2.5, 3, 4} and a memory in GB is distributed from the same set of values.

The number of available PMs and VMS is set as the initial values for simulating data centers of various sizes.

Table I shows the parameter settings for two sets of experiments. Sets of random input data are generated for each series of experiments. Each experiment was performed 20 times. The results are averaged.

TABLE I. THE VALUES OF THE PARAMETERS OF THE EXPERIMENT

A series of experiments	The size of the data center
The effectiveness of allocation	50 PM, 50 VM
Scalability	50~250 VMs and PMs

In each set of experiments, the proposed multi-criteria algorithm was compared with the practical heuristic algorithms for the bin packing problem.

Best Fit Decreasing algorithm (BFD)—the list of servers is sorted in descending order according to the performance of the processor (bfd_cpu) or memory (bfd_mem), then each VM is assigned to the server so that the remaining resource used (processor or memory) is minimal.

First Fit Decreasing algorithm (FFD)—the list of servers is sorted in descending order according to processor performance (ffd_cpu) or memory (ffd_mem), then each VM is assigned to the first matching server in the list.

When solving the optimization problem using the convolution method, the values of the vector of weight coefficients changed in the range (0,1; 0,9) ~ (0,9; 0,1) in increments of 0.2.

The minimum and maximum values for each criterion were also calculated, and the normalized values of the criteria were calculated relative to them. Results on criteria for SLA-violation and resource wastage for the bfd_cpu, bfd_mem, ffd_cpu, ffd_mem, and α_1, α_2 = 0,1; 0,3; 0,5; 0,7; 0,9 shown in Fig. 4 and Fig. 5.

As it can be seen, in accordance with the bfd_cpu algorithm, virtual machines were placed more densely on the servers and, as a result, there is the highest level of violation of SLA agreements. Also, the algorithms bfd_cpu, bfd_mem, ffd_cpu, ffd_mem, being single-criterion, demonstrate inefficient loading of several types of resources (Fig. 5). In this regard, convolution is the preferred method, giving an efficient solution on two criteria. Based on the results of computational experiments, we can conclude that solving the problem of optimizing the placement of virtual machines as a transport problem allows us to find the optimal value according to two criteria in comparison with heuristic algorithms.

The second series of experiments was aimed at determining the run time depending on the dimension of the problem. The transport problem was solved by the simplex method implemented in the lpSolve package for the R language on a Pentium(R) Dual-Core CPU E5700 3 GHz 4 GB RAM. The solution time depending on the problem dimension is shown in Table II.

The dependence of the problem solution time on the dimension is quadratic $O(n^2)$ with the approximation confidence value R² = 0.9916 and is shown in Fig. 6, that is better than Hungarian algorithms. The computational complexity of the algorithm for solving this problem is optimal relative to the order of complexity [30].

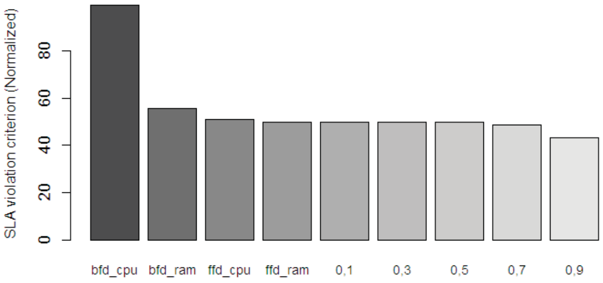

Fig. 4. Normalized values of the SLA violation criterion for various algorithms

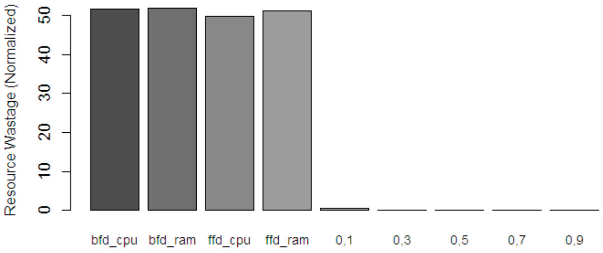

Fig. 5. The normalized values of the resource wastage criterion

TABLE II. THE DEPENDENCE OF THE SOLUTION TIME ON THE PROBLEM DIMENSION.

The size of the data center	The computation time (s.)
50 VMs, 50 PMs	0.73
100 VMs, 100 PMs	1.72
150 VMs, 150 PMs	4.84
200 VMs, 200 PMs	11.44
250 VMs, 250 PMs	22.17

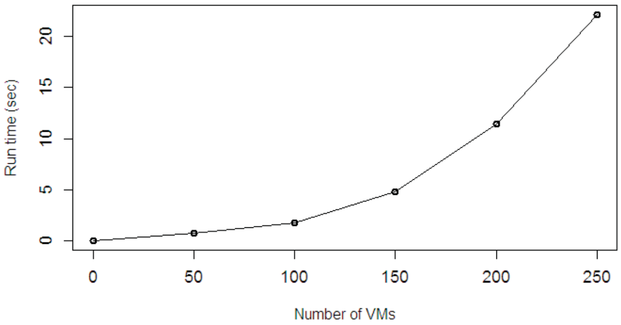

Fig. 6. Solution computation time for large problem instances

VII. CONCLUSIONS

To increase profits cloud providers have to apply resource management strategies, such as dynamic consolidation of VMs and switching idle servers to power-saving modes. However, such consolidation is not trivial, as it can result in violations of the SLA negotiated with customers. The process of dynamic resource allocation includes three stages: monitoring servers for detecting critical situations, VM selection for migration and destination server selection. This paper focuses on the third stage: destination server selection for hosting virtual machines. The problem of multi-criteria optimization of virtual machine placement is set in the form of assignment problem.

We have evaluated the proposed algorithms through simulations. Two series of experiments were conducted to evaluate the proposed multi-criteria approach to VM placement in terms of placement efficiency and scalability. It was shown that proposed method outperforms widely used FFD and BFD heuristics. Furthermore, the computational complexity of the algorithm is quadratic $O(n^2)$ that is better than Hungarian algorithm.

Reducing the main assignment problem to a closed transport problem made it possible to solve the problem of virtual machine placement under many criteria in real time and significantly increase its dimension, which makes it possible to maintain the quality of modern cloud services in the conditions of rapid growth of physical and virtual resources of data centers.

The developed mathematical statement of the problem and the results of computational experiments can be included in the mathematical support of data center resource management system.

REFERENCES

[1] J. Xu and J. Fortes, “A multi-objective approach to virtual machine management in datacenters”, *Proceedings of the 8th ACM international conference on Autonomic computing*, 2011, pp. 225-234.

[2] A. Beloglazov and R. Buyya, “Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers”, *Concurrency and Computation: Practice and Experience*, 2012, Vol. 24, №. 13, pp. 1397-1420.

[3] A.S. Vorozhtsov, N.V. Tutova, and A.V. Tutov, “Dynamic computing resource allocation in data centers”, *T-Comm*, Vol. 10, No.7, 2016, pp. 47-51. (in Russian).

[4] R. Buyya et al. “A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade”, *ACM Computing Surveys*, Volume 51, No. 5, Article No. 105, Pages: 1-38, ISSN 0360-0300, ACM Press, New York, USA, January 2019.

[5] A. V. Toutov “Models and methods of resource allocation of infocommunication system in cloud data centers”, *Aviation, space-rocket hardware*, 2018, Vol. 10, №. 6, p. 100.

[6] J. Xu and J. Fortes, “Multi-objective Virtual Machine Placement in Virtualized Data Center Environments”, *Proceedings of the 2010 IEEE/ACM Conference on Green Computing and Communications*, 179-188.

[7] A.S. Vorozhtsov, N.V. Tutova, and A.V. Tutov, “Optimal cloud servers placement in data centers”, *T-Comm*, Vol 9, No.6, 2015, pp. 4-8. (in Russian).

[8] R. S. Camati, A. Calsavara, and Jr L. Lima, “Solving the virtual machine placement problem as a multiple multidimensional knapsack problem”, *ICN 2014*, 2014, pp. 253-260.

[9] M. H. Ferdaus et al., “Virtual machine consolidation in cloud data centers using ACO metaheuristic”, *European conference on parallel processing*, Springer, Cham, 2014, pp. 306-317.

[10] A. Gulati et al., “Vmware distributed resource management: Design, implementation, and lessons learned”, *VMware Technical Journal*, 2012, Vol. 1, №. 1, pp. 45-64.

[11] F. F. Moges and S. L Abebe, “Energy-aware VM placement algorithms for the OpenStack Neat consolidation framework”, *Journal of Cloud Computing*, 2019, Vol. 8 (1). – P. 2.

[12] E. Feller, L. Rilling and C.Morin, “Energy-aware ant colony based workload placement in clouds”, *In Proceedings of the 2011 IEEE/ACM 12th International Conference on Grid Computing*, IEEE Computer Society, 2011, pp. 26–33.

[13] Y. Gao, H. Guan, Z. Qi, Y. Hou and L.Liu, “A multi-objective ant colony system algorithm for virtual machine placement in cloud computing”, *Journal of Computer and System Sciences*, 2013, Vol. *79,* (8), 1230-1242.

[14] A. S. Vorozhtsov, N. V. Toutova and A. V. Toutov, “Resource control system stability of mobile data centers”, *2018 Systems of Signals Generating and Processing in the Field of on Board Communications*, Moscow, 2018, pp. 1-4, doi: 10.1109/SOSG.2018.8350579.

[15] A. V. Toutov, A. S. Vorozhtsov and N. V. Toutova, “Analytical approach to estimating total migration time of virtual machines with various applications”, *International Journal of Embedded and Real-Time Communication Systems (IJERTCS),* Vol. 11, №. 2, 2020, pp. 58-75.

[16] A. V. Toutov, A. S. Vorozhtsov and N. V. Toutova, “Estimation of Total Migration Time of Virtual Machines in Cloud Data Centers", *2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies" (IT&QM&IS)*, St. Petersburg, 2018, pp. 389-393.

[17] A. Murtazaev and S. Oh, “Sercon: Server consolidation algorithm using live migration of virtual machines for green computing”, *IETE Technical Review*, 2011, Vol. 28, №. 3, pp. 212-231.

[18] E. Feller, C. Morin and Armel Esnault, “A case for fully decentralized dynamic vm consolidation in clouds”, *In Proceedings of the 4th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2012)*, IEEE, 2012, pp, 26–33.

[19] A. A. Mikryukov and R. Hantimirov, “Initial resource provisioning in IaaS clouds based on the analytic hierarchy process”, *Statistics and Economics*, 2015;(4), pp. 184-187. (In Russ.).

[20] M.C.S. Filho, C.C. Monteiro, P.R.M. Inacio, M.M. Freire, “Approaches for optimizing virtual machine placement and migration in cloud environments: A survey”, *J. Parallel Distrib. Comput*, 2017, Vol. 111, pp. 222-250.

[21] R. Shaw, E. Howley, and E. Barrett, “An energy efficient anti-correlated virtual machine placement algorithm using resource usage predictions”, *Simulation Modelling Practice and Theory*, 2019, Vol. 93, pp. 322-342.

[22] F. Alharbi et al., “An ant colony system for energy-efficient dynamic virtual machine placement in data centers”, *Expert Systems with Applications*, 2019, Vol. 120, pp. 228-238.

[23] A. Beloglazov, J. Abawajy, and R. Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing”, *Future generation computer systems*, 2012, Vol. 28, №. 5, pp. 755-768.

[24] S. Farzai, M. H. Shirvani, and M. Rabbani, “Multi-objective communication-aware optimization for virtual machine placement in cloud datacenters”, *Sustainable Computing: Informatics and Systems*, 2020, p. 100374.

[25] M. H. Ferdaus, M. Murshed, R. N. Calheiros and R. Buyya. “An algorithm for network and data-aware placement of multi-tier applications in cloud data centers”, *Journal of Network and Computer Applications*, 2017, Vol. 98, pp. 65-83.

[26] V. D. Nogin, “Linear convolution of criteria in multi-criteria optimization”, *Artificial intelligence and decision making*, 2014, №. 4, pp. 73-82. (in Russian).

[27] Kuhn H. W. “The Hungarian method for the assignment problem” *Naval research logistics quarterly*, 1955, Vol. 2, № 1-2, pp. 83-97.

[28] J. Munkres “Algorithms for the assignment and transportation problems”, *Journal of the society for industrial and applied mathematics*, 1957, Vol. 5, № 1, pp. 32-38.

[29] V. F. Krotov (ed.), *Fundamentals of optimal control theory: textbook*, Moscow: Higher school, 1990, 431 p. (in Russian).

[30] V. I. Khokhlyuk, *Parallel algorithms for integer optimization*, Moscow: Radio and communications, 1987. (in Russian).