Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

UDC 004.75
DOI: https://doi.org/10.30837/ITSSI.2020.13.025

V. DAVYDOV, D. HREBENIUK

DEVELOPMENT OF THE METHODS FOR RESOURCE REALLOCATION IN CLOUD COMPUTING SYSTEMS

The **subject** matter of the article is development of the models and methods of load and resource balancing and reallocation in cloud computing systems based on the infrastructure as a service model. The **goal** of the work is to increase the efficiency of available resources usage in cloud computing systems (such as RAM, disk space, CPU, network) by developing the model for adaptive management of resource reallocation. This will allow new virtual machines to be launched with minimal performance degradation for already running applications. The following **tasks** were solved in the article: development of an complex approach to manage resource reallocation in cloud systems, including decomposition of the cloud computing system into zones (based on the defining features of the resources provided in each zone), initial resource allocation (based on the hierarchy analysis method) and resources reallocation within cloud computing system (based on the developed method); development of a method for computing resources reallocation in cloud computing systems; evaluation of the effectiveness of the developed method. To solve the set tasks, approaches and **methods** of dynamic load balancing were used, as well as methods of theoretical research, which are based on the scientific provisions of the theory of artificial intelligence, static, functional and system analysis. The following **results** were obtained – on the basis of existing load balancing methods in cloud computing systems analysis, the main features of existing resource allocation methods were identified, their advantages and disadvantages were given. On the basis of the conducted analytical study, the necessity of improving the existing methods of resource reallocation has been proved. A method and an algorithm for computing resources reallocation within cloud computing systems have been developed. This make it possible to reduce the values of the coefficient of computing resources uneven usage while minimizing the cost of moving them. The results obtained have been confirmed by experiments carried out using software for creating private infrastructure cloud services and cloud storages. **Conclusions**: the improvement of the reallocation and load balancing method in cloud computing systems has increased the ability of these systems to launch new virtual machines with a minimum decrease in the performance of already running applications.

Keywords: infrastructure as a service; cloud computing; resource reallocation.

Introduction

The development of the information society adds qualitative adjustments to the perception of users about exactly how information services should be provided. Of paramount importance are qualitative indicators – timely satisfaction of business needs, ease of use, speed of standard operations, etc.

In connection with the above, the defining approach in the provision of computing power is the model "Infrastructure as a Service" (IaaS, Infrastructure-as-a-Service) [1-3], which makes it possible to minimize the interaction of the provider with the consumer of computing resources on technical issues, to reduce the number of incidents and the time of their processing, to provide businesses with greater opportunities to adapt their work to their needs, as well as reduce financial and operating costs.

IaaS is provided as an opportunity to use the cloud infrastructure to independently manage processing and storage resources, networks and other fundamental computing resources. For example, a consumer may install and run any software, which may include operating systems, platform and application software [2].

The cloud computing system [3-5] includes a wide range of computing resources: servers, storage systems, network devices. Resources can be of the same type or heterogeneous in performance, instruction set, ratio of the number of processor cores to the amount of RAM, etc. This exacerbates the problem of sub-optimal use of available resources and leads to a significant increase in the cost and complexity of management or the complete lack of ability to use cloud computing systems IaaS to solve various problems on one set of equipment [6, 7].

Thus, the urgency of the problem of optimal resource management in cloud computing systems to increase the efficiency of their use and minimize costs. At the same time both real needs of applications, and indicators of use of available resources taking into account their specificity have to be considered.

Analysis of recent research and publications

Existing approaches to primary allocation and subsequent resource reallocation in cloud computing systems can be divided into the following three types: manual resource assignment, resource scheduling by the cloud computing manager, and resource scheduling in a virtualization environment [5, 8].

The main ideas of IaaS are described in the literature [1, 2], namely:

1. Lack of information on the administrator about the real needs of applications that are inside the instances (most often - virtual machines).

2. The administrator does not have the ability to assign resources to instances manually.

These features of cloud computing systems determine the specifics of resource reallocation in such an environment. Some elements of this issue are common to the problems of virtualization systems, and some have features that are characteristic of cloud computing systems. The process of resource reallocation in such an environment is influenced by the following:

- different classes of equipment and combination of resources;
- uneven load of each of the resources;
- lack of information on the real needs of applications in resources;

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

- different resource needs of applications;
- discrepancies between the resources requested and the resources consumed;
- lack of ability to assign resources manually.

The influence of the above factors leads to irrational use of resources and inefficient operation of applications in the cloud, which, in turn, leads to reduced productivity.

Thus, the analysis showed that cloud computing systems have a number of characteristics and problems that are not inherent in other models of resource reallocation. This necessitates the improvement of resource reallocation methods for these environments.

Analysis of the literature [9, 10] revealed the features of the method of distributed resource reallocation (DRS). The DRS method solves the problem of distributing virtual machine workloads on nodes within a virtualization cluster, and tracks available resources. In addition, depending on the level of automation, DRS provides maximum performance by automatically transferring virtual machines to other nodes within the cluster. As a result of the DRS method, it is possible to optimally distribute the load between the virtual environment hypervisors - the least loaded virtual machines are consolidated on some hypervisors, and the most loaded - on others. First, it increases the peak performance of virtual machines, and second, it improves resource efficiency in the virtualization environment. However, this method only estimates the CPU load and does not predict the change in load on other parameters (for example, the number of I/O operations and network congestion). That is why the DRS method is quite primitive and is not suitable for use in cloud computing systems, where the administrator of the cloud computing system is not able to predict all the undesirable consequences of the redistribution of resources in this way.

Analysis of the literature [9, 10] revealed the features of the method of dynamic resource reallocation (DPM). The DPM method optimizes power consumption at the cluster or node level. When the DPM method is initialized, the resources of the node or cluster are compared to the needs of the virtual machine, including preliminary needs statistics, according to which the virtualization nodes are put into standby mode. As resource requirements increase, the DPM method starts free nodes and connects additional workloads to them. The study of this method showed that it solves the problem of reducing energy consumption in the virtualization environment, but does not solve the problem of maximizing the efficient use of available resources.

The analysis [8, 11-13] showed that in cloud computing systems based on the IaaS model, the problems of optimal resource planning arise both at the stage of their reallocation and in the process of use. These problems cannot be solved manually, as in the case of virtualization, for two reasons. First, the cloud infrastructure administrator does not know what needs applications have to use them effectively. Second, the presence of "static" platform - the lack of response to changes in these needs over time. That is, the IaaS model does not solve the problem of optimal redistribution of computing resources in cloud systems.

For example, one of the most popular cloud infrastructure management solutions currently OpenStack supports only 3 methods of resource reallocation: random, random within the availability zone and simple (resource reallocation occurs in turn) [8, 14, 15]. As for the redistribution of load with its unevenness – such approaches in cloud computing systems do not yet exist.

Selection of previously unsolved parts of the overall problem. The purpose of the work

The analysis of existing methods of resource reallocation in cloud computing systems showed that the currently used methods do not meet the real needs of owners and consumers of IaaS cloud service on a number of parameters and do not allow optimal use of available resources for heterogeneous applications existing within instances.

The aim of the work is to increase the efficiency of the use of computing resources in cloud computing systems based on the IaaS model by developing a method of intelligent control of their distribution.

To achieve this goal, the following tasks are defined:
- development of an integrated approach to intelligent management of resource reallocation in the cloud;
- development of a method for the reallocation of computing resources in cloud systems;
- evaluation of the effectiveness of the developed methods.

Materials and methods

To form an intelligent management of the reallocation of resources in the cloud, it is proposed to use an integrated approach, which includes the following three stages:

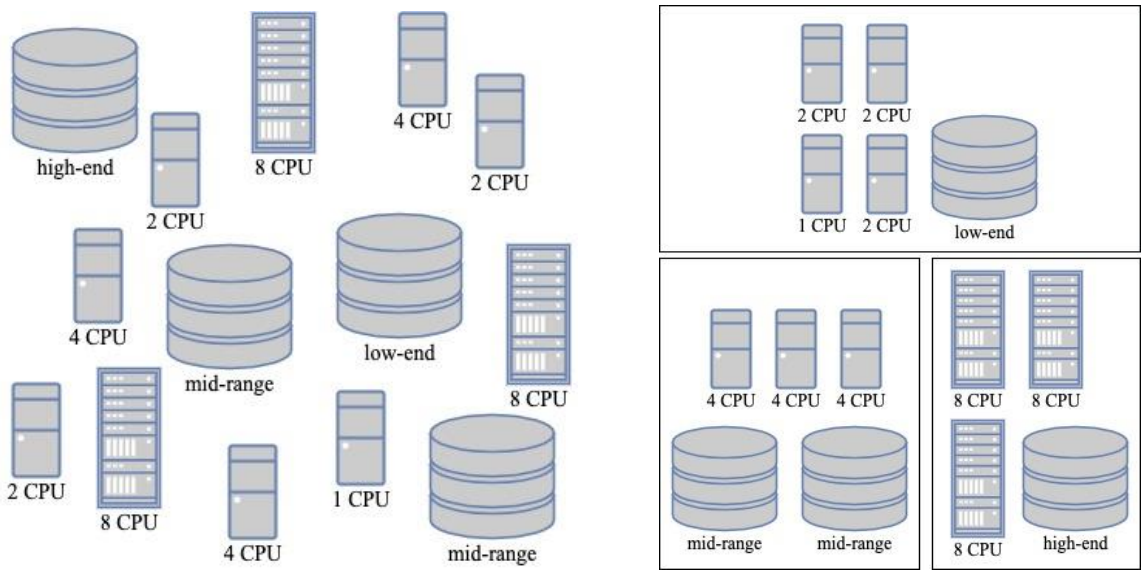
1. Decomposition of the cloud computing system into zones.
2. Initial allocation of resources.
3. Redistribution of resources of the cloud computing system.

Stage 1. At this stage, the system of cloud computing is decomposed into zones, based on the outstanding features provided in each zone of resources (fig. 1). Each of the zones includes hosts and storage resources with similar characteristics. For example, a cloud computing system can be divided into:
- high load area (eight-processor servers, 40-gigabit network, Hi-End storage system);
- medium load area (four-processor servers, 10-gigabit network, Midrange storage);
- low load area (dual-processor servers, gigabit network, local drives).

This simplifies the task as follows: during initialization, the instance will be placed in the area whose resources it is most likely to need, based on the initial characteristics (number of CPU cores, RAM, storage capacity). The choice is made using simple logical

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13) ISSN 2524-2296 (online)

expressions. The administrator of the cloud computing system performs decomposition based on his expert knowledge of the features of the equipment and its compatibility.



a) *b)*

Fig. 1. Decomposition of the cloud computing system into zones. Cloud system before decomposition into zones (a) and after decomposition into zones (b)

Stage 2. At this stage, the problem of primary allocation of resources is solved (fig. 2), i.e. there is a primary analysis of the possible needs of the application and launch the instance using those resources that are most likely to meet the needs of the application. The initial allocation of resources for each initialized instance occurs using the primary allocation model based on the method of hierarchy analysis. Prediction of values of dynamic parameters of functioning of system of cloud calculations is carried out on the basis of model of reception of forecast values with use of the mathematical device of Elman's neural networks, artificial immune systems and clustering by a method of fuzzy c-means. Redistribution of resources in the system of cloud computing is carried out on the basis of the model of dynamic redistribution of resources using the algorithm to reduce the uneven use of resources.

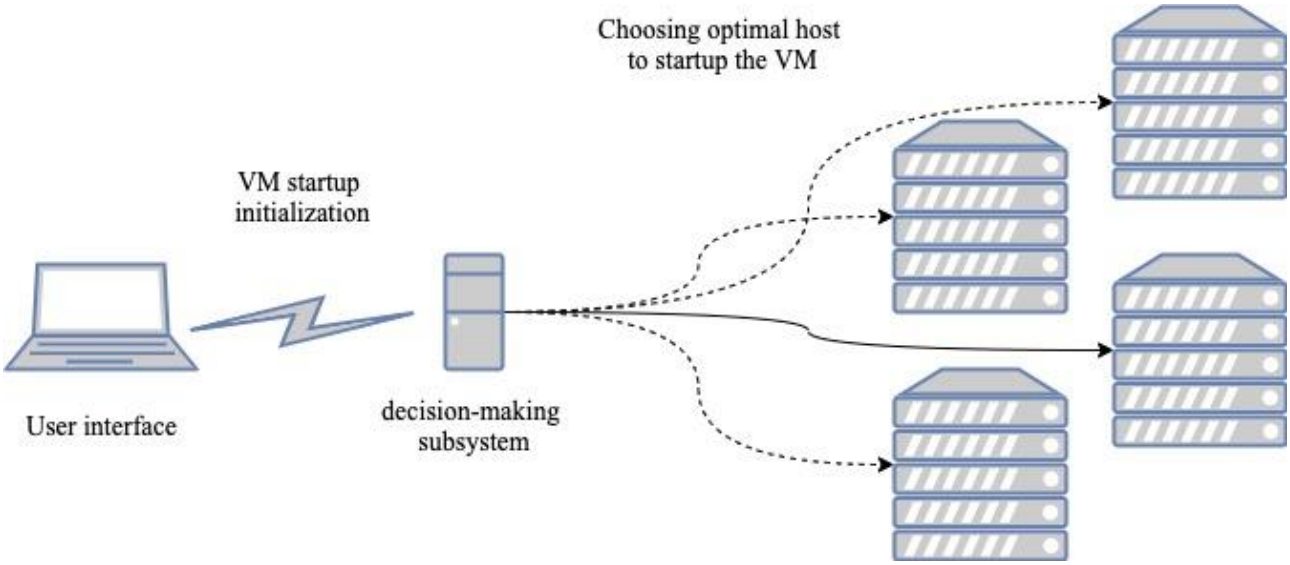


Fig. 2. The process of initial resource reallocation for each initialized virtual machine

Stage 3. At this stage, the problem of redistribution of computing resources in the system of cloud computing is solved, which is divided into the following two subtasks:

- determining the feasibility of redistribution of resources, i.e. the identification of overloaded resources, or resources for which excessive load is projected; analysis of expediency of initialization of dynamic redistribution at the moment; collection and analysis of metrics of sources of the provided computing resource. The decision on the reallocation of resources is made based on the results of the analysis;

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

- dynamic redistribution of resources (fig. 3). As a method of load balancing in the system of cloud computing, it is proposed to use the approach based on "live migration", which means the technology of migration of virtual machines between hosts or storage resources with zero downtime.

The basic sequence of the process of live migration in the event of a change of host is as follows:

1) suspension of the virtual machine;

2) transfer of parameters of the virtual machine from the current server to the target;

3) transfer the image of RAM from the current server location of the virtual machine to the target server location;

4) creating a virtual domain and placing the image of the RAM of the virtual machine in the RAM of the target location server;

5) initialization of the virtual machine on the target location server.

If you change the repository, the virtual machine image is transferred from one storage resource to another.

For efficient allocation of resources, it is proposed to use the concept of uneven use of resources of the *Rp* -th server (N_R^p). Let *n* is the number of considered resources, and – the projected load of the -th resource. Let's define the predicted uneven loading of server resources *p* as:

$$N_R = \sqrt{\sum_{i=1}^{n}\left(\frac{r_i - \bar{r}}{\bar{r}}\right)^2}, \quad (1)$$

where $\bar{r}$ is the average projected load of all server *p* resources. In practice, not for all types of resources, performance degradation affects the efficiency of the system as a whole, so only significant resources need to be considered in the calculation (e.g. CPU load, memory, network disk subsystem, etc.). By minimizing unevenness, it is possible to combine different types of workloads and improve the overall use of server resources in cloud computing systems.

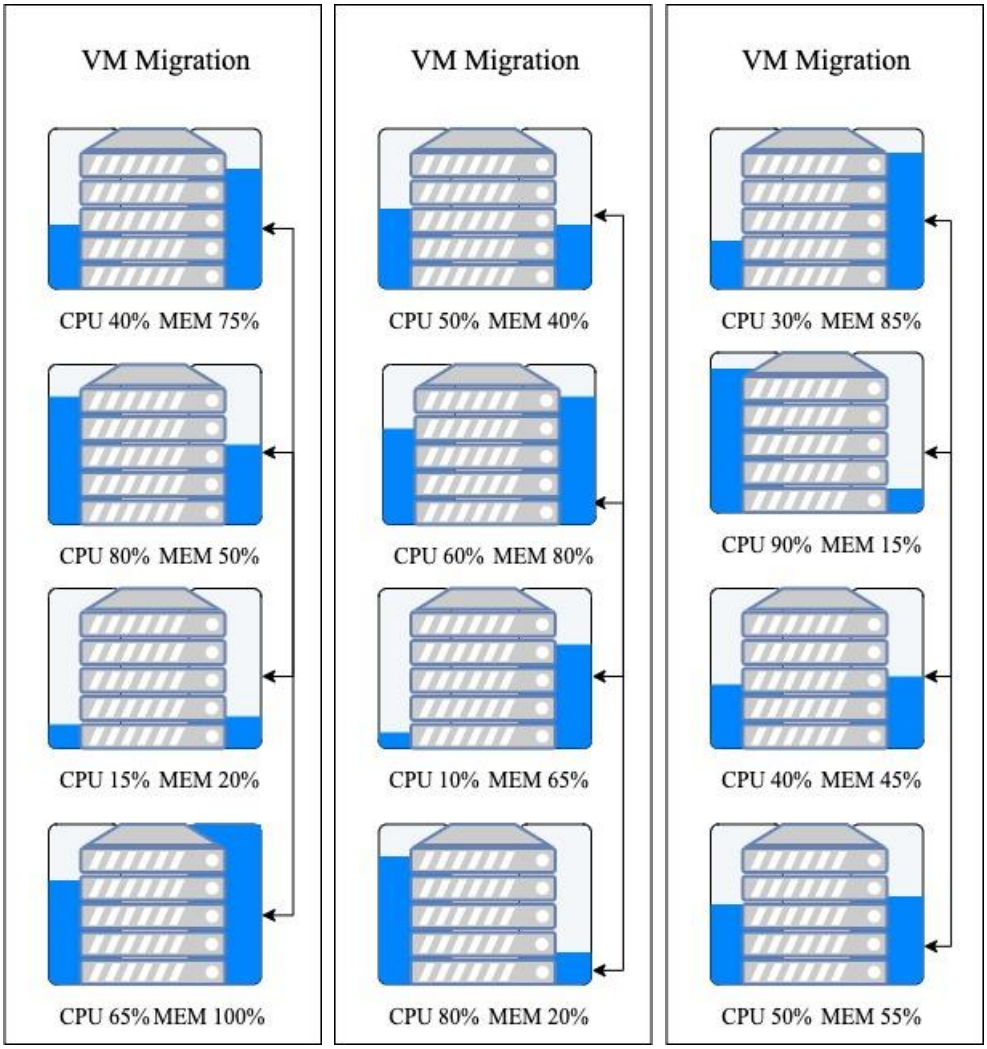


Fig. 3. Dynamic redistribution of resources for different types of host machines

The algorithm is performed periodically to assess the state of the allocated resources on the basis of the performed forecast of the needs of the instances. A server is called a "hot spot" if the use of any of its resources is above the so-called "valid mark" predefined for each type of resource. This indicates that the host is overloaded, so some instances must be moved from it to other hosts. The "temperature" $t*$ of a hot spot is defined as the square

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13) ISSN 2524-2296 (online)

sum of the use of all its resources above the "allowable mark":

$$t^* = \sum_{r \in R} \left(r - r_t\right)^2, \quad (2)$$

where R is a set of overloaded server resources p, and r_t is a "permissible mark" of the resource r (only overloaded resources are taken into account in the calculations). The hotspot temperature reflects the server overload level. If the server is not a hotspot, its temperature is zero.

Different types of resources have different valid labels. For example, for CPU usage and RAM usage, they can be defined as 90% and 80%, respectively. Thus, the server becomes a hot spot when this load is reached.

Given the above, the task of identifying servers with irrational use of computing resources is to identify "hot spots" according to the described algorithm.

The load balancing algorithm in the cloud computing system includes the following set of steps:

1. Sorting the list of servers - "hot spots" p in descending order of temperature t^* (i.e. the hottest point becomes the first in the list). The goal is to eliminate all hot spots if possible, or keep their temperature as low as possible.

2. Defining for each server p a list of instances e for which migration should be performed.

3. Sorting the list of instances e based on the resulting host temperature t^*_{res}, which is determined after the migration of the instance (virtual machine). The goal is to migrate the instance that will reduce the server p_n temperature to a minimum $t^*_{res}\left(p_n\right) \to \min$.

4. The choice of a virtual machine e', the migration of which will minimize the unevenness of the server.

5. Defining for the selected virtual machine e' the ability to find the target server $p_{t\arg}$ to host it. The server is considered possible for migration if after transferring the virtual machine e' the unevenness of the target server will be less than the unevenness on the current server.

6. Selecting a server $p'_{t\arg}$ from this list, the non-uniformity of which will be minimized after the transfer of this virtual machine. It should be noted that the unevenness of the target server can be increased. In this case, you must select a server for which such an increase will be minimal.

7. If the target server $p'_{t\arg}$ is found, the migration of the virtual machine to this server is initiated, and the load forecast for all available servers is updated. Otherwise, the next virtual machine is selected from the list, and the target server for it is searched.

Till the moment the target server can be found for any virtual machine of the "hot spot" server, the execution of the algorithm for this server continues. As soon as this becomes impossible, the transition to the next "hot spot" [16].

Load prediction using an artificial neural network before starting the algorithm allows you to ignore insignificant, short-term load peaks, which leads to a significant reduction in the number of false positives of the algorithm to eliminate hotspots and, thus, reduce computing costs for moving virtual machines.

Research results and their discussion

To study the developed algorithm of load redistribution in the system of cloud computing, a cluster consisting of 5 servers of similar configuration with 30 running instances was created. At the time of execution of the algorithm, the predicted parameters of the operation of this system of cloud computing are described by the dynamic values given in table 1.

Table 1. *Dynamic parameters of the cloud computing system*

Parameter	Server 1	Server 2	Server 3	Server 4	Server 5
Load					
CPU load L_{CPU} (%)	10	40	60	2	81
MEM load L_{MEM} (%)	10	99	50	50	30
Disk subsystem load L_{disk} (%)	1	30	60	70	90
Network load L_{net} (%)	10	20	51	5	30

In the graph shown in fig. 4, it is seen that the resources of servers 2, 4 and 5 in this example are used unevenly, which can lead to inefficient use of resources.

Let the "allowable mark" be set to 70% of the load of any of the host resources. Thus, we receive the list of "hot spots" servers, we calculate their "temperature" t^* and we sort these servers on descending temperature:

$$t^*\left(p_5\right) = \sum_{r \in R} \left(r - r_t\right)^2 = \left(L_{ЦП_5} - 0.7\right)^2 + \left(L_{диск_5} - 0.7\right)^2 =$$

$$= \left(0.81 - 0.7\right)^2 + \left(0.9 - 0.7\right)^2 = 0.0121 + 0.04 = 0.0521$$

$$t^*\left(p_2\right) = \sum_{r \in R} \left(r - r_t\right)^2 = \left(L_{ЦП_2} - 0.7\right)^2 = t^*\left(p_2\right) =$$

$$= \left(0.9 - 0.7\right)^2 = 0.04$$

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

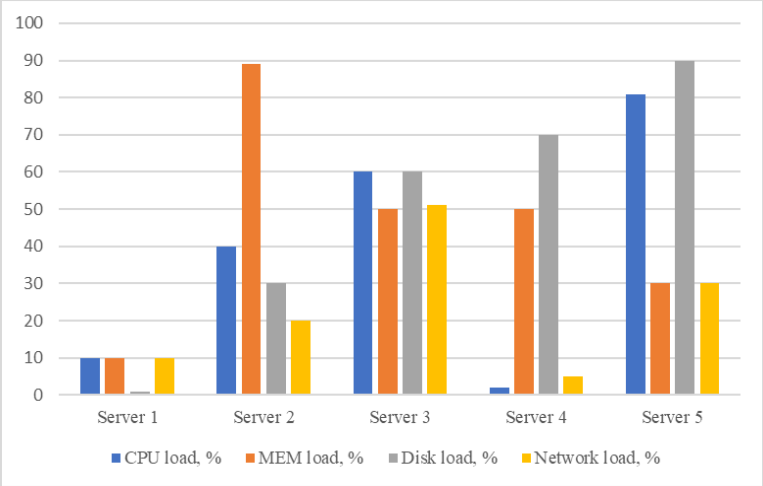


Fig. 4. Dynamic parameters of the cloud computing system

Therefore, the server 5 has the highest temperature (0.0521). Therefore, the next step will initiate the migration of one instance from this server to another cloud computing server. The list of instances of server 5 is presented in table 2, the share of server resource usage by each instance is clearly shown in fig. 5.

Table 2. *List of server instances 5*

Parameter	Instance 1	Instance 2	Instance 3
Using a host CPU L_{CPU} (%)	14	36	31
Using host MEM L_{MEM} (%)	10	15	5
Using the host disk subsystem L_{disk} (%)	43	30	17
Using a host network L_{net} (%)	8	12	10

In this case, the resulting server 5 temperature $t^*_{res}(p_5)$ is equal to 0 when migrating any instance from the server, because there will be no resource that exceeds the "hot mark". Therefore, the next step is to determine the unevenness of the server after the migration of each instance and select for the migration the instance that will minimize the unevenness of the server.

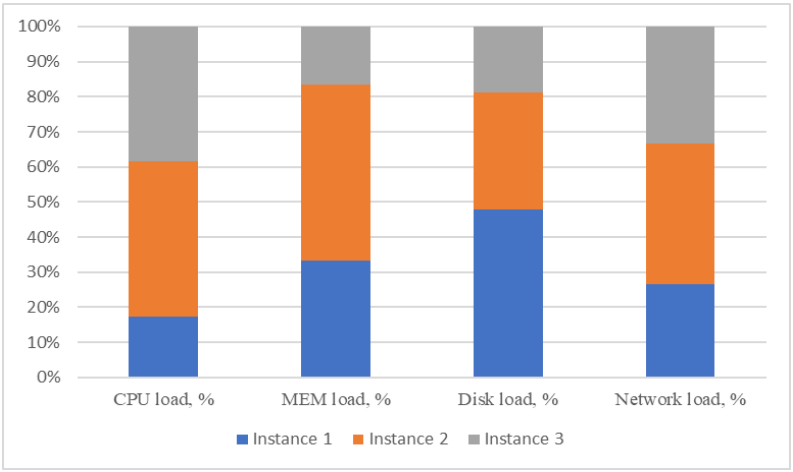


Fig. 5. Partitions of server 5 instance resource usage

Uneven server 5 after the migration of the first instance:

$$\bar{r}_1 = \frac{36+15+30+12+31+5+17+10}{8} = 19.5$$

$$N_R(p_5)_1 = \sqrt{\sum_{i=1}^{n}\left(\frac{r_i - \bar{r}}{\bar{r}}\right)^2} = \sqrt{\sum_{i=1}^{8}\left(\frac{r_i - 19.5}{19.5}\right)^2} = 1.54$$

Uneven server 5 after migrating the second instance:

$$\bar{r}_2 = \frac{14+10+43+8+31+5+17+10}{8} = 17.25$$

$$N_R(p_5)_2 = \sqrt{\sum_{i=1}^{n}\left(\frac{r_i - \bar{r}}{\bar{r}}\right)^2} = \sqrt{\sum_{i=1}^{8}\left(\frac{r_i - 17.25}{17.25}\right)^2} = 2.01$$

Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Uneven server 5 after the migration of the third instance:

$$\bar{r}_3 = \frac{36+15+30+12+14+10+43+8}{8} = 21$$

Based on the calculations, we conclude that the migration of the first instance will minimize the unevenness of the server 5, therefore, its migration must be initiated.

Next, a search is made for the server $p'_{t\arg}$ to which instance 1 will be migrated in the future. The server is selected on the basis of the maximum decrease in the unevenness after the acceptance of the instance (or the minimum increase in the unevenness of the server, if this increase is less than the change in the unevenness of the source server 5 after migration). In this case, only server 1 is suitable for migration, because the resources of other servers are not enough to accept the instance. The following indicators were calculated for him:

- output value N_R: 2.513;
- value of N_R after acceptance of the instance 1: 2.146;
- reduction of unevenness: 0.367.

As a result of the calculations for instance 1, migration to server 1 was initiated, which reduced the average uneven use of resources by hosts of the cloud computing system, and thus increase the efficiency of use of these resources.

The evaluation of the efficiency of the initial allocation of resources was performed by measuring the rate of uneven use of resources of the server p, which eventually launched the instance, using different methods. The parameter of uneven N_R loading of server p resources r is subject to comparison:

$$N_R = \sqrt{\sum_{i=1}^{n}\left(\frac{r_i - \bar{r}_t}{\bar{r}_t}\right)^2}, \qquad (3)$$

where r_i is a load on the current server resource p, $\bar{r}_t$ is an average current load of all server p resources. Each of the methods performed 100 test runs of an instance with arbitrary characteristics in a working cloud computing system without changing any load parameters. The evaluation results are presented in fig. 6.

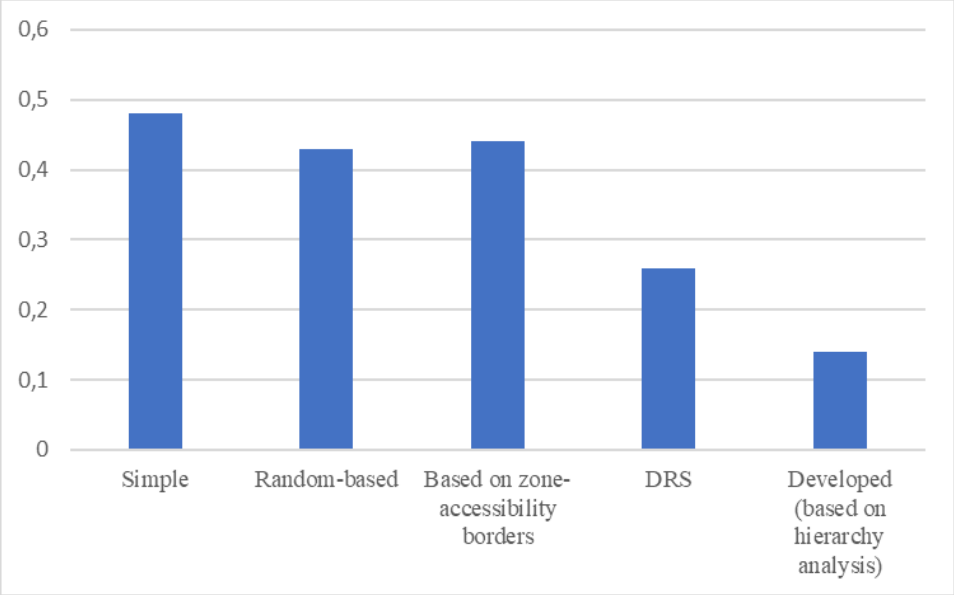


Fig. 6. Estimation of efficiency of various algorithms of primary allocation of resources in system of cloud calculations. On the OY axis is the value of the coefficient of uneven distribution (the smaller is the better)

As can be seen from the table above, the scatter of values between simple and random algorithms and the algorithm "within the reach" is small and is purely statistical. This is due to the fact that the reallocation of resources does not take into account any parameters of the cloud computing system [8, 15].

The DRS algorithm showed a much better result due to such a parameter as CPU load.

The developed algorithm based on the method of hierarchy analysis showed the best result of reducing the unevenness of host resources, as it takes into account the whole set of performance indicators of the host in relation to the characteristics of the running instance.

Thus, the developed resource reallocation algorithm has demonstrated its effectiveness in the initial launch of the instance in a functioning cloud computing system.

Conclusions and prospects for further development

1. A systematic analysis of existing methods of load distribution and balancing in corporate virtualization environments and cloud computing systems. The main differences in the features of resource reallocation are revealed, the necessity of using other approaches to resource reallocation of cloud computing systems instead of those used in virtualization is proved. Problems have been identified that hinder the optimal use of cloud

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative Technologies and Scientific Solutions for Industries. 2020. No. 3 (13)

resources in existing approaches to their planning and distribution.

2. The method and algorithm of distribution of computing resources in cloud systems which allow to allocate resources effectively with use of the forecast of loading with the minimum expenses of a computing resource on their movement are created.

3. Developed a method of planning and redistribution of resources, which provides a comprehensive consideration of indicators that affect the quality of service delivery and effective management of cloud computing systems.

As a result of the application of the created methods and algorithms of timely redistribution of resources in the dynamics of applications within instances it was possible to achieve a significant increase in cloud resource efficiency, as well as increase the ability of cloud computing systems to launch new instances with minimal performance. The obtained results can be applied as a resource reallocation subsystem for existing cloud computing systems.

The direction of further research is to further improve the methods of load distribution and balancing in virtualization environments, based on obtaining predictive values of dynamic parameters of cloud systems.

References

1. Dimitri, N. (2020), "Pricing cloud IaaS computing services", *Journal of Cloud Computing*, No. 9. DOI: https://doi.org/10.1186/s13677-020-00161-2
2. Soh, J., Copeland, M., Puca, A., Harris, M. (2020), "Overview of Azure Infrastructure as a Service (IaaS) Services", Mi*crosoft Azure*, P. 21–41. DOI: https://doi.org/10.1007/978-1-4842-5958-0_2
3. Kudriavtsev, A., Koshelev, V., Izbyshev, A., Dudina, I., Kurmangaleev, Sh., Avetisian, A., Ivannikov, V., Velihov, V., Riabinkin, Ye. (2013), "Design and Implement Cloud for High Performance", *Works ISP RAS*, No. 1, P. 13–33, available at : https://cyberleninka.ru/article/n/razrabotka-i-realizatsiya-oblachnoy-sistemy-dlya-resheniya-vysokoproizvoditelnyh-zadach (last accessed: 25.09.2020).
4. Vyshnivskyi, V., Vasylenko, V., Hrynkevych, H., Kuklov V. (2016), "Implement advanced cloud computing within data centers", *Information security*, No. 3 (23), P. 118–125.
5. Agavanakis, K., Karpetas, G., Taylor, M., Pappa, E., Michail, C., Filos, J., Trachana, V., Kontopoulou, L. (2019), "Practical machine learning based on cloud computing resources", *Technologies and Materials for Renewable Energy, Environment and Sustainability (TMREES19)*.
6. Alshamrani, S. (2018), "An Efficient Allocation of Cloud Computing Resources", *AICCC '18: Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference*, P. 68–75. DOI: https://doi.org/10.1145/3299819.3299828
7. Zhu, Y., Wang, Y. (2013), "A Model of Cloud Computing Resources", *Proceedings of the 2013 International Conference on Computer Sciences and Applications*, P. 684–686. DOI: https://doi.org/10.1109/CSA.2013.165
8. Srinivasan, J., Suresh Gnana Dhas, C. (2020), "Cloud management architecture to improve the resource allocation in cloud IAAS platform", *Journal of Ambient Intelligence and Humanized Computing*. DOI: https://doi.org/10.1007/s12652-020-02026-7
9. Hrebeniuk, D. (2018), "Analysis of methods of distribution of resources in the virtualization media", *Control, navigation and communication systems*, No. 6 (52), P. 98–103. DOI: https://doi.org/10.26906/SUNZ.2018.6.098
10. Gulati, A., Holler, A., Ji, M., Shanmuganathan, G., Waldspurger, C., Zhu, X. (2012), "VMware distributed resource management: Design, implementation and lessons learned", *VMware Technical Journal*, No. 1, P. 45–64.
11. Calcavecchia, N. M., Biran, O., Hadad, E., Moatti, Y. (2012), "VM Placement Strategies for Cloud Scenarios", *2012 IEEE Fifth International Conference on Cloud Computing*, P. 852–859. DOI: https://doi.org/10.1109/CLOUD.2012.113
12. Wu, G., Tang, M., Tian, Y., Li, W. (2012), "Energy-Efficient Virtual Machine Placement in Data Centers by Genetic Algorithm", *International Conference on Neural Information Processing*, P. 315–323. DOI: https://doi.org/10.1007/978-3-642-34487-9_39
13. Pasko, D., Molchanov, H., Davydov, V. (2018), "Unlimited cloud storage management", *Advanced Information Systems*, Vol. 2, No. 3, P. 49–53. DOI: https://doi.org/10.20998/2522-9052.2018.3.08
14. Sagala, A., Hutabarat, R. (2016), "Private Cloud Storage Using OpenStack with Simple Network Architecture", *Indonesian Journal of Electrical Engineering and Computer Science*, No. 4, P. 155–164. DOI: https://doi.org/10.11591/ijeecs.v4.i1.pp155-164
15. Shevchenko, V., Chengar, O., Kokodey, T. (2020), "Information technology for the deployment of the OpenStack cloud environment", *IOP Conference Series: Materials Science and Engineering*, No. 734:012131. DOI: https://doi.org/10.1088/1757-899X/734/1/012131
16. Luo, S., Ren, B. (2016), "The monitoring and managing application of cloud computing based on Internet of Things", *Computer Methods and Programs in Biomedicine*, No. 130, P. 154–161. DOI: https://doi.org/10.1016/j.cmpb.2016.03.024.

Received 25.08.2020

Відомості про авторів / Сведения об авторах / About the Authors

Давидов В'ячеслав Вадимович – кандидат технічних наук, Національний технічний університет "Харківський політехнічний інститут", доцент кафедри обчислювальної техніки та програмування, Харків, Україна; email: vyacheslav.v.davydov@gmail.com; ORCID: https://orcid.org/0000-0002-2976-8422.

Давыдов Вячеслав Вадимович – кандидат технических наук, Национальный технический университет "Харьковский политехнический институт", доцент кафедры вычислительной техники и программирования, Харьков, Украина.

Davydov Viacheslav – PhD (Engineering Sciences), National Technical University "Kharkiv Polytechnic Institute", Associate Professor of the Department of Computer Engineering and Programming, Kharkiv, Ukraine.

Гребенюк Дарина Сергіївна – магістр, Національний технічний університет "Харківський політехнічний інститут", аспірант кафедри обчислювальної техніки та програмування, Харків, Україна; email: darina.gg1@gmail.com; ORCID: https://orcid.org/0000-0001-5331-2444.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2020. № 3 (13) *ISSN 2524-2296 (online)*

Гребенюк Дарина Сергеевна – магистр, Национальный технический университет "Харьковский политехнический институт", аспирант кафедры вычислительной техники и программирования, Харьков, Украина.

Hrebeniuk Daryna – Master's Degree, National Technical University "Kharkiv Polytechnic Institute", Postgraduate Student of the Department of Computer Engineering and Programming, Kharkiv, Ukraine.

РОЗРОБЛЕННЯ МЕТОДІВ РОЗПОДІЛУ РЕСУРСІВ У СИСТЕМАХ ХМАРНИХ ОБЧИСЛЕНЬ

Предметом дослідження в статті є моделі та методи балансування та розмежування навантаження і ресурсів в системах хмарних обчислень, що базуються на моделі надання послуг інфраструктури як сервісу. **Метою** роботи є підвищення ефективності використання наявних ресурсів в системах хмарних обчислень (таких, як оперативна пам'ять, дисковий простір, ЦПУ, мережа) шляхом розробки моделі адаптивного управління розмежуванням ресурсів. Це дозволить запускати нові віртуальні машини з мінімальним зниженням продуктивності вже функціонуючих програм. У статті вирішуються наступні **завдання**: розробка комплексного підходу управління розмежуванням ресурсів у хмарних системах, яка включає в себе декомпозицію системи хмарних обчислень на зони (виходячи з визначальних особливостей надаваних в кожній зоні ресурсів), первинне виділення ресурсів (що базується на основі методу аналізу ієрархій) та розмежування ресурсів системи хмарних обчислень (на основі розробленого методу); розробка методу розмежування обчислювальних ресурсів у системах хмарних обчислень; оцінка ефективності розробленого методу. Для вирішення поставлених завдань були використані підходи і **методи** динамічного балансування навантаження, а також методи теоретичних досліджень, які засновані на наукових положеннях теорії штучного інтелекту, статичного, функціонального і системного аналізів. Отримані наступні **результати**: на основі проведеного аналізу існуючих методів розмежування і балансування навантаження в системах хмарних обчислень були виявлені основні особливості існуючих методів розподілу ресурсів, наведені їх переваги та недоліки. На основі проведеного аналітичного дослідження доведено необхідність вдосконалення існуючих методів розмежування ресурсів. Створено метод і алгоритм розмежування обчислювальних ресурсів в системах хмарних обчислень, що дозволяють зменшити значення коефіцієнта нерівномірності використання обчислювальних ресурсів при мінімізації витрат на їх переміщення. Отримані результати підтверджені проведеними експериментами при використанні програмного забезпечення для створення приватних інфраструктурних хмарних сервісів і хмарних сховищ. **Висновки**: вдосконалення методу розмежування і балансування навантаження в системах хмарних обчислень дозволило підвищити здатність цих систем запускати нові віртуальні машини з мінімальним зниженням продуктивності вже функціонуючих програм.

Ключові слова: інфраструктура як сервіс; хмарні обчислення; перерозподіл ресурсів.

РАЗРАБОТКА МЕТОДОВ РАЗГРАНИЧЕНИЯ РЕСУРСОВ В СИСТЕМАХ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ

Предметом исследования в статье являются модели и методы балансировки и разграничения нагрузки и ресурсов в системах облачных вычислений, базирующихся на модели предоставления услуг инфраструктуры как сервиса. **Целью** работы является повышение эффективности использования имеющихся ресурсов в системах облачных вычислений (таких, как оперативная память, дисковое пространство, ЦПУ, сеть) путем разработки модели адаптивного управления разграничением ресурсов. Это позволит запускать новые виртуальные машины с минимальным снижением производительности уже функционирующих приложений. В статье решаются следующие **задачи**: разработка комплексного подхода к управлению разграничением ресурсов в облачных системах, которая включает в себя декомпозицию системы облачных вычислений на зоны (исходя из определяющих особенностей предоставляемых в каждой зоне ресурсов), первоначальное выделение ресурсов (базирующееся на основе метода анализа иерархий) и разграничение ресурсов системы облачных вычислений (на основе разработанного метода); разработка метода разграничения вычислительных ресурсов в системах облачных вычислений; оценка эффективности разработанного метода. Для решения поставленных задач были использованы подходы и **методы** динамической балансировки нагрузки, а также методы теоретических исследований, которые основаны на научных положениях теории искусственного интеллекта, статического, функционального и системного анализов. Получены следующие **результаты**: на основе проведенного анализа существующих методов разграничения и балансировки нагрузки в системах облачных вычислений были выявлены основные особенности существующих методов распределения ресурсов, приведены их достоинства и недостатки. На основе проведенного аналитического исследования доказана необходимость совершенствования существующих методов разграничения ресурсов. Созданы метод и алгоритм разграничения вычислительных ресурсов в системах облачных вычислений, позволяющие уменьшить значения коэффициента неравномерности использования вычислительных ресурсов при минимизации расходах на их перемещение. Полученные результаты подтверждены проведенными экспериментами при использовании программного обеспечения для создания частных инфраструктурных облачных сервисов и облачных хранилищ. **Выводы**: усовершенствование метода разграничения и балансировки нагрузки в системах облачных вычислений позволило повысить способность этих систем запускать новые виртуальные машины с минимальным снижением производительности уже функционирующих приложений.

Ключевые слова: инфраструктура как сервис; облачные вычисления; перераспределение ресурсов.

Бібліографічні описи / Bibliographic descriptions

Давидов В. В., Гребенюк Д. С. Розроблення методів розподілу ресурсів у системах хмарних обчислень. *Сучасний стан наукових досліджень та технологій в промисловості*. 2020. № 3 (13). С. 25–33. DOI: https://doi.org/10.30837/ITSSI.2020.13.025.

Davydov, V., Hrebeniuk, D. (2020), "Development of the methods for resource reallocation in cloud computing systems", *Innovative Technologies and Scientific Solutions for Industries*, No. 3 (13), P. 25–33. DOI: https://doi.org/10.30837/ITSSI.2020.13.025.