Adaptation of cloud computing as optimization of the process of rendering services to users in the conditions of limited computing resources

© Aleksandr Matov

Institute for Information Recording of NAS of Ukraine, Kyiv, Ukraine

matov@ipri.kiev.ua

Abstract. We consider the cloud computing (CC) infrastructure as an object of adaptation and the process of cloud computing adaptation as an optimization. The general formulation of the problem of adaptation of the discipline of providing computing resources to the users of CC is outlined. The technology of dynamic adaptive mixed discipline of providing computing resources to users of CC is offered. The direction of solving the problem of optimization of dynamic adaptive mixed discipline is given.The well-known optimization functionality is proposed, based on the assumption that the results of the use of computing resources by the user (solving user problems) are depreciated in proportion to their time in the queue for the solution and the solution itself in the CC system. Other functionalities with time constraints are also possible. This is relevant for today's global real-time information and analytics systems using cloud computing technology and can be critical with limited computing resources. It is stated that the optimization problem is solved by an iterative method using the appropriate analytical models of the operation of CC.The description of such models is given.The stochastic nature of the main factors and the need to quantify mass processes based on probability theory determines the use of queuing theory. It is proposed to develop analytical models of cloud computing as a queuing system with mixed service discipline. Models should consider failures and different features of operation and, where possible, have arbitrary distribution laws for certain probable processes. Then it is possible and appropriate to use the technology of dynamic adaptive mixed discipline of providing computational resource to the users of CC as a mechanism of adaptation of CC. The mathematical formulation and method of solving such tasks are given.

Keywords: cloud computing, discipline of providing computing resources, absolute and relative priorities, adaptation and optimization of service disciplines, adaptation efficiency, mixed service discipline, mathematical model.

Introduction

Creating adaptive cloud computing infrastructures that are able to adapt dynamically to constantly changing conditions of operation, and developing appropriate computing organization methods is an important area of development of modern global information and analytical systems using cloud computing technologies.

Adaptation as control is secondary to the main control loop. If the management fulfills the basic goals, the realization of which ensures the functioning of the object, then the adaptation ensures the quality of this functioning. Therefore, when there is a need to improve (or maintain at the required level) the quality of the facility, there is always a need for adaptation.

The peculiarity of the adaptation system is that it is necessary to work in the conditions of considerable uncertainty of the environment and the handling of the object.

The ambiguity of the environment and the object is a feature that allows you to consider adaptation as a specific type of control. In this case, the degree of uncertainty determines the importance of solving the adaptation problem: the greater the uncertainty, the greater the need for adaptation.

Cloud computing as an object of adaptation

Cloud computing is an object with a high degree of uncertainty in the operation process. Here, the external uncertainty of the flow of computational resource (CR) requests (environment) is complemented by the internal uncertainty of the CC (object) associated with the presence or absence of the required CR, the random failures of the CC system, and the need to provide certain temporal characteristics for the many customers. This is what determines the need for introducing adaptation into the process of functioning of CC.

In addition, the introduction of adaptation to the process of functioning of the CC is associated with the need to maintain the system in an optimal and sometimes simply operational state, regardless of the numerous external and internal factors that bring the CC to the desired target state.

All of the above can equally be attributed to the computational process as an object of adaptation, because it develops in CC and is an integral attribute of it.

The notion of adaptation as an active action (control) is usually embedded in two meanings: adapting an object to a fixed environment (passive adaptation) and finding an environment appropriate to that object (active adaptation) [1]. In the first case, the adaptable entity functions to fulfill its goal in the best possible environment, that is, to maximize its effectiveness in that environment. Active adaptation, on the contrary, implies a change of environment in order to maximize the performance of the object.

With respect to CC, as a queuing system), active adaptation can be seen as a change in the intensity or quantity of incoming application flows, as well as the laws of the distribution of the application process.

Passive adaptation is most commonly used in the operation of CC, in which adaptive influence may have different character. It may change either the parameters of the adaptation object (parametric adaptation) or its structure (structural adaptation).The intensity and laws of the distribution of the application service process, restrictions on the waiting time (stay) of applications in the queue (system), the order of service of applications (service discipline), etc. can be considered as managed parameters of CC. An example of a structural adaptation that changes the number of servicing devices and the relationship between them is the reconfiguration of multi-server CC. Structural adaptation is more radical and is usually accompanied by parametric adaptation, because each structure has its own parameters.

Depending on whether the model is an object of adaptation or not, there are two very important types of adaptation: adaptation with and without model (search adaptation), which differ significantly from each other [1].

In the presence of an adequate object model, it is sufficient to measure the state of the environment for the synthesis of the adaptive impact, and using the model to determine the impact that should put the object in the desired state.

However, very often, the object of adaptation is so complex that it is impossible to build a model of it, and an adequate model is all the more so.

At the same time, it is probably not possible to use the adaptation method with the model, which forces to resort to search adaptation. This type of adaptation is distinguished by the presence of search, a specially organized process that allows you to determine the necessary adaptive impact without having an object model. Search engine adaptation is characterized by experiments with an object, in the process of which they obtain information about its properties. This information determines the adaptive impact of the object's performance.

The search engine adaptation process itself is a consistent, multi-stage process - steps are taken at each stage to improve the performance of the facility (as opposed to adapting to a one-stage adaptation model).

If, when adapting to a model, the state of the object is to be measured only to adjust its model and not required for the adaptation itself, then in the search for adaptation the state of the object carries the basic information for forming the influencing adaptation.

The difficulty of adapting to a model lies in the synthesis of the model of the object, and the adaptation itself is the solution of the optimization problem of selecting such an impact formation that would satisfy the adaptation goals. Search engine adaptation has other difficulties - you need to experiment with the object at the same time and adapt it.

In all cases, when it is possible to build an adequate model of the object, the question of choosing the type of adaptation weighs unequivocally in favor of adaptation with the model, because only the presence of the model allows you to quickly adapt the object.

Cloud computing adaptation as optimization

The solution to the problem of adaptation is to determine the kind of control (adaptive) impact that maximizes the performance of the object in the current situation.

The situation is characterized by two factors: the state of the environment in which the object is located and the state of the adaptation object itself.

For CC, as a queuing system, the state of the environment can be understood, for example, the intensity of incoming requests, and the state of the object (system) - the number or time of waiting (stay) of requests in the queue (system), or the malfunction of the serving device, system boot level, etc.

Depending on the current situation, an adaptive effect should be formed that minimizes the average number or average waiting time (stay) of applications in the queue (system), or the time of entering the system in a steady state, or the total cost for the system operation, or the probability of losing applications, etc. e. The purpose of the adaptation may be to maximize revenue from service requests, eliminate system overload, and maintain it in a stationary mode.

Thus, the adaptation of CC can be considered as a process of optimizing work in the current situation.

General statement of the task of adaptation of the discipline of service

The task of adapting the discipline of service in the CC is due to unforeseen and uncontrolled changes in the environment and system, which inevitably alter the optimal setting of the discipline of service, if one was implemented in the system. Therefore, the systematic adjustment (adaptation) of the discipline of service is inevitable if you wish to maintain the system in optimal mode, regardless of changes occurring in the environment and system.

We formulate in general terms the task of adapting the discipline of service [2].

Let X and E be the controlled and uncontrolled states of the medium. The {X, E} pair uniquely describes the environment in which the CC is located. For example, X is the passport data of service requests and E is the intensity of their receipt.

Similarly, the pair {Y, H} describes the state of the system. Here, Y and H are respectively controlled and uncontrollable factors. For example, Y is the length of the application queues, and H is the service intensity or system failure rate.

The performance of the system is extreme. It is defined on the controlled states of the environment and system:

$$\acute{Y} = \acute{Y}(X, Y) \,. \tag{1}$$

The system performance indicators may be the average time (waiting) of applications in the system (queues), the average length of the application queue, the average total cost of waiting (staying) applications in the queue (system), etc.

The status of system Y depends on X, E and H, as well as on the discipline of servicing S:

$$Y = F(X, E, H, S) \,, \tag{2}$$

where F is the system operator.

Service discipline refers to the rule of selecting service requests depending on the state of the environment and system:

$$S = S(X, Y)\,. \tag{3}$$

In most cases, the optimality of discipline S is related to the extremisation of the performance of the system (1).This means that for the synthesis of optimal discipline, the following optimization problem must be solved:

$$\acute{Y}[X, F(X, E, H, S)] \rightarrow \underset{S \in S^*}{extr} \Rightarrow S^0\,, \tag{4}$$

where - restrictions imposed on the choice of discipline of service *S*.

These restrictions may be related, for example, with a certain set of predefined service disciplines, etc.

Obviously, it is impossible to solve problem (4) at the stage of designing a computer system , because a priori unknown factors *E* and *H*. Averaging over these factors cannot be entered because they can be non-stationary.

Therefore, the problem of synthesis of optimal discipline S^0 should be solved by the adaptation of CC, that is, in the mode of their operation. Then adaptation is reduced to solving the problem

$$\acute{Y}(S) \rightarrow \underset{S \in S^*}{extr} \Rightarrow S^0$$

in different disciplines:

$$\hat{\acute{Y}}_1 = \hat{\acute{Y}}(S_1), ..., \hat{\acute{Y}}_\xi = \hat{\acute{Y}}(S_\xi).$$

The adaptation algorithm must specify the sequence of transition from one discipline to another: which leads to a solution that is optimal in the current situation.

Use for dynamic adaptive technology adaptationthe mixed discipline of providing computing resourcesusers of CC

Currently, a large number of different service disciplines are known. Of these, the disciplines of service with relative and absolute priorities are widely used in CC. However, these disciplines are static and therefore have a number of significant disadvantages that reduce the efficiency of computing systems (processes) in the uncertainty of the environment and the behavior of the systems themselves.

When using discipline with relative priority, the selection of a regular service request can only be made after the completion of the current service, even if the service request has a lower priority. As a result, the length of your stay at the CC may be unacceptably long for some of your most important applications. Reducing the delay in servicing important applications is achieved by interrupting, that is, introducing absolute priority for these applications. However, the duration of low priority applications is increased by the CC, and in some cases, with the intensive receipt of important applications, the process of servicing low priority applications may be blocked, which also reduces the effectiveness of the CC as a whole.

In order to compensate for the disadvantages inherent in the disciplines of service with relative and absolute priorities and taking into account their advantages, it is advisable to implement in the mixed disciplines of service that use both relative and absolute priorities.

Consider one of the mixed service disciplines. Let the *N* input streams of applications according to their importance and urgency in service be divided into *M* groups, between which there is an absolute priority, and inside a relative one. This means that requests from any stream from a group $m(m = \overline{1, M})$ interrupt the service of requests belonging to streams from groups with numbers. Each group contains N_m threads whose requests do not interrupt each other $\overline{m + 1, M}$. It is obvious that $\sum_{m=1}^{M} N_m = N$. The priority of any application in a system with such a service discipline can be described by a pair of numbers m and n, $n = \overline{1, N_m}$, where it determines the number of the request flow in the group with the number m.

The described mixed discipline of service allows to adapt more flexibly to various situations arising during the functioning of the CC due to its adaptation. In this case, the adaptation of the discipline consists in changing the number and position of the boundaries that separate the flow of applications into groups of absolute priority, that is, in changing the number of groups and the number of flows in groups. Grouping options will be called breakdowns. The total number of breakdowns of F is determined by the number of requests streams $N : \hat{O} = 2^{N-1}$. Each breakdown $\varphi(\varphi \in \hat{O})$ is given by a set of numbers $\{N_1, N_2, \ldots, N_M\}$.

Such a discipline is rightly called the dynamic adaptive mixed discipline of providing computing resources to users of CC

Introduced mixed discipline of service is a known practical interest, because in the optimal selection of the breakdown of flows into groups, in principle, provides no worse service compared to "pure" disciplines (with relative and absolute priorities). Thus $M = N,\ N_m = 1$, for all $m = \overline{1, M}$, there is a discipline of service with absolute priority, and when $M = 1,\ N_1 = N$ - with relative.

Tasks of dynamic adaptive mixed discipline providing computing resources with the model

Let's consider two practical problems of dynamic adaptive mixed discipline of providing computing resources (mixed discipline of service) with the model.

One of the main indicators of the effectiveness of CC is indicators based on the temporal characteristics of these systems. Such metrics can be set by the contract between the supplier and the user of the CR CC and are of particular importance for real-time systems.

Due to the random nature of the computing process, there are additional delays in the processing of information, violating the permissible restrictions on its time in the CC, which adversely affects the effectiveness of solving targeted user tasks.

In such situations, it is necessary to maintain the time characteristics of the system at a predetermined level in order to ensure the necessary efficiency of the CC. In conditions of scarcity of computing resources, this is possible only by improving the efficiency of the computing process, in particular, by adapting the discipline of servicing. At the same time, there is a problem of the most efficient use of available computing resources at each moment of time of operation of the control CC. This task can also be addressed by adapting the discipline of service.

In view of the foregoing, we will choose the average total cost of provisioning time (waiting in queues and time of use, ie staying in CCs as in queuing system) CR as a measure of the performance of the CC according to the requests (requirements) of users. To do this, we use the well-known functional [1] - the average total cost of time to provide the CR:

$$C^{(S)} = \sum_{i=1}^{n} \alpha_i \lambda_i v_i^{(s)} ,$$

what do we have

$$C^{(\varphi)} = \sum_{m=1}^{M} \sum_{n=1}^{N} \alpha(m,n) \lambda(m,n) v^{(\varphi)}(m,n) , \qquad (5)$$

where

α_i - is the cost per unit of time CR for the i-th type of user requests;

λ_i - the intensity of the *i*-th flow of applications;

$v_i^{(S)}$ - the average time for submitting the CR applications to the *i*-th stream;

n - is the number of application types;

s - parameter characterizing the method of organization of the computing process;

$v^{(\varphi)}(m,n)(m=\overline{1,M}, n=\overline{1,N_m})$ - the average time of the provision of the CR in the CC application (*m*, *n*)-th flow;

α (*m*, *n*) - is the unit cost of the time the CR is submitted to the CC (*m*, *n*)-th request;

λ (*m*, *n*) is the intensity of the (*m*, *n*) flow of the CR in the CC.

The performance indicator is based on the assumption that the results of the user's use of the CR are depreciated in proportion to their time in the CC system. Then the goals of adaptation of the mixed discipline of service will be either to satisfy the requirements of the timely stay (m, n) of applications in the system, which are set by acceptable values of this time $v_{\ddot{A}}(m,n)$, or to minimize the functional (5). This goal is achieved by finding the appropriate optimal breakdowns φ^0, that is, the tasks of adapting a mixed service discipline with relative absolute priority are optimization problems, the general formulation of which is discussed above.

Since the above goals of adapting mixed discipline to service can be achieved with several different breakdowns of requests flow into groups of absolute priority, there is a need to introduce additional restrictions on the choice of breakdown φ .

The presence of an absolute priority in CC requires some technological losses of the CR, which are proportional to the number of groups (levels) of the absolute priority. In this regard, it is optimal to consider such a breakdown that ensures that the adaptation goals are attained with a minimum number of absolute M priority groups.

Then the considered problems of adaptation of mixed discipline of service can be formally set as follows:

$$\begin{gathered} v^{(\varphi)}(m,n) \le v_{\ddot{A}}(m,n) \Rightarrow \varphi^{0}, \\ \varphi \in \boldsymbol{\Phi}, \\ M = \min . \end{gathered} \tag{6}$$

$$\begin{gathered} C^{(\varphi)} \to \min \Rightarrow \varphi^{0}, \\ \varphi \in \boldsymbol{\Phi}, \\ M = \min . \end{gathered} \tag{7}$$

It is not possible to solve the problems of finding the optimal partition (6) and (7) using known analytical optimization methods. The only way to solve these problems is a heuristic approach, which has no formal justification, but relies solely on the specifics of the mathematical models [3…11] and related understanding.

It follows from expressions (5) - (7) that the achievement of the goals of adaptation of the mixed discipline of service is combined with the need to estimate the value of the average residence time in the application system (*m*, *n*) -type - *v*(*m*, *n*). Therefore, there is a need to synthesize a mathematical model of CC with a mixed discipline of maintenance [3].

Characterization of analytical models and mathematical formulation of tasks in queuing theory

One of the main indicators of the effectiveness of CC is the indicators based on the assessment of the time characteristics of these systems. Violation of permissible time constraints, for example, the response time of the CC, affects the effectiveness of the solution of user targets, which is of particular importance for real-time systems. First of all it concerns special information systems, which are built using private CC.

The stochastic nature of the main factors and the necessity of quantification of mass processes on the basis of the theory of probability determines the use of the theory of mass service. Then it is possible and appropriate to use the technology of the dynamic adaptive mixed discipline of providing CR (maintenance) to users of the CC as mechanisms of adaptation of the CC [1].

Analytical models for calculation of time characteristics are offered in the conditions of the features of the functioning of the CC using a mixed discipline of service with absolutely relative priorities and taking into account failures [3].

It is proposed to develop analytical models of cloud computing as a queuing system with mixed resource delivery discipline. Models should consider failures and different features of operation and have arbitrary distribution laws for some probable processes.

The general description of the models is as follows. Let the input of the CC system, in which the discipline of service with a relatively absolute priority is implemented, arrive *N* Poisson flows of applications of intensity $\lambda(m,n) \quad (m=1,M, \quad n=1,N_m)$. These flows are aligned with *N* priorities [1].

The duration of the maintenance of applications of priority (*m*, *n*) is a random variable with a distribution function $B_{m,\ n}(t)$, the first *b* (*m, n*) and the second $b^{(2)}(m,n)$ start point.

An application of priority (*m*, *n*) whose service is interrupted by applications from groups with $1,m-1,$ numbers is returned to the queue. Updating its service is possible either after servicing all interrupted applications (maintenance discipline A), or after servicing all interrupted applications and all applications for accumulated flows, the *m* group with $(m,1),(m,n-1)$ numbers (discipline of service upgrade B) .

The serving device (CC) fails in accordance with the Poisson law with the λ_0 parameter. The period of recovery of the device is a random variable that has an arbitrary distribution law *B_o(t)* with the first b_0 and second b_0^2 initial moments.

During the restoration of the service device, requests of some streams in the queue are accepted, while others are not accepted. This condition is given by the matrix-row of coefficients $n_i, i=1,N$, and in the case if requests of the $n_i=1$ stream are accepted in the queue, and if requests $n_i=0$ are denied.

Adaptation to bounce will be that in the period of recovery device incoming applications can either accumulate in the queue (discipline replenishment queue I), or receive a refusal and leave the system (discipline replenishment queue II).

Failure of the servicing device can occur both during its free state and during service of the application. In the latter case, the renewal of the service is carried out either from the interrupted application, if there are no applications interrupting its service, (the discipline of the renewal of service C), or from applications of the senior relative priority of the corresponding group, if any (discipline of renewal of service D).

In case of repeated receipt of the servicing device, the interrupted application shall be maintained from the place where it was interrupted. Within one priority, applications are served in the order of receipt.

The combination of service updating disciplines and queue replenishment allows you to consider independent models of different types of systems that have the proper designation. Different features of functioning consist of various combinations of disciplines A, B, C, D, I and II.

Let CC be in stationary mode, which $R_M < K_r$ condition is for systems of type I, and for systems of type II - $R_M < 1$. Here $R_M = \sum_{m=1}^{M}\sum_{n=1}^{N}\rho(m,n)$ - total loading of the device applications ($(\rho(m,n) = \lambda(m,n)b(m,n)$ - loading of the device (m, n) - applications), and $K_r = 1/(1+\rho_0)$ - the system readiness coefficient ($(\rho_0 = \lambda_0 b_0$ - loading the device with refusals).

It is necessary to determine the average $v(m,n)$ time spent in the system of applications of each (m, n) -priority, ie, the response time of the system CC. To determine the average time of applications in the system (time response systems) use the known direct method [1].

Conclusions

- To develop the basics of creating adaptive cloud computing infrastructures capable of dynamically adapting to current changes in operating conditions. Dynamic adaptive discipline for providing cloud computing resources is proposed.
- Cloud computing is an object with a high degree of randomness of the process of operation, the main factors of which are: the probability of flow of requests for computing resources; availability of necessary resources and accidental timing of their use by consumers; accidental failure of the CC infrastructure and time of their elimination.
- Due to the random nature of the computing process, there are additional delays in the processing of information, violated permissible restrictions on its time in the system (at the time of the CC response), which adversely affects the effectiveness of solving targeted user tasks. This is relevant for real-time systems and, above all, for special information systems built using private clouds, and can be critical with limited CCcomputing resources.
- It is possible to reduce or reduce the impact of beneficial phenomena on the functioning of the CCby introducing adaptation to the functioning of the CC infrastructure. In addition, the introduction of adaptation is associated with the need to support CC in the optimal (efficient use of resources) and sometimes just in working order, regardless of the many factors that drive the CC infrastructure from the desired target state.
- Depending on the situation, some adaptive influence should be generated, for example, minimizing the average number or average waiting time (stay) of applications in a queue (system), or the time of entering the system in a steady state, or the total cost for system operation, or the likelihood of losing applications, etc. .. The purpose of adaptation may be to maximize revenue from the service of applications, eliminate overloading the system and maintain it in a stationary mode of operation. Thus, the adaptation of the CC infrastructure can be considered as a process of optimizing the work in the current situation.
- The adaptation problem can be solved by using the adaptive discipline (order) of providing computing resources to users. Unforeseen and uncontrolled changes in the

environment and the system will inevitably alter the optimal setting of the discipline, if one has been implemented in the system. Therefore, systematic adjustments (adaptation) of the discipline are inevitable if you wish to maintain the system in optimal mode, regardless of changes occurring in the environment and system.

- Currently, a large number of different disciplines are known. Of these, disciplines with relative and absolute priorities are widely used for SS infrastructure. However, these disciplines are static and, as a result, have a number of significant drawbacks that reduce the efficiency of CC processes in the context of environmental uncertainty and the behavior of the systems themselves.
- The discipline of providing computing resources to data center users as an object of adaptation and the process of adaptation as an optimization is considered. The general statement of the adaptation problem as optimization is outlined.
- Finding the best discipline is not always linked to the extremisation of the data center performance indicator. The purpose of adaptation may also be to satisfy the performance constraint given by equals or inequalities. In any case, this formulation of the adaptation problem implies the need to implement dynamically changing several different or one mixed discipline of providing computing resources to CC users.
- The technology of dynamic adaptive mixed discipline for providing computing resources to CC users is proposed. The direction of solving the problem of optimization of dynamic adaptive mixed discipline is given.
- The mixed discipline of providing computing resources allows it to respond more flexibly to the various situations that arise in the functioning of CC due to its adaptation. In this case, the adaptation of the discipline consists of an optimal change in the number and position of boundaries that divide the flows of user requests for resources into groups of absolute priority, within which the relative priority, ie in the change of the number of groups and the number of flows in groups, operates.
- A well-known optimization functional is proposed, based on the assumption that the results of the use of computing resources by the user (solving user problems) are depreciated in proportion to their time in the CC. Other functionalities with time constraints are also possible. Such metrics can be specified by the agreement between the supplier and the user of computing resources.
- It is not possible to solve the problems of finding the optimal breakdown into groups of flows of providing computing resources to users using known analytical methods of optimization. The only way to solve these problems is a heuristic approach, which has no formal justification but relies on the specifics of the problems (mathematical models) that should be able to determine the temporal characteristics of the CC, such as response time.
- CC analytical models are proposed as multi-threaded and multi-priority queuing systems with mixed service discipline. Models take into account failures and different features and have arbitrary distribution laws for some probable processes. The mathematical formulation and the proposed method of solving service problems for obtaining analytical expressions of temporal characteristics, including the response time of CC, which allow to realize dynamic adaptive discipline of providing computing resources to users of cloud computing, are made.

References

1. Matov A.Y., Shpilev V.N., Komov A.D. et al.: Organization of computational processes in ACS. Ed. A.Ya.Matov. Kiev 200s. (1989). (in Russian)
2. Matov A.Y.: Optimization of the provision of computing resources with adaptive cloud infrastructure. Data recording, storage and processing. T.20, N.3, 83-90 (2018). (in Ukrainian).
3. Matov Aleksandr: Mathematical models of cloud computing with absolute-relative priorities of providing of computer resources to users in conditions of functioning features and failures. CEUR Workshop Proceedings Vol-2318. Selected Papers of the XVIII International Scientific and Practical Conference on Information Technologies and Security (ITS 2018) PP. 150-159. [http://ceur-ws.org/Vol-2318/paper13.pdf]
4. Mokrov E.V., Samuilov K.E.: Cloud computing system model in the form of a queuing system with multiple queues and with a group of requests. https://cyberleninka.ru/article/n/model-sistemy-oblachnyh-vychisleniy-v-vide-sistemy-massovogo-obsluzhivaniya-s-neskolkimi-ocheredyami-i-s-gruppovym-postupleniem-zayavok., last accessed 2019/10/26.
5. Tsai J.M., Hung S.W.: A novel model of technology diffusion: system dynamics perspective for cloud computing. Journal of Engineering and Technology Management. V. 33. P. 4762. (2014).
6. Singh P., Dutta M.: Aggarwal N. A review of task scheduling based on meta-heuristics approach in cloud computing / Knowledge and Information Systems. V. 52. N 1. (2017).
7. .Grusho A.A., Zabezhailo M.I., Zatsarinny A.A.: Information flow monitoring and controlling the cloud computing environment. Informatics and Applications.Vol. 9.No 4. P. 91–97 (2015). (in Russian).
8. Gudkova I.A., Maslovskaya N.D.: Probability model for analyzing impact of delays due to monitoring on mean service time in cloud computing. T-Comm: Telecommunications and Transport. No 6.P. 13–15 (2014). (in Russian)
9. Gorbunova A.V., Zaryadov I.S., Matyushenko S.I., Samuylov K.E., ShorginS.Ya.: Approximation of the response time of a cloud charge system. Computer science and its applications. (2015). (in Russian)
10. Bezzateev S.V., Elina T.N., Mylnikov V.A.: Modeling the processes of selecting parameters of cloud systems to ensure their stability, taking into account reliability and security. Scientific and technical bulletin of information technologies, mechanics and optics. 2018.Vol. 18. No. 4. P. 654–662. (2018). (in Russian)
11. Gudkova I.A., Maslovskaya N.D.: A probabilistic model for analyzing access latency to a cloud computing infrastructure with a monitoring system / T-Comm: Telecommunications and Transport. No. 6. S. 13-15. (2014). (in Russian)