Adaptive management of the order in which resources are provided to cloud users

© Aleksandr Matov

Institute for information recording of NAS of Ukraine, Kyiv, Ukraine

Abstract. The review considers the issues of further development of the principles of creating adaptive infrastructures of cloud computing, capable of dynamically adapting to user requirements and current features and changes in operating conditions. Methods and analytical conditions for adapting the provision of resources to users of cloud computing have been developed. These conditions provide an opportunity to develop technology (mechanisms and algorithms) for the use of adaptive discipline (order) of providing computing resources to users. In turn, this allows you to meet the time requirements of different users to obtain timely computational results or make the most efficient use of available cloud computing resources. . This is relevant for real-time systems and, above all, for special information systems built using private clouds, and can be critical with limited computing resources of cloud computing.

Analytical (formulaic) conditions of adaptation are developed on the basis of the corresponding indicators of efficiency and mathematical models of cloud calculations. The stochastic nature of the main factors and the need to quantify mass processes based on probability theory determines the use of the analytical model of cloud computing as a multi-threaded and multi-priority queuing system with queues with mixed service discipline. The model takes into account probable failures and various features and has arbitrary distribution laws for some probable processes. The model allows to calculate the time characteristic - the response time of the system in terms of features of operation and failures of cloud computing.

Keywords: cloud computing, mathematical model, discipline of computing resources provision, mixed service discipline, absolute and relative priorities, time characteristics, response time, efficiency of cloud computing.

1 Introduction

Creating adaptive infrastructures that are able to adapt to changing operating conditions and maintain systems in optimal, and sometimes just in working order, is an important direction in the development of modern global information and analytical systems using cloud computing (CC) technologies. For such adaptation, a dynamic adaptive mixed discipline of providing computing resources to users of CC is proposed [1,3].

Consider two practical problems of dynamic adaptation of a mixed discipline of resource provision with relative-absolute priorities. One of the main indicators of the effectiveness of CC are indicators based on the assessment of the temporal characteristics of these systems and which must be maintained at a given level. Such indicators can be set by agreement between the supplier and user of CC and are especially important for systems primarily for special information systems based on private clouds. Due to the random nature of the computational process, there are additional delays in information processing, the permissible limitations for the time of its stay in the CC are violated, which negatively affects the effectiveness of solving target tasks of users.

To ensure the required efficiency of CC in such situations, it is necessary to maintain the time characteristics of the system at a given level. Given the shortage of computing resources, this is possible only by increasing the efficiency of the computing process, in particular, by adapting the discipline of service.

Along with this, there is the problem of the most efficient use of available computing resources at any time during the operation of the management of CC. This problem can also be solved by adapting the discipline of service.

2 Indicators of resource efficiency for users of cloud computing

The aim of the work is to develop methods and analytical conditions, adaptation of the provision of computing resources to users of CC to ensure the time characteristics of information and analytical systems and optimize the use of resources of CC.

As an indicator of the effectiveness of CC we take the average total cost (fine) of response time CC, time delay in the queue, waiting time in queues and time to provide resources, ie stay in the CC as in the queuing system (QMS)) on applications (requirements) of users. To do this, use the known functionality [3]:

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

from what we have

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

where

α_i - cost (fine) per unit of response time of CC (delays, stay in CC) of applications of the i-th stream;

λ_i - intensity of the i-th stream of applications;

$v_i^{(S)}$ - the average response time of CC applications of the *i*-th stream;

n - is the number of types of applications;

s - is a parameter that characterizes the method of organizing the computational process;

$v^{(\varphi)}(m,n)(m=\overline{1,M}, \quad n=\overline{1,N_m})$ - the average response time of CC applications (m, n) -th stream;

α (*m*, *n*) - the unit cost of response time CC (delay in HO) applications (*m*, *n*) -th stream;

λ (*m*, *n*) - intensity (*m*, *n*) -flow.

This efficiency indicator is based on the assumption that the results of the use of resources by the user are depreciated in proportion to the time of their delay in the CC, ie stay in the CC as in the QMS. Then the purposes of adaptation of the mixed discipline of service will be either satisfaction of requirements of timely stay (m, n) of applications in the system set by admissible values of this time, or minimization of functional (1). These goals are achieved by finding the appropriate optimal breakdowns into relative and absolute priorities, ie the problems of adaptation of a mixed service discipline with a relative-absolute priority are optimization problems, the general formulation of which is discussed above.

Since the above objectives of adapting a mixed service discipline can be achieved with several different breakdowns of application flows into groups of absolute priority, it is necessary to introduce an additional restriction on the choice of breakdown.

The presence of absolute priority in HO requires some technological losses of resources, which are proportional to the number of groups (levels) of absolute priority. In this regard, it is necessary to consider the optimal breakdown, which ensures the achievement of adaptation goals with a minimum number of groups of absolute priority M.

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

$$v^{(\varphi)}(m,n) \le v_{\text{Д}}(m,n) \Rightarrow \varphi$$

$$\begin{matrix} \varphi \in \Phi \\ M = \min \\ C^{(\varphi)} \to \min \end{matrix} \Rightarrow \varphi^0 \quad ; \tag{2}$$

$$\begin{matrix} \varphi \in \Phi \\ M = \min \end{matrix} \quad . \tag{3}$$

It is not possible to solve the problems of finding the optimal breakdown (2) and (3) using known analytical optimization methods. The only way to solve these problems is a heuristic approach, which has no formal justification, but is based only on the specifics of problems (mathematical models) and related understandings.

From expressions (1) - (3) it follows that the achievement of the goals of adaptation of the mixed service discipline is associated with the need to assess the value of the average response time of CC (stay in CC) applications (m, n) -type v (m, n) on resources. Therefore, there is a need to synthesize a mathematical model of CC with a mixed discipline of providing computing resources (maintenance).

3 Cloud infrastructure model class

Development of mathematical models of cloud computing or information systems created using clouds is an important area for identifying and improving their characteristics [2…10]. Cloud computing is an object with a high level of uncertainty in the operation process. Here, the external uncertainty of the flow of requests for computing resources (CR) (environment) is complemented by the internal uncertainty of the CC (object), which is associated with the presence or absence of the necessary CR, accidental failures of the CC system, as well as the need to provide certain time characteristics for many clients. . This determines the need for the introduction of adaptation into the functioning of the CC.

In addition, the introduction of adaptation into the process of functioning of CC is associated with the need to maintain the system in optimal and sometimes simply operational condition, regardless of the many external and internal factors that remove CC from the required target state.

Cloud computing (CC) is an object with a high level of uncertainty in the functioning process, the main factors of which are [1]:

-probability of the flow of requests for computing resources (CR);

-the presence of the necessary PR and the randomness of the time of their use by customers;

-accidental failures of the infrastructure of CC and the time of their elimination;

-the need to provide certain time characteristics for a number of customers, for example, the response time of CC;

-the need for optimal use of CR depending on the cost of delay time ordered by customers, the results of calculations and operating conditions;

-the need to introduce adaptation into the process of operation of the CC in order to provide certain time characteristics for a number of customers and the optimal use of CR.

The stochastic nature of the main factors and the need to quantify mass processes based on probability theory determines the use of queuing theory. Then it is possible and expedient to use the technology of dynamic adaptive mixed discipline of providing PR (service) to users of CC as mechanisms of adaptation of CC [1].

Analytical models for subtraction of time characteristics in the conditions of features of functioning of CC with use of mixed discipline of service with absolutely - relative priorities and the account of failures are offered. Models are based on works [2, 3].

4 Mathematical description of multi-threaded and multi-priority model of cloud infrastructure operation with queues with mixed service discipline and failure adaptation.

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 $\lambda(m,n)$ $(m=1,M,$ $n=1,N_m)$. These flows are aligned with N priorities [2].

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 $R_M < K_{r_M}$ condition is for systems of type I, and for systems of type II - $R_M<1$. Here $R_M=\sum_{m=1}\sum_{n=1}\rho(m,n)$ - 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.

5 Definition of time characteristics of a model of a system of type AS-I.

To determine the average time of applications in the system (time response systems) type AS-I use the known direct method [3].

Let some application (j, k) be a priority in the system. The average duration of this application in the system v (j, k) consists of the average waiting time in the queue w (j, k) and the average service time b (j, k):

$$v(j,k) = w(j,k) + b(j,k) .$$

The average waiting time in the queue w (j, k) consists of the average waiting time before service and the average standby time in the interrupted state u (j, k):

$$w(j,k) = w_H(j,k) + u(j,k) .$$

The last term in this formula is due to the interruptions in the maintenance of the application (j, k) -priority of applications from groups $\overline{1, j-1}$ and denials, that is:

$$u(j,k) = u_3(j,k) + u_0(j,k) .$$

Average time from the beginning of service (j, k) - application to completion is the average full time of service:

$$\Theta(j,k) = b(j,k) + u(j,k) . \qquad (4)$$

Let's start with the calculation u (j, k), for which we apply the approach described in [2].

During the service (j, k) -supply on average will occur $b(j,k)\Lambda_{j-1}$ interruptions

where $\Lambda_{j-1} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\lambda(m,n)$ the intensity of the total flow of interrupted applications.

As a result of these interruptions (j, k), the application returns to the queue and waits for the termination of service interruptions that will continue in $b(j,k)R_{j-1}$ average units of time

where $R_{j-1} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\lambda(m,n)b(m,n)$. (5)

During this time, applications from groups $\overline{1, j-1}$ will be received, which will lead to an increase in waiting time (j, k) - applications for value $b(j,k)R_{j-1}^2$. In addition, the service of these applications will be accompanied by additional accumulation of applications of the same priorities, requiring service before (j, k) -payment. This process is endless, with supplements to the waiting time (j, k) -positions form a declining geometric progression with a denominator $R_{j-1} < 1$. The sum of members of such geometric progression is the mean time of all service interruptions (j, k) -request:

$$T^{(1)} = b(j,k)\frac{R_{j-1}}{1-R_{j-1}}. \quad (6)$$

In the mean time $T^{(1)}$, the device will fail $T^{(1)}\lambda_0$, resulting in it will be restored within $T^{(1)}\lambda_0 b_0 = T^{(1)}\rho_0$ units of time. Since in the system type AS-I during the period of recovery the device again receives applications that continue to accumulate in the queue, then after the device is restored, the average waiting time (j, k) -supply in the interrupted state will increase by $T^{(2)} = T^{(1)}\rho_0 \frac{R_{j-1}}{1-R_{j-1}} = b(j,k)\rho_0 \frac{R_{j-1}^2}{(1-R_{j-1})^2}$. (7)

During this time there may be a refusal of the device, the restoration of which will be accompanied by the accumulation of new applications served before (j, k) -payments, etc.

The total time of all applications service interruptions (j, k) -priority of $\overline{1, j-1}$ application groups, taking into account device refusals $u_3(j,k) = T^{(1)} + T^{(2)} + \ldots + T^{(\infty)}$. This expression represents the sum of two infinitely decreasing geometric progressions. After calculating the sum of the members of each of them and compiling the results, we get: $u_3(j,k) = b(j,k)\frac{R_{j-1}}{K_r - R_{j-1}}$. (8)

Similarly, the average waiting time (j, k) is determined in the interrupted state due to device refusals $u_0(j,k)$. The only difference is the beginning of reasoning. During the service (j, k) -supply, the device will fail on $b(j,k)\lambda_0$ average, which will result in its restoration within $b(j,k)\rho_0$ units of time. Taking into account the possibility of accumulation in the period of device renewal and priority service of applications with absolute priority from $\overline{1, j-1}$ group, the average waiting time (j, k) -payments will increase by $b(j,k)\rho_0 \frac{R_{j-1}}{1-R_{j-1}}$.

During this time, the device can again be denied, which additionally increases the waiting time (j, k) - request for value $b(j,k)\rho_0 \frac{R_{j-1}}{1-R_{j-1}}$ etc.

In the final analysis, we get: $u_0(j,k) = b(j,k)\frac{K_r\rho_0}{K_r - R_{j-1}}$. (9)

Then the total average waiting time $R_{j-1}(j,k)$ request in the interrupted state: $u(j,k) = b(j,k)\frac{R_{j-1}}{K_r - R_{j-1}}$, (10)

and the total average service time (j, k) -request:

$$\Theta(j,k) = b(j,k)\frac{1}{K_r - R_{j-1}} . \quad (11)$$

Now calculate $w_H(j,k)$. Before (*j*, *k*) -request entered the system for the first time, the following should be done:

1) the device is restored

2) an application has been served from $\overline{1,j}$ or groups of submissions of the served application from the $\overline{j+1,M}$ groups;

3) service requests from $\overline{2,j}$ groups interrupted by applications from $\overline{1,j-1}$ groups;

4) service requests from $\overline{1,j}$ groups interrupted by denials of the device;

5) existing requests for streams with numbers $\overline{(1,1),(j,k)}$ are served;

6) service requests flowed with numbers $\overline{(1,1),(j,k-1)}$ received during the waiting time (*j*, *k*) -request, taking into account device refusals.

$$w_H(j,k) = \sigma_0 + \sigma(j,k) + \eta(j,k) + \eta_0(j,k) + \sum_{m=1}^{j-1}\sum_{n=1}^{N_m} w_H(m,n)\rho(m,n) + \sum_{n=1}^{k} w_H(j,n)\rho(j,n) +$$

For the average duration of these events, we write the equation:

$$+[\sigma_0 + z_H(j,k)]\frac{R_{j,k-1}}{K_r - R_{j,k-1}} + z_H(j,k)\frac{K_r\rho_0}{K_r - R_{j,k-1}} . \quad (12)$$

Here $\sigma_0 = K_r\rho_0\Delta_0$ - average time for updating the device in the presence (j, k) -position: $K_r\rho_0$ - probability of recovery of the device [2], $\Delta_0 = b_0^{(2)} / 2b_0$;

$\sigma(j,k) = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\rho(m,n)\Delta(m,n)$ - average time for the maintenance of the application by the device in the presence (*j*, *k*) -request: $\Delta(m,n) = b^{(2)}(m,n) / 2b(m,n)$;

$\eta(j,k) = \sum_{m=2}^{j-1}\sum_{n=1}^{N_m}\frac{R_{m-1}}{K_r - R_{m-1}}\rho(m,n)\Delta(m,n)$ - average time to receive applications from $\overline{2,j}$ groups interrupted by applications from groups $\overline{1,j-1}: \frac{K_{m-1}}{K_r - R_{m-1}}\rho(m,n)$ - probability of staying in queue (*m*, *n*) - applications, interrupted by applications from $1,m-1$ groups. This probability is determined by the formula (8), taking into account the intensity $\lambda(m,n)$ of the flow (*m*, *n*) -payments;

$\eta_0(j,k) = \sum_{m=1}^{j}\sum_{n=1}^{N_m} \frac{K_r\rho_0}{K_r - R_{m-1}}\rho(m,n)\Delta(m,n)$ - average time of subscription of applications from $\overline{1,j}$ groups interrupted by device refusals

$\frac{K_r\rho_0}{K_r - R_{m-1}}\rho(m,n)$ - the probability that the queue has (m, n)-applications, interrupted by the denial of the device. This probability is determined on the basis of (9) with account $\lambda(m,n)$;

$z_H(j,k)$ - average waiting time (j, k) - application, equal to the sum of the considered components without accounting σ_0;

$R_{j,k-1} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\rho(m,n) + \sum_{n=1}^{k-1}\rho(j,n)$.

Note that in each queue there can be no more than one application interrupted by application (j, k) with absolute priority by denial.

After simple transformations from equation (12) we obtain the following recurrence relation:

$$w_H(j,k) = \frac{1}{K_r - R_{j,k}}\left[K_r^2\rho_0\Delta_0 + \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\frac{1}{K_r - R_{m-1}}\times\right.$$
$$\times\rho(m,n)\Delta(m,n) + \sum_{m=1}^{j-1}\sum_{n=1}^{N_m} w_H(m,n)\rho(m,n) +$$
$$\left. + \sum_{n=1}^{k-1} w_H(j,n)\rho(j,n)\right] \qquad (13)$$

where $R_{j,k} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\rho(m,n) + \sum_{n=1}^{k}\rho(j,n)$.

To obtain a formula for explicit determination, we analyze the relation (13) for "pure" service disciplines with a relative and absolute priority.

For the discipline of service with a relative priority we receive:

- for the first flow $w_H(1,1) = \frac{K_r^3\rho_0\Delta_0 + \sum_{n=1}^{N_1}\rho(1,n)\Delta(1,n)}{K_r[K_r - \rho(1,1)]}$,

- for the second flow. $w_H(1,2) = \frac{K_r^3\rho_0\Delta_0 + \sum_{n=1}^{N_1}\rho(1,n)\Delta(1,n)}{[K_r - \rho(1,1)]\times[K_r - \rho(1,1) - \rho(1,2)]}$.

These formulas allow us to assume a general solution in the form:

$$w_H(1,k) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{n=1}^{N_1} \rho(1,n)\Delta(1,n)}{(K_r - R_{1,k-1})(K_r - R_{1,k})}, \qquad (14)$$

Where $R_{1,k-1} = \sum_{n=1}^{k-1} \rho(1,n), \quad R_{1,k} = \sum_{n=1}^{k} \rho(1,n)$.

For the discipline of service with absolute priority $(M = N, \quad N_m = 1$ for all $m = \overline{1, M})$ of the expression (13) we obtain:

• for the flow of the first group $w_H(1,1) = \frac{K_r^3 \rho_0 \Delta_0 + \rho(1,1)\Delta(1,1)}{K_r[K_r - \rho(1,1)]}$;

• for the flow of the second group $w_H(2,1) = \frac{K_r^3 \rho_0 \Delta_0 + \rho(1,1)\Delta(1,1) + \rho(2,1)\Delta(2,1)}{[K_r - \rho(1,1)][K_r - \rho(1,1) - \rho(2,1)]}$.

Then on the basis of these equalities we get the general expression:

$$w_H(j,1) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{m=1}^{j} \rho(m,1)\Delta(m,1)}{(K_r - R_{j-1,1})(K_r - R_{j,1})} \qquad (15)$$

where $R_{j-1,1} = \sum_{m=1}^{j-} \rho(m,1), \quad R_{j,1} = \sum_{m=1}^{j} \rho(m,1)$.

Analyzing the expression (14) and (15), it is easy to assume the general form of the formula for determining $w_H(j,k)$ for a mixed discipline of service:

$$w_H(j,k) = \frac{K_r^3 \rho_0 \Delta_0 + \sum_{m=1}^{j} \sum_{n=1}^{N_m} \rho(m,n)\Delta(m,n)}{(K_r - R_{j,k-1})(K_r - R_{j,k})} \qquad (16)$$

Substituting formula (16) in (13) and making simple transformations, we can verify the validity of this assumption.

By expressions (11) and (16) we calculate the required average time of stay (j, k) - request v (j, k) in the AS-I system

Similarly, as for the system type AC-I, formulas can be derived for determining the temporal characteristics for the remaining systems type AC-II, BD-I, BD-II.

6 Methods and analytical conditions for adapting the provision of resources to users of CC

Adaptation of the mixed service discipline with the **CC** model is to find the optimal breakdown of application flows by groups (levels) of absolute priority (φ^0), ie such a set of numbers $\{N_m\} m = \overline{1, M}$ at which the temporal characteristics of the **CC** model would provide equality according to problem (2):

$$\varphi^0 = opt\{N_1, N_2, \ldots, N_M / v^{(\varphi)}(m,n) \le v_{Д}(m,n), \varphi \in \Phi, M = \min\}, \quad (17)$$

and in accordance with problem (3) equality:

$$\varphi^0 = opt\{N_1, N_2, \ldots, N_M / C^{(\varphi)} = \min, \varphi \in \Phi, M = \min\}. \quad (18)$$

Since the number of application streams N is finite, the problem of finding the optimal breakdown φ^0 can be solved by a complete search of all possible breakdowns and choosing from them one that satisfies equations (17) and (18). However, this path for real-time **CC** is unacceptable, because the number of all possible breakdowns $\Phi = 2^{N-1}$ at large N is large and the implementation of the method of complete search requires significant time. Therefore, there is a need to develop such methods of adaptation that allow to obtain the optimal breakdown as a result of considering a limited number of grouping options.

To find the breakdown φ^0 that provides equality (17), a method is proposed, the essence of which is to alternate the requirements for the time of stay of applications in the system, starting with the first stream, by sequentially forming first the first, then the second, etc. groups of absolute priority. The adaptation process begins with a breakdown that corresponds to the discipline of service with a "pure" relative priority ($M = 1, N1 = N$). In this regard, the first of the breakdowns, in which the purpose of adaptation is fulfilled, is characterized by the minimum possible number of groups of absolute priority M, ie is optimal.

To find a breakdown φ^0 that satisfies equality (18), a method of adaptation is proposed, the essence of which is the purposeful formation of groups of absolute priority, starting with the latter, based on the analysis of the sign of increment of the average total cost of applications in the system $\Delta C^{(\varphi)}$. When forming the next group, the flow requests of the formed groups are excluded from consideration, because they do not affect the average time spent in the flow request system of the previous groups of absolute priority. The process of adaptation in this case begins with a breakdown that corresponds to the discipline of service with a "pure" absolute priority ($M = N, Nm = 1$), which also provides a minimum number of groups M in fulfilling the goal of adaptation.

Let's define $\Delta C^{(\varphi)}$. Assume that the previous q-breakdown has the form $N_1 = N_2 = \ldots = N_{l+1} = 1;\ N_{l+2} = N_j = P - l - 1$, where P is the number of application streams considered at the stage of formation of the next group with number j. When

$P = N - \sum_{m=1}^{j-1} N_m$ numbering groups from the latter . The following $\varphi-$ breakdown differs from the *q*-breakdown in that the application streams of the last two groups are combined into one $N_1 = N_2 = \ldots = N_l = 1, \quad N_{l+1} = N_j = P - l$.

When $\Delta C^{(q,\varphi)}$ the transition from *q*-breakdown to $\varphi-$ breakdown is defined as follows:

$$\Delta C^{(q,\varphi)} = C^{(\varphi)} - C^{(q)} = \sum_{i=1}^{N} \alpha_i \lambda_i \Delta \nu_i^{(q,\varphi)} \tag{19}$$

where $\Delta\nu_i^{(q,\varphi)} = \nu_i^{(\varphi)} - \nu_i^{(q)}, \quad i = \overline{1,N}$.

From formula (19) it follows that the $\varphi-$ breakdown is considered better compared to the *q*-breakdown, if $\Delta C^{(q,\varphi)} \le 0$. In this case $\Delta C^{(q,\varphi)} > 0$, the *q*-breakdown is preferred. The expression $\Delta C^{(q,\varphi)} = 0$ means that the $\varphi-$ breakdown by the criterion of the average total cost of applications in the system is not worse than the *q*-breakdown, but provides fewer groups of absolute priority M.

Let's calculate $\frac{\Delta C^{(q,\varphi)}}{b_i}$ on an example of system of type AC-I for which on the basis of expressions (11) and (16) it is possible to write down:

$$\nu_i^{(\varphi)} = \begin{cases} \dfrac{b_i}{K_r - R_{i-1}} + \dfrac{K_r^3 \rho_0 \Delta_0 + \sum_{r=1}^{P} \rho_r \Delta_r}{(K_r - R_{i-1})(K_r - R_l)}, & i = \overline{1,l}; \\ \dfrac{b_i}{K_r - R_l} + \dfrac{K_r^3 \rho_0 \Delta_0 + \sum_{r=1}^{P} \rho_r \Delta_r}{(K_r - R_{i-1})(K_r - R_i)}, & i = \overline{l+1, P}. \end{cases} \tag{20}$$

In the transition from *q*-breakdown to $\varphi-$ breakdown, the increase in the average residence time of applications in the system $\Delta\nu_i^{(q,\varphi)}$

$$\Delta \nu_i^{(q,\varphi)} = \begin{cases} 0, & i = \overline{1,l}; \\ \dfrac{\sum_{r=l+2}^{P} \rho_r \Delta_r}{(K_r - R_l)(K_r - R_{l+1})}, & i = l+1; \\ -\dfrac{b_i \rho_{l+1}}{(K_r - R_l)(K_r - R_{l+1})}, & i = \overline{l+2, P}. \end{cases} \tag{21}$$

Then write the increase $\Delta C^{(q,\varphi)}$ in the form

$$\Delta C^{(q,\varphi)} = \alpha_{l+1}\lambda_{l+1}\Delta\nu_{l+1}^{(q,\varphi)} + \sum_{i=l+2}^{P}\alpha_i\lambda_i\Delta\nu_i^{(q,\varphi)} =$$

$$= \frac{\rho_{l+1}}{2(K_r - R_l)(K_r - R_{l+1})}\sum_{i=l+2}^{P}\lambda_i b_i^{(2)}\left(\frac{\alpha_{l+1}}{b_{l+1}} - \frac{2}{1+\varphi_i^2}\frac{\alpha_i}{b_i}\right), \quad (22)$$

where $\psi_i = \sqrt{D[t_i]}/b_i$ - is the coefficient of variation of the service time of the applications of the *i*-th stream ($D[t_i]$ - variance of the service time). At indicative law service of applications of the *i*-th stream $\psi_i = 1$, and at deterministic service - $\psi_i = 0$.

Analysis of expression (22) shows that the feasibility of the transition from *q*-breakdown to $\varphi-$ breakdown is determined by the sign of the input::

$$\Delta C_l = \sum_{i=l+2}^{n}\lambda_i b_i^{(2)}\left(\frac{\alpha_{l+1}}{b_{l+1}} - \frac{2}{1+\psi_i^2}\frac{\alpha_i}{b_i}\right). \quad (23)$$

It follows from this equality:

1) if $\Delta C_l \le 0$ for all $l = \overline{0, N-2}$, the optimal is the discipline of service with a "pure" relative priority;

2) when $\Delta C_l > 0$ for all l the optimal is the discipline of service with "pure" absolute priority;

3) in the case of $\Delta C_l \le 0$ or $\Delta C_l > 0$ not all l optimal is a mixed discipline of service.

Thus, to determine the feasibility of the transition from *q*-breakdown to $\varphi-$ breakdown, it is sufficient by formula (23) to calculate ΔC_l and analyze the result. Changing the breakdown is appropriate if $\Delta C_l \le 0$.

Conclusions

1. Further development of the principles of creating adaptive infrastructures of cloud computing, able to dynamically adapt to user requirements and current features and changes in operating conditions. This scientific direction remains relevant and requires further research.

2. Cloud data centers are objects with a high level of randomness of the operation process, the main factors of which are: the probability of the flow of requests for computing resources; availability of necessary resources and randomness of time of their use by consumers; randomness of infrastructure failures and time of their elimination.

Due to the random nature of the computational process there are additional delays in processing information, violate the permissible restrictions on the time of its stay in the system (at the time of system response), which negatively affects the effectiveness of solving target tasks of users. This is relevant for real-time systems and, above all, for special information systems built using private clouds, and can be critical with limited computing resources.

3. It is possible to get rid of or reduce the impact of favorable phenomena on the functioning by introducing adaptation into the process of functioning of the infrastructure. In addition, the introduction of adaptation is associated with the need to maintain the CC in the optimal (efficient use of resources), and sometimes just a working condition, regardless of the many factors that bring the data center infrastructure out of the required target state. The purpose of adaptation can be to maximize revenue from customer service, eliminate system overload and maintain it in a stationary mode of operation.

4. The problem of adaptation can be solved by using the adaptive discipline (order) of providing computing resources to users. Unforeseen and uncontrolled changes in the environment and system inevitably change the optimal setting of the discipline, if such was implemented in the system. Therefore, systematic adjustments (adaptation) of the discipline are inevitable if you want to maintain the system in the optimal mode, regardless of changes in the environment and system. For adaptation, a dynamic adaptive mixed discipline with absolute relative priorities of providing computing resources to users of cloud computing was used, one of the options for creating the technology of which was considered by the author in [1,4]. The adaptation of the discipline consists of an optimal change in the number and position of the boundaries that divide the flow of user requests for resources into groups of absolute priority, within which the relative priority, ie in changing the number of groups and the number of flows in groups.

5. The development of analytical conditions for the adaptation of the provision of resources to users of cloud computing is performed on the basis of the analytical model. Analytical conditions allow to develop mechanisms and algorithms of adaptation of CC. These mechanisms and algorithms take into account the physical properties of CC, such as instantaneous elasticity (dynamic migration, allocation and release of resources for rapid scaling according to needs) and measurement services (management and optimization of resources using measurement tools). The moment of activation of the adaptation algorithm is determined by the control system of the CC in case of violation of acceptable limits on response time, change of controlled parameters of the system (for example, its total load) or system efficiency indicator above the limit values.

6. The stochastic nature of the main factors and the need to quantify mass processes based on probability theory determines the use of the analytical model of cloud computing as a multi-threaded and multi-priority queuing system with a mixed service discipline. The model takes into account probable failures and various features and has arbitrary distribution laws for some probable processes. Then as a mechanism of adaptation of CC it is possible and expedient to use the technology of dynamic adaptive mixed discipline of providing resources to users of CC.

References

1. Alexander Matov. Adaptation of cloud computing as optimization of the process of rendering services to users in the conditions of limited computing resources. // Selected Papers of the XIX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2019). CEUR Workshop Proceedings. - Vol-2577. - pp 210-221.

2. Matov O.Ya. Analytical models of multi-priority cloud data centers with a mixed discipline of service provision, taking into account the peculiarities of operation and possible failures. Registration, storage and data processing. 2019. T.21, №1 P.32 - 45.2 (in Ukraine)
3. Matov A.Ya., Shpilev VN, Komov AD et al. Organization of computational processes in ACS. Ed. Matov A.Ya. Kiev, 1989. - 200p. (in Russian).
4. Mokrov EV, Samuilov KE 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. Russ. https://cyberleninka.ru/article/n/model-sistemy-oblachnyh-vychisleniy-v-vide-sistemy-massovogo-obsluzhivaniya-s-neskolkimi-ocheredyami-i-s-gruppovym-postupleniem-zayavok
5. Bezzateev SV, Elina TN, Mylnikov VA 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. (in Russian).
6. Grusho AA, Zabezhailo MI, Zatsarinny AA Information flow monitoring and control in the cloud computing environment. Informatics and Applications, 2015. Vol. 9. No 4. P. 91–97. (in Russian).
7. Singh P., Dutta M., Aggarwal N. A review of task scheduling based on meta-heuristics approach in cloud computing // Knowledge and Information Systems. 2017. V. 52. N 1.
8. Gudkova I.A., Maslovskaya N.D. A probabilistic model for analyzing the delay in access to cloud computing infrastructure with a monitoring system // T-Comm: Telecommunications and Transport. 2014. No. 6. S. 13-15 (in Russian).
9. Tsai J.M., Hung S.W. A novel model of technology diffusion: system dynamics perspective for cloud computing. Journal of Engineering and Technology Management. 2014. V. 33. P. 4762.
10. Gorbunova A.V., Zaryadov I.S., Matyushenko S.I., Samuylov K.E., Shorgin S.Ya. Approximation of the response time of a cloud charge system. Computer science and its applications. 2015. (in Russian).