Mathematical Models of Cloud Computing with Absolute-relative Priorities of Providing of Computer Resources to Users in Conditions of Functioning Features and Failures

Aleksandr Matov[1]

[1] Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine
matov@ipri.kiev.ua

Abstract. Analytical models of cloud systems (CS) are developed as queueing systems with a mixed discipline of resource allocation. The models take into account failures and various functioning features and have arbitrary distribution laws for many stochastic processes. Such models for the CS are used for the first time. One of the main indicators of the effectiveness of the CS are indicators based on the evaluation of the time characteristics of these systems. Violation of the permissible time constraints, for example, the response time of the cloud system, adversely affects the efficiency of solving the target tasks of the user. This is particularly important for real-time systems and, first of all, for specific information systems built using private cloud systems.

General description of the models is as follows. The input of the cloud system, which implements a mixed queuing discipline (with relative and absolute priorities), receives N Poisson flows of requests for resources with corresponding N priorities. The duration of requests queuing of various flows has their own arbitrary distribution laws. Are quest with relative priority interrupted by requests with absolute priority, returns to the queue. Two disciplines of the resumption of A and Bqueuing are considered. Within the same priority, requests are processed on a first-come, first-served basis.

The CS fails according to the Poisson law, and is restored under an arbitrary law. During the recovery period, elements of adaptation to failures are used: requests of some flows to the queue are accepted and accumulated, while others are not accepted (the discipline of the queue replenishment, I and II, respectively,). The failure of service device can occur both during of its free state and during of the requests queuing. Two disciplines of queuing resumption after restoration C and D are considered. An interrupted request is processed from the point of its interruption. The combination of queuing resumption and replenishment of queue disciplines allows us to consider independent models of various types of systems that have the respective designation.

Different functioning features consist of various combinations of disciplines A, B, C, D, I and II.

Keywords: Cloud Computing, Mathematical Model, Interruption, Discipline of Providing Computing Resources, Adaptation and Optimization of Queuing Disciplines, Efficiency of Adaptation, Mixed Queuing Discipline

1 Introduction

The development of mathematical models of cloud computing or information systems created using clouds is an important area for identifying and improving their characteristics [1], [3], [10]-[12]. Cloud computing (CC) is an object with a high level of uncertainty in the functioning process, the main factors of which are [1], [3]:

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

-presence of the required CR and the accidental time of their use by customers;

-accidently failure of the infrastructure of the CC and the time of their elimination;

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

-necessity of optimal use of PR depending on the cost of time delay of customers ordered results of calculations and operating conditions;

-necessity of introduction of adaptation in the process of functioning CC in order to provide certain time characteristics for a number of clients and optimal use OP.

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 CR, 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 queueing theory. Then it is possible and appropriate to use the technology of the dynamic adaptive mixed discipline of providing CR (queuing) to users of the CC [1] as mechanisms of adaptation of the CS.

Analytical models for calculation of time characteristics are offered in the conditions of the features of the functioning of the CC using a mixed queuing discipline with absolutely relative priorities and taking into account failures. Models are based on works [2], [4] – [7], [9].

2 Description of the Model of Cloud Infrastructure Operation with Mixed Queuing Discipline and Adaptation to Failures.

Let the input of the CC system, in which the queuing discipline with a relatively absolute priority is implemented, arrive N Poisson flows of requests of intensity $\lambda(m,n) \quad (m=\overline{1,M}, \quad n=\overline{1,N_m})$. These flows are aligned with N priorities [2].

The duration of the maintenance of requests 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.

Are quests of priority (m, n) whose service is interrupted by requests from groups with $\overline{1, m-1},$ numbers is returned to the queue. Updating its service is possible either after servicing all interrupted requests (queuing discipline A), or after servicing all interrupted requests and all requests for accumulated flows, the m group with $\overline{(m,1),(m,n-1)}$ numbers (queuing discipline upgrade B) .

The service device (SD) fails in accordance with the Poisson law with the λ_0 parameter. The recovery period of the service device is a random variable that has an arbitrary distribution law Bo(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 $n_i, i = \overline{1, N}$, , 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 recovery period service device incoming requests can either accumulate in the queue (discipline of replenishment of queueI), or receive a refusal and leave the system (discipline of replenishment of queue II).

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

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

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

Let SD 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 service device requests ($(\rho(m,n) = \lambda(m,n) b(m,n)$ - loading of the service device(m, n)-requests), and $K_r = 1/(1+\rho_0)$ – the system readiness coefficient ($\rho_0 = \lambda_0 b_0$ – loading the service device with refusals).

It is necessary to determine the average $v(m,n)$ time spent in the system of requests of each (m, n)-request, namely the response time of the system CC.

3 Definition of Time Characteristics of a Model of a System of Type AS-I.

To determine the average time of requests in the system (time response systems) type AS-I use the known direct method [2].

Let some request (j, k) be a priority in the system. The average duration of this request 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):

$$v(j,k) = w(j,k) + b(j,k)\,. \tag{1}$$

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$u(j,k)$:

$$w(j,k) = w_H(j,k) + u(j,k)\,. \tag{2}$$

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

$$u(j,k) = u_3(j,k) + u_0(j,k)\,. \tag{3}$$

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

$$\Theta(j,k) = b(j,k) + u(j,k)\,. \tag{4}$$

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

During the queuing of *(j, k)* – request,$b(j,k)\Lambda_{j-1}$interruptions will occur on average, 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 requests.

As a result of these interruptions, *(j, k)*-request 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
$$R_{j-1} = \sum_{m=1}^{j-1}\sum_{n=1}^{N_m}\lambda(m,n)b(m,n)\,. \tag{5}$$

During this time, requests from groups $\overline{1, j-1}$will be received, which will lead to an increase in waiting time of (j, k)-requests for value$b(j,k)R_{j-1}^2$. In addition, the service of these requests will be accompanied by additional accumulation of requests of the same priorities, which are require of queuing of *(j, k)*-request before. This process is endless, with supplements to the waiting time of (j, k)-requests 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 interruptions of queuing of (j, k)-request:

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

In the mean time $T^{(1)}$, the of service 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 recovery period the service device again receives requests that continue to accumulate in the queue, then after the service device is restored, the average waiting time (j, k) -supply in the interrupted state will increase by

$$\mathrm{T}^{(2)} = \mathrm{T}^{(1)}\rho_0\frac{R_{j-1}}{1-R_{j-1}} = b(j,k)\rho_0\frac{R_{j-1}^2}{(1-R_{j-1})^2}. \tag{7}$$

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

The total time of all interruptions of queuing of *(j, k)*-request of $\overline{1, j-1}$ request groups, taking into account refusals of service device $u_3(j,k) = T^{(1)} + T^{(2)} + \ldots + T^{(\infty)}$. 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:

$$u_3(j,k) = b(j,k)\frac{R_{j-1}}{K_r - R_{j-1}}. \tag{8}$$

Similarly, the average waiting time of (j, k)-request is determined in the interrupted state due to service device refusals $u_0(j,k)$. The only difference is the beginning of reasoning. During the queuing of (j, k)-request, the service 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 service device renewal and priority service of requests with absolute priority from $\overline{1, j-1}$ group, the average waiting time of (j, k)-requests will increase by

$$b(j,k)\rho_0\frac{R_{j-1}}{1-R_{j-1}}.$$

During this time, the service device can again be denied, which additionally increases the waiting time (j, k)-request for value $b(j,k)\rho_0^2\frac{R_{j-1}}{1-R_{j-1}}$ etc.

In the final analysis, we get:

$$u_0(j,k) = b(j,k)\frac{K_r\rho_0}{K_r - R_{j-1}}. \tag{9}$$

Then the total average waiting time *(j, k)*-request in the interrupted state:

$$u(j,k) = b(j,k)\frac{R_{j-1} + K_r\rho_0}{K_r - R_{j-1}}. \tag{10}$$

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

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

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

1) the service device is restored

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

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

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

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

$$\begin{aligned} w_H(j,k) &= \sigma_0 + \sigma(j,k) + \eta(j,k) + \eta_0(j,k) + \\ &+ \sum_{m=1}^{j-1}\sum_{n=1}^{N_m} w_H(m,n)\rho(m,n) + \sum_{n=1}^{k} w_H(j,n)\rho(j,n) + \\ &+ [\sigma_0 + z_H(j,k)]\frac{R_{j,k-1}}{K_r - R_{j,k-1}} + z_H(j,k)\frac{K_r\rho_0}{K_r - R_{j,k-1}} \end{aligned}. \tag{12}$$

Here

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

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

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

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

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

$z_H(j,k)-$ average waiting time (j, k)-request, 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 request interrupted by requests with absolute priority or denial.

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

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

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" queueing disciplines with a relative and absolute priority.

For the queueing discipline with a relative priority we receive:

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

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

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

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

where $R_{1,k-1} = \sum\limits_{n=1}^{k-1}\rho(1,n), \quad R_{1,k} = \sum\limits_{n=1}^{k}\rho(1,n)$.

For the queueing discipline with absolute priority ($M = N, N_m = 1$ 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:

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

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

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 queueing discipline:

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

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.

The models take into account the physical properties of the CC, such as instantaneous elasticity (dynamic allocation and release of resources for fast scaling according to needs) and measuring service (management and optimization of resources with the help of measuring instruments).

References

1. A.Ya.Matov,"Optimization of the provision of computing resources with adaptive cloud infrastructure",*Data recording, storage and processing*,vol. 20, no.3, pp.83-90,(in Ukrainian), 2018.
2. A.Ya.Matov, V.N.Shpilev, A.D.Komov et al.,*Organization of computational processes in ACS*. Ed. A.Ya.Matov. Kiev, Ukraine, (in Russian), 1989.
3. A.Ya. Matov, I.O.Khramova, "Problems of mathematics and mathematical modeling of old ones are counted for the integration of information and analysis of the system and power management",*Data recording, storage and processing,* vol. 12,no.2, pp. 113-127, (in Ukrainian), 2010.
4. A. Ya.Matov,"Two modes of continuous completion of a queue when the instrument is restored in a servicing system with a relative priority",*Avtomat. i Telemekh.*, pp. 66-70, 1974.
5. A. Ya.Matov,"Two priority system with an unreliable device and period of servicing",*Engineering Cybernetics,* no. 10 (5), pp. 849-852, 1973.
6. A. Ya.Matov,"Two continuous queue disciplines for service-resumption period in a non-preemptive-priority queuing system",*Automation and remote control*, no. 35(4), pp. 575-578, 1974
7. A.Ya.Matov, N.F. Tishchenko,"Mathematical models of computing systems with priority denial of service",*Izv. Academy of Sciences of the USSR. Technical cybernetics*, no.3, pp. 190-194, (in Russian), 1980.
8. A.Ya.Matov, V.N.Shpilev,"The use of combined priorities to improve the efficiency of computing processes in the ACS",*Mechanization and automation of management*, no. 4, pp. 58-60, (in Russian), 1983.
9. A Ya. Matov, V. I. Zhluktenko, K. A. Chernous, N. F. Tishchenko, "Two continuous queuing disciplines in mixed priority systems", *Cybernetics and Systems Analysis*, no. 14(3), pp. 421-426, 1978.
10. E.V. Mokrov, K.E. Samuilov, "Cloud computing system model in the form of a queuing system with multiple queues and with a group of requests", (in Russian). [Online]. Available: https://cyberleninka.ru/article/n/model-sistemy-oblachnyh-vychisleniy-v-vide-

sistemy-massovogo-obsluzhivaniya-s-neskolkimi-ocheredyami-i-s-gruppovym-postupleniem-zayavok. Last accessed: November 14, 2018.

11. J.M.Tsai, and S.W.Hung,“A novel model of technology diffusion:system dynamics perspective for cloud computing”, J*ournal of Engineering and Technology Management*, vol. 33, pp. 47-62, 2014.
doi:10.1016/j.jengtecman.2014.02.003

12. P.Singh, M.Dutta, N.Aggarwal,“A review of task scheduling based on meta-heuristics approach in cloud computing”,*Knowledge and Information Systems*, vol. 52, nol.1, 2017. doi:10.1007/s10115-017-1044-2