Efficiency Of Servicing Heterogeneous Traffic When Allocating Cluster Nodes For Redundant Execution Of Latency-Critical Requests

Vladimir A. Bogatyrev [1,2], Stanislav V. Bogatyrev [1,3] and Anatoly V. Bogatyrev [3].

[1] *ITMO University, Kronverksky Pr. 49, bldg. A, Saint-Petersburg, 197101, Russia*
[2] *Saint-Petersburg State University of Aerospace Instrumentation, 67, Bolshaya Morskaia str. St Petersburg, Russia*
[3] *JSC NEO Saint Petersburg Competence Center, 1-Ya Sovetskaya, house 6 str. St. Petersburg, Russia*

Abstract
For computer systems of cluster architecture operating in real-time, criteria for the overall efficiency of servicing requests of traffic that is heterogeneous in terms of criticality to delays are proposed and justified. The possibilities of increasing the overall efficiency of servicing heterogeneous traffic are analyzed. The replication of waiting-critical requests and the division of cluster resources between requests with different limits of acceptable delays are considered as mechanisms for improving service efficiency. The number of cluster nodes (resources) allocated for servicing requests of different criticality to waiting is determined based on the ratio of the allowed waiting time for critical and other requests, the rigidity of the requirements for fulfilling the waiting time limits for them, as well as the ratio of the intensity of receipt of these requests. As a generalized indicator of the efficiency of servicing heterogeneous traffic by a computer cluster, the profit from the provision of information services is selected. An analytical model is proposed and the efficiency of redundant service options with possible separation of cluster nodes for solving requests of different criticality to waiting is determined. It is shown that there is a region of efficiency of reserved servicing of latency-critical requests when dividing cluster nodes into groups designed to service requests of different latency criticality. The expediency of setting and solving the optimization problem of determining the multiplicity of reserving various types of requests and dividing the cluster into groups is shown.

Keywords 1
Real-time, query replication, cluster, probability of timely maintenance.
Paper template, paper formatting, CEUR-WS

1. Introduction

For information and communication systems operating in real-time, including as part of cyber-physical systems [1-4], the key is to ensure high reliability and fault tolerance when fulfilling restrictions on the execution time of the request flow [5-9]. In some cases, to support the functional reliability of computer systems, it is necessary to ensure the continuity of the computing process and the timeliness of servicing requests in case of accidental failures and malicious influences [10-12].

Modern real-time info-communication systems are often characterized by heterogeneity of the flow of requests in terms of functionality and criticality to reliability and acceptable waiting time.

For computer systems, the efficiency of functioning is achieved when resources are consolidated as a result of their association into clusters. For real-time systems, in the case of traffic heterogeneity in the allowable waiting time, the reliability and timeliness of servicing a set of requests in a cluster can

Proceedings of VI International Scientific and Practical Conference Distance Learning Technologies (DLT–2021), September 20-22, 2021, Yalta, Crimea
EMAIL: vabogatyrev@corp.ifmo.ru (A. 1); stanislav@nspcc.ru (A. 2); anatoly@nspcc.ru (A. 3)
ORCID: 0000-0003-0213-0223 (A. 1); 0000-0003-0836-8515 (A. 2); 0000-0001-5447-7275 (A. 3)

be maintained by prioritizing them, balancing a load of nodes, replicating the most latency-critical requests, and adaptive allocation of cluster resources for servicing requests of different types [11-15].

The complexity of choosing mechanisms to ensure the timeliness of servicing heterogeneous traffic in terms of criticality to delays is associated with the justification of the criterion for the effectiveness of servicing a set of different types of requests. The criterion should contribute to the justification of the decision on the choice of the discipline of servicing requests of heterogeneous traffic, so that, based on a compromise, to ensure the overall efficiency of servicing requests of different criticality to delays.

The purpose of this article is to justify the choice of efficiency criteria and to study the possibilities of increasing the overall (vector) efficiency of servicing requests of heterogeneous traffic.

The replication of waiting-critical requests and the division of cluster resources between requests with different limits of acceptable delays are considered as mechanisms for improving service efficiency.

The number of cluster nodes (resources) allocated for servicing requests of different criticality to waiting is determined based on the ratio of the allowed waiting time for critical and other requests, the rigidity of the requirements for fulfilling the waiting time limits for them, as well as the ratio of the intensity of receipt of these requests.

2. Criteria for the Effectiveness of Reserved Services

Let's analyze the options for organizing redundant request servicing in a cluster containing *n* identical computer nodes (servers), with and without distributing its resources between requests of different criticality to waiting in the queue.

Consider heterogeneous traffic with the allocation of *z* types (streams) of requests for which the allowed waiting times in queues are t_1, t_2,…, t_z,, and their shares are equal to g_1, g_2, … , g_z, while

$$\sum_{i=1}^{z} g_i = 1.$$

The efficiency of real-time servicing of requests of the i-th stream is determined by the probability of not exceeding the expectation of the maximum allowable time ti [16-19].

Determining the efficiency of servicing a total heterogeneous flow is associated with finding a compromise to resolve the contradiction of achieving the efficiency of servicing all flows [16-19].

The efficiency of servicing a heterogeneous flow of requests in [19] is determined by the probability that waiting in the queue for requests of each of the z types does not exceed the maximum allowable time for each of them t_i,

$$P_м = \prod_{i=1}^{z} P_i \cdot ,$$

Such a criterion allows us to reduce the vector problem of evaluating the total efficiency by a multiplicative scalar criterion, but its application is limited to the case of the same importance of the probability of timely servicing of all types of traffic requests. The criterion does not allow taking into account the probabilities of requests of different criticality to the waiting time.

The scalar criterion allows us to take into account the influence of the probabilities of various types of requests for heterogeneous traffic on the final probability of timely servicing of the total flow

$$A = \sum_{i=1}^{z} g_i P_i ,$$

corresponding to the mathematical expectation of the probability that any type of request of an inhomogeneous stream will be executed in a timeless than the maximum allowable time for it t_i, it is possible to modify the last additive scalar criterion in the form

$$A_1 = \sum_{i=1}^{z} \alpha_i g_i P_i ,$$

where α_i -sets the importance of timely servicing of requests of the *i*-th stream

$$\sum_{i=1}^{z} \alpha_i = 1, \ \sum_{i=1}^{z} g_i = 1.$$

The profit from the provision of information services can be chosen as a generalized indicator of the efficiency of the system's functioning. So, if the profit from timely servicing of the request of the *j*-th stream is c$_j$, and the penalties for late servicing are s_j, then the mathematical expectation of profit from executing one request and receiving profit per unit of time from providing information services (profit intensity) of the total flow will be [18]

$$C = \sum_{j=1}^{N} g_j \left(R_j c_j - (1 - R_j) s_j \right),$$

$$D = \Lambda \sum_{j=1}^{N} g_j \left(R_j c_j - (1 - R_j) s_j \right),$$

where Λ is the total intensity of the request flow.

In the simplest case, with absolute reliability and error-free operation of the system, the R_j indicator can be defined as the probability of timely servicing of j-type requests (with an acceptable waiting time for t_j requests.

3. Evaluation of the Efficiency of Reserving Wait-Critical Requests in a Cluster without Splitting Nodes into Groups

The cluster is represented by a set of n single-channel queuing systems with infinite queues of the M/M/1 type [20, 21]. Under such assumptions, the probability that the delay in the queue of an unserved request of the *i*-th thread is less than the maximum allowable time t_i for the average execution time of requests of all *z* types equal to v is calculated as

$$P_i = 1 - \frac{\Lambda \ v}{n} e^{\left(\frac{\Lambda}{n} - \frac{1}{v} \right) t_i} \qquad (1)$$

where ...

$$\Lambda_0 = \sum_{i=1}^{z} k_i g_i \Lambda .$$

Let's distinguish two gradations of requests in the heterogeneous flow according to criticality to expectation - critical and non-critical. Critical requests are duplicated, non-critical requests are not replicated. The share of critical requests *g*.

The profit per unit of time from the provision of information services is calculated as

$$C = \Lambda \left\{ g \left[c_1 P_1 - s_1 (1 - P_1) \right] + (1 - g) \left[c_2 P_2 - s_2 (1 - P_2) \right] \right\}.$$

If there is no replication of requests, then P_1 and P_2 are determined by (1).

If only critical requests are duplicated, then

$$P_1 = 1 - \left[\frac{\Lambda(1+2g)v}{n} e^{\left(\frac{\Lambda(1+2g)}{n} - \frac{1}{v} \right) t_1} \right]^2,$$

$$P_2 = 1 - \frac{\Lambda(1+2g)v}{n} e^{\left(\frac{\Lambda(1+2g)}{n} - \frac{1}{v} \right) t_2} .$$

The dependence of the intensity of the profit received when servicing requests of the total flows on their intensity Λ without dividing *n*=12 cluster nodes into groups is shown in Fig.1. The calculations were performed at the average query execution time *v*=0.01 s and the maximum allowable waiting time for critical queries t_1=0.01 s and non-critical queries t_2=0.1 s. The values of profit and penalties for timely and untimely servicing of requests of various criticality to waiting are set as c_1=2, y. e. c_2=0.2 y.

e., s_1=-3, y. e. s_2=-0.02 y. e. Curves 1-3 correspond to the duplication of critical requests with the proportions of critical requests g =0.7, 0.5, 0.2. It can be seen from the graphs that there are areas of efficiency of duplicated requests that are prone to waiting. At the same time, as the intensity of the total flow of requests increases, this efficiency first increases, and then begins to fall. At the same time, there is a value of Λ, above which duplication of critical requests becomes impractical. Indeed, the replication of waiting-critical requests causes an increase in the overall cluster load and a decrease in the probability of timely servicing of non-critical requests, which can lead to a decrease in the intensity of profit from the provision of information services.

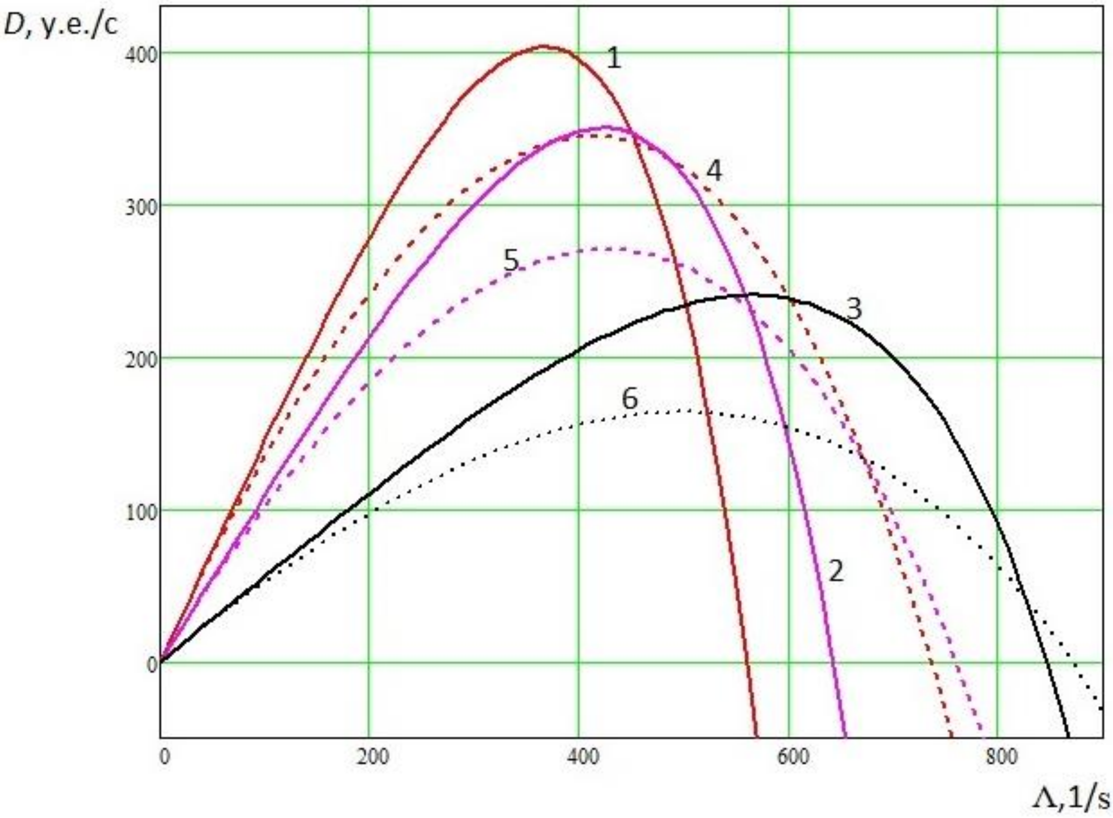

Figure 1: Dependence of the profit intensity when executing total flow requests without dividing the cluster nodes into groups

4. Maintenance with the Separation of Cluster Nodes Between Requests of Different Criticality to Waiting

Let's analyze the efficiency of dividing the cluster into two groups, including n_1 and n-n_1 nodes, the first of which is allocated for servicing critical waiting requests, and the second for the remaining requests.

The probability of timely execution of critical and non-critical waiting requests without their replication will be:

$$P_1 = 1 - \frac{\Lambda g v_1}{n_1} e^{\left(\frac{\Lambda g}{n_1} - \frac{1}{v_1}\right) t_1},$$

$$P_2 = 1 - \frac{(1-g)\Lambda v_2}{n - n_1} e^{\left(\frac{\Lambda(1-g)}{n-n_1} - \frac{1}{v_2}\right) t_2}.$$

The profit per unit of time from the provision of information services is calculated as

$$C = \Lambda\left\{g\left[c_1P_1 - s_1\left(1-P_1\right)\right] + \left(1-g\right)\left[c_2P_2 - s_2\left(1-P_2\right)\right]\right\}.$$

The dependence of the profit intensity when servicing total flow requests on their intensity Λ with and without dividing *n*=12 cluster nodes into groups is shown in Fig.2. Fig. 2 *a* reflects the case when the proportion of critical requests is *g*=0.7 and Fig. 2 *b* when *g*=0.1. The calculations are performed at *g*=0.7, the average query execution time is *v*=0.01 s, and the maximum allowable waiting time for critical requests is t_1=0.01 s, and for non-critical t_2=0.05 s. In this case, c_1=2 y. e., c_2=0.2 y. e., s_1=-3, y. e. s_2=-0.02 y. e. Curve 1 corresponds to a cluster without dividing nodes into groups. Curves 1-6 correspond to the allocation for servicing the waiting-critical requests n_1=2, 3, 4, 7, 8 nodes. From the presented dependencies, it is clear that there is an area for which it is advisable to divide the cluster resources into groups. With an increase in the intensity of the total flow of requests, the growth and then the fall in profits from the provision of information services is visible first. The presented graphs show the significance of the influence of the number of nodes allocated to groups on the overall efficiency of servicing heterogeneous traffic. Moreover, some options for splitting the cluster may not be effective compared to the basic option without dividing the cluster into groups.

The dependence of the intensity of the profit received when servicing requests of the total flow on the number of nodes allocated for servicing latency-critical requests is shown in Fig. 3. The calculation was carried out for the previously specified initial data with the proportion of requests critical to delays g=0.7, t1=0.01 s, t2=0.05 s. Curves 1-5 correspond to the intensities of the total flow Λ=600, 500, 400, 350, 200 1/s. To compare the efficiency of group allocation, Figure 2 shows the values of the profit intensity without dividing the cluster into groups, while lines 6 and 7 correspond to Λ=600 1/s and Λ= 200 1/s. These dependencies confirm the effectiveness of dividing the cluster into groups when there is an optimal variant of allocating cluster resources that allows you to get maximum profit from the provision of information services.

The presented dependencies allow us to conclude that reserving the most critical requests for delays in queues can significantly increase the efficiency of servicing not only these requests but also the total flow as a whole. It is shown that varying the number of cluster nodes included in the group for servicing critical requests makes it possible to increase the efficiency of providing services both critical to delays in request queues and the entire total flow. Thus, the expediency of setting and solving the optimization problem of determining the multiplicity of reserving requests and dividing the cluster into groups to maximize the probability of timely execution of requests of a heterogeneous flow is shown.

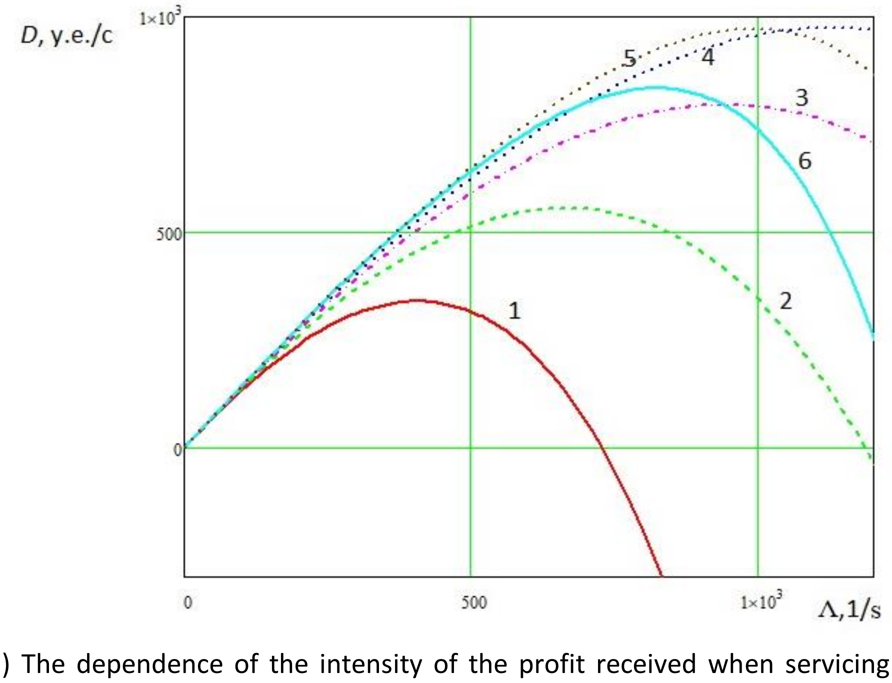

Figure 2: a) The dependence of the intensity of the profit received when servicing the total flow requests on their intensity with and without dividing the cluster nodes into groups

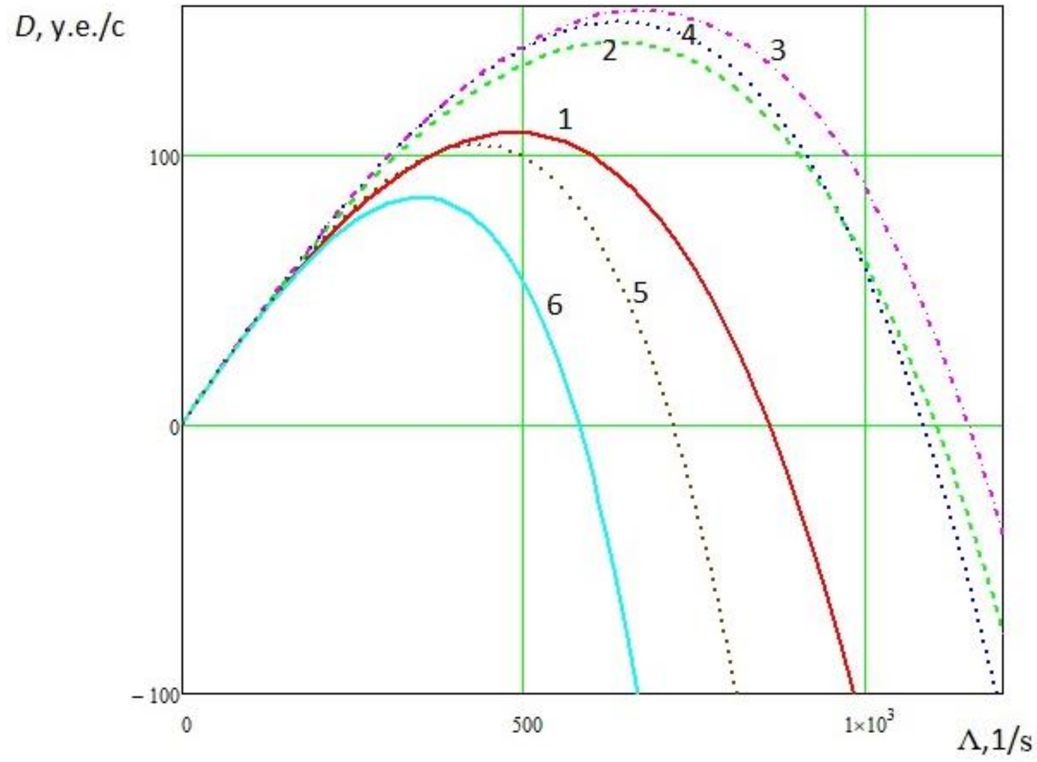

Figure 3: b) The dependence of the intensity of the profit received when servicing the total flow requests on their intensity with and without dividing the cluster nodes into groups

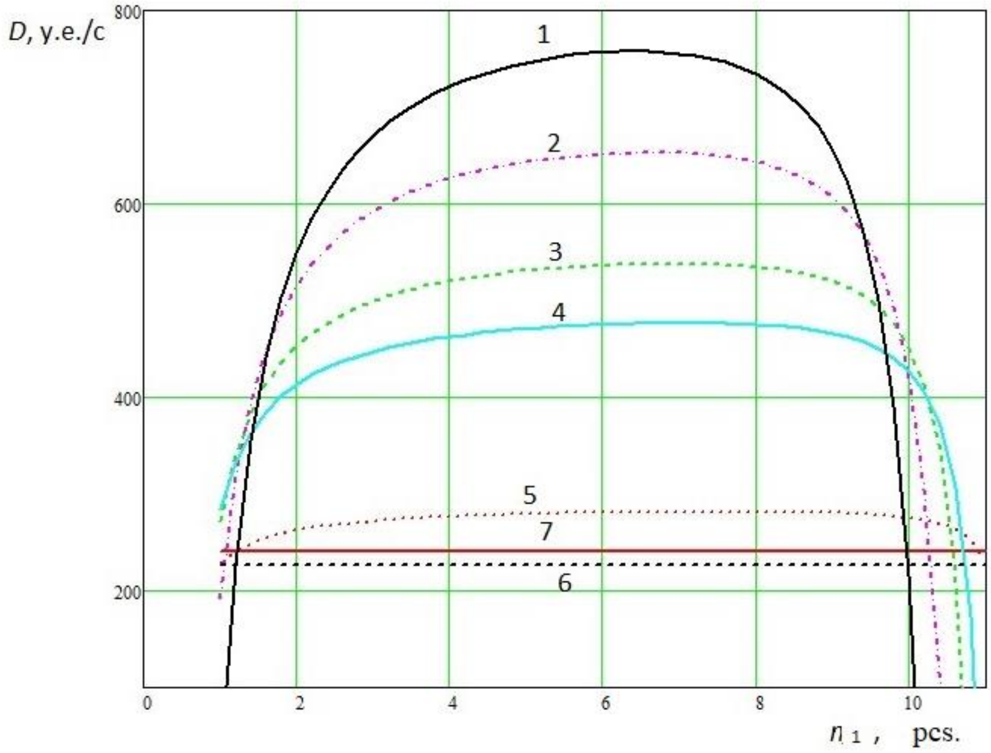

Figure 4: The dependence of the intensity of the profit received when servicing requests of the total flow on the number of nodes allocated for servicing latency-critical requests

5. Conclusions

For computer systems of cluster architecture operating in real-time, efficiency criteria are proposed and the possibilities of increasing the overall efficiency of servicing requests of heterogeneous traffic are shown based on the replication of waiting-critical requests and the allocation of a group of nodes for their servicing of certain types of requests for acceptable waiting delays.

An analytical model is proposed and the efficiency of redundant service options is determined with the possible allocation of cluster nodes to solve the most critical waiting requests in queues.

It is shown that there is a region of efficiency of reserved servicing of latency-critical requests when dividing cluster nodes into groups designed to service requests of different latency criticality.

References

[1] I. Koren, Fault-tolerant systems. Morgan Kaufmann publications, San Francisco 2009 378 p.

[2] H. Aysan, Fault-tolerance strategies and probabilistic guarantees for real-time systems Mälardalen University, Västerås, Sweden. 2012. 190 p.

[3] T. Astakhova, N.Verzun, M. Kolbanev, A. Shamin, A model for estimating energy consumption seen when nodes of ubiquitous sensor networks communicate information to each other. In Proceedings of the 10th Majorov International Conference on Software Engineering and Computer Systems, Saint Petersburg, Russia, December 20-21 (2018).

[4] D.A. Zakoldaev, A.G. Korobeynikov, A.V. Shukalov, I.O. Zharinov, O.O. Zharinov, Industry 4.0 vs Industry 3.0: the role of personnel in production//IOP Conference Series: Materials Science and Engineering, 2020, Vol. 734, No. 1, pp. 012048. DOI 10.1088/1757-899X/734/1/012048

[5] M. Lanctot, A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning / M. Lanctot et al. // Advances in Neural Information Processing Systems. – 2017. – P. 4190-4203.

[6] M. Bennis, M Debbah,.; Poor, H.V. Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale. *Proc. IEEE* 2018, *106*, 1834–1853. DOI**:** 10.1109/JPROC.2018.2867029.

[7] H.Ji, S. Park, J. Yeo, Y. Kim, J. Lee, B. Shim, Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects. *IEEE Wirel. Commun.* 2018, *25*, 124–130. DOI:10.1109/MWC.2018.1700294.

[8] J. Sachs, G. Wikström, T. Dudda,;, R. Baldemair.; K. Kittichokechai, 5G Radio Network Design for Ultra-Reliable Low-Latency Communication. *IEEE Netw.* 2018, *32*, 24–31. DOI:10.1109/MNET.2018.1700232

[9] S. Samarasinghe, Neural Networks for Applied Sciences and engineering: from Fundamentals to Complex Pattern Recognition / S. Samarasinghe. – Boston: Auerbach publications, 2016. – 570 p.

[10] M. Siddiqi1, H. Yu, J. Joung, 5G Ultra-Reliable Low-Latency Communication Implementation Challenges and Operational Issues with IoT Devices Electronics 2019, 8, 981; doi:10.3390/electronics8090981

[11] I S. Kim, Y. Choi, Constraint-aware VM placement in heterogeneous computing clusters. Cluster Comput 23, 71–85 (2020). https://doi.org/10.1007/s10586-019-02966-6.

[12] B. Sovetov, T. Tatarnikova, V. Cehanovsky, a Detection system for threats of the presence of the hazardous substance in the environment. Proceedings of 2019 22nd International Conference on Soft Computing and Measurements, SCM 2019 (2019) 121-124. DOI: 10.1109/SCM.2019.8903771.

[13] E.D.Poymanova, T. M. Tatarnikova, Models and Methods for Studying Network Traffic. In 2018 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF), pp. 1-5 (2018). doi:10.1109 / WECONF.2018.8604470

[14] S.Sahni, V.Varma, A hybrid approach to live migration of virtual machines // Proc. IEEE Int. Conf. on Cloud Computing for Emerging Markets (CCEM 2012). Bangalore, India, 2012. P. 12–16. DOI: 10.1109/CCEM.2012.6354587

[15] V. Malik, C.R. Barde, Live migration of virtual machines in cloud environment using prediction of CPU usage // International Journal of Computer Applications. 2015. V. 117 N 23. P. 1–5. DOI: 10.5120/20691-3604

[16] V.A. Bogatyrev, A.V. Bogatyrev, S.V. Bogatyrev, Redundant Servicing of a Flow of Heterogeneous Requests Critical to the Total Waiting Time During the Multi-path Passage of a

Sequence of Info-Communication Nodes. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. Vol. 12563. pp. 100-112. DOI 10.1088/1742-6596/1864/1/012094.

[17] V.A. Bogatyrev, S.V. Bogatyrev A.N. Derkach, Timeliness of the Reserved Maintenance by Duplicated Computers of Heterogeneous Delay-Critical Stream. CEUR Workshop Proceedings. 2019. Vol. 2522. pp. 26-36.

[18] V.A. Bogatyrev, S.V. Bogatyrev, A.V. Bogatyrev, Replication of requests when dividing cluster nodes between threads of different criticality to delays in queues CEUR Workshop Proceedingsthis link is disabled, 2020, 2893

[19] V.A. Bogatyrev, S.V. Bogatyrev, A.V. Bogatyrev, Redundant multi-path service of a flow heterogeneous in delay criticality with defined node passage paths // Journal of Physics: Conference Series, Volume 1864, 13th Multiconference on Control Problems (MCCP 2020) 6-8 October 2020, Saint Petersburg, Russia 2021 J. Phys.: Conf. Ser. 1864 012094 - 2021, Vol. 1864, 012094, No. 1, pp. 012094

[20] Kleinrock, L. Queueing Systems: Volume I. Theory. New York: Wiley Interscience. 1975 p. 417.

[21] Kleinrock, L. Queueing Systems: Volume II. Computer Applications. New York: Wiley Interscience. 1976 p. 576.