Inter-cluster Redistribution of Requests with Redundancy Depending on their Criticality to Delays

Vladimir Bogatyrev [a,b], Stanislav Bogatyrev [b] and Anatoly Bogatyrev [b]

[a] *University Saint-Petersburg State University of Aerospace Instrumentation, 67, Bolshaya Morskaia str., Saint-Petersburg, 190000, Russia*
[b] *JSC NEO Saint Petersburg Competence Center, 6, 1st Sovetskaya Str., Saint- Petersburg, 191036, Russia*

Abstract

A distributed computer system containing clusters connected through a network is considered, each of which receives a separate stream of requests. Possibilities, based on the redistribution of requests between clusters, are investigated to increase the probability of timely execution of a non-uniform flow of requests of different criticality to service delays. The efficiency of inter-cluster redistribution of requests through the network with replication of the most latency-critical requests is shown.

When replicating queries, the condition for timeliness of computations is to service at least one replica of a query in a time less than the maximum allowable time. An indicator of the efficiency of the systems under consideration is the readiness to timely fulfill all requests of a non-uniform flow in terms of criticality.

An analytical model is proposed and the efficiency of increasing the availability and timely servicing of a flow that is heterogeneous in terms of criticality to latency requests is shown on the basis of redistribution of requests between clusters through the network with possible replication of the most latency-critical requests. The influence on the efficiency of a multicluster computer system, the proportions of requests redistributed between clusters, is established, and the existence of their optimal values is shown, which make it possible to achieve the maximum readiness of a multicluster system for timely execution of a heterogeneous flow of requests.

Keywords

Cluster, Allowable Latency, Real Time, Heterogeneous Flow

1. Introduction

High demands are placed on distributed computer systems in real time, in terms of reliability [1-3], fault tolerance, performance and acceptable service delays [4-6]. The problems of increasing fault tolerance, reliability, timeliness and continuity of calculations are especially acute for distributed cyberphysical systems [7-12].

High indicators of reliability and timeliness of computations in distributed systems uniting many clusters are achievable as a result of the consolidation and sharing of resources [13-16], including when redistributing requests between clusters [17-20]. Redistribution of requests allows you to adapt to resource outages and traffic changes. For real-time systems that are critical to service delays, redistribution of requests can increase the reliability and the likelihood of their timely execution.

Additional possibilities for ensuring the reliability and timeliness of the execution of requests in infocommunication systems are provided by their redundant service [17-18], in which copies are created for critical to the waiting requests and distributed for execution to different computer nodes. If the input stream of requests is heterogeneous, the multiplicity of reservation of requests can be assigned differently, taking into account their criticality to the permissible waiting time [19-21].

Proceedings of the 12th Majorov International Conference on Software Engineering and Computer Systems, December 10-11, 2020, Online & Saint Petersburg, Russia
EMAIL: vladimir.bogatyrev@gmail.com (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)

Redistribution of replicas of requests through the network can be based on multipath routing technologies [17-18] and multipath redundant transmissions] [21-22].

The purpose of this article is to study the possibilities of increasing the reliability and the probability of timely servicing of a flow that is heterogeneous in terms of criticality to the latency of requests based on the redistribution of requests between clusters through the network with the possible replication of the most latency-critical requests.

When replicating requests, the condition for a timely service request is to service at least one replica in a time less than the maximum allowable time.

The efficiency of redundant servicing of heterogeneous traffic can be determined by the probability that all types of requests will be serviced taking into account the constraints on admissible queuing delays specified for them.

The effect of increasing the likelihood of timely service is achieved as a result of redistributing requests with varying their redundancy rate depending on the constraints on the allowable waiting time in queues.

The proposed research is aimed at resolving a technical contradiction. arising during the reservation of a part of the requests, which leads both to an increase in the probability of timely execution of at least one replica of the reserved requests, and to an increase in the overall load of the system, which negatively affects the increase in this probability.

The investigated effect of increasing the probability of timely servicing is achieved when replication of the most critical to waiting requests does not lead to a violation of the conditions of timely execution of all other requests.

The article solves the problem of analyzing the influence of redistribution of requests during their replication on increasing the probability of timely servicing of all types of flow requests that differ in the allowable waiting time (criticality).

For real-time systems, when the system operability condition is determined by the requirement of timely servicing of a non-uniform flow of requests, the proposed redundant redistribution of the most critical to waiting requests leads to an increase in the availability (functional reliability) of the system.

The novelty of the proposed research consists in analyzing the possibilities and efficiency of increasing the probability of timely execution of a non-uniform flow and the reliability (availability) of multicluster systems based on redistribution of the flow of requests between clusters with replication of the most latency-critical requests.

The proposed research is aimed at solving the practical problem of increasing the reliability of systems that require timely servicing of a heterogeneous flow of requests, different criticality to the permissible waiting time.

It should be noted that in practice, under conditions of traffic variability during operation, the investigated redundant redistribution of requests should be based on monitoring traffic [23, 24], as well as the load and availability of system resources.

2. Probability of timely servicing of a non-uniform flow with redistribution of requests

Consider a distributed computer system containing clusters connected through a network, each of which receives a separate stream of requests. Adaptation of the system to server failures and overload is achieved by redistributing requests between clusters.

The probability that the waiting time for a request in a node (server) of a cluster, rep-resented by a single-channel queuing system of the M / M / 1 type [25, 26], is less than the maximum allowable time t is calculated as

$$P(\Lambda, t) = 1 - \Lambda v e^{\left(\Lambda - \frac{1}{v}\right)t},$$

where is the intensity of requests, arriving at the node, v is the average query execution time.

The probability of timely execution of queries, taking into account their possible redistribution between clusters for queries arriving in the first and second clusters, we find, respectively, as

$$P_1 = \delta_1 g\left(1-\Lambda_1 v e^{\left(\Lambda_1-\frac{1}{v}\right)t_1}\right)+\delta_2(1-g)\left(1-\Lambda_2 v e^{\left(\Lambda_2-\frac{1}{v}\right)(t_1-D)}\right), \quad (1)$$

$$P_2 = \delta_1(1-d)\left(1-\Lambda_1 v e^{\left(\Lambda_1-\frac{1}{v}\right)(t_2-D)}\right)+\delta_2 d\left(1-\Lambda_2 v e^{\left(\Lambda_2-\frac{1}{v}\right)t_2}\right),$$

where g and d are the proportions of requests of the first and second flows executed in the first and second clusters without redistribution through the network, D is the delay in transmitting a request through the network, t_1 and t_2 are the maximum allowable waiting times for requests of the first and second flows, the intensities of which, respectively, Λ and $\Lambda\beta$.

The intensity of requests served in the first and second clusters containing n and m nodes, taking into account the redistribution, are, respectively

$$\Lambda_1 = \frac{\left[g+\beta(1-d)\right]\Lambda}{n},$$

$$\Lambda_2 = \frac{\left[(1-g)+\beta d\right]\Lambda}{m}.$$

Condition for stationarity of the service mode

$$\delta_1 = 1, if\ \Lambda_1 v < 1, \text{ else } \delta_1 = 0,$$

$$\delta_2 = 1, if\ \Lambda_2 v < 1, \text{ else } \delta_1 = 0.$$

In the simplest case, when realizing redistribution through one communication node, represented as a single-channel queuing system of the M / M / 1 type [25, 26]

$$D = \frac{v_s}{1-\Lambda_s v_s},$$

where v_s is the average time of transmission of a request through the network, the intensity of the flow of requests redistributed through the network

$$\Lambda_s = \left[(1-g)+(1-d)\beta\right]\Lambda.$$

If there are n_s nodes in the network involved in the redistribution of requests, each of them receives requests with an intensity

$$\Lambda_s = \left[(1-g)+(1-d)\beta\right]\frac{\Lambda}{n_s}.$$

The probability of timely execution of requests of both threads is defined as

$$R = P_1 P_2 \ . \quad (2)$$

The results of calculating the probability of timely execution of requests of the first and second flows R are shown in Fig. 2, where curves 1-6 correspond to d = 0.1; 0.3; 0.5; 0.7; 0.9. Figure 1 a) represents the case t_1 = 0.4 s and t_2 = 0.2 s, and Figure 1 b case - t_1 = 0.4 s and t_2 = 0.4 s. The calculations were performed at v = 0.1 s, the intensity of the first and second flow of requests Λ = 50 1/s (β = 1) and the delay in the transmission of requests through the network D = 0.01 s.

The presented dependencies confirm that inter-cluster re-distribution of requests significantly affects the probability of timely servicing of requests, and allows you to maximize the probability of timely execution of requests of different criticality to the time of acceptable waiting. Moreover, the optimal shares of requests redistributed through the network significantly depend on the ratio of the criticality of requests of different flows to the delays, which is confirmed by a comparison of the graphs in Figure 1 a and b.

a)

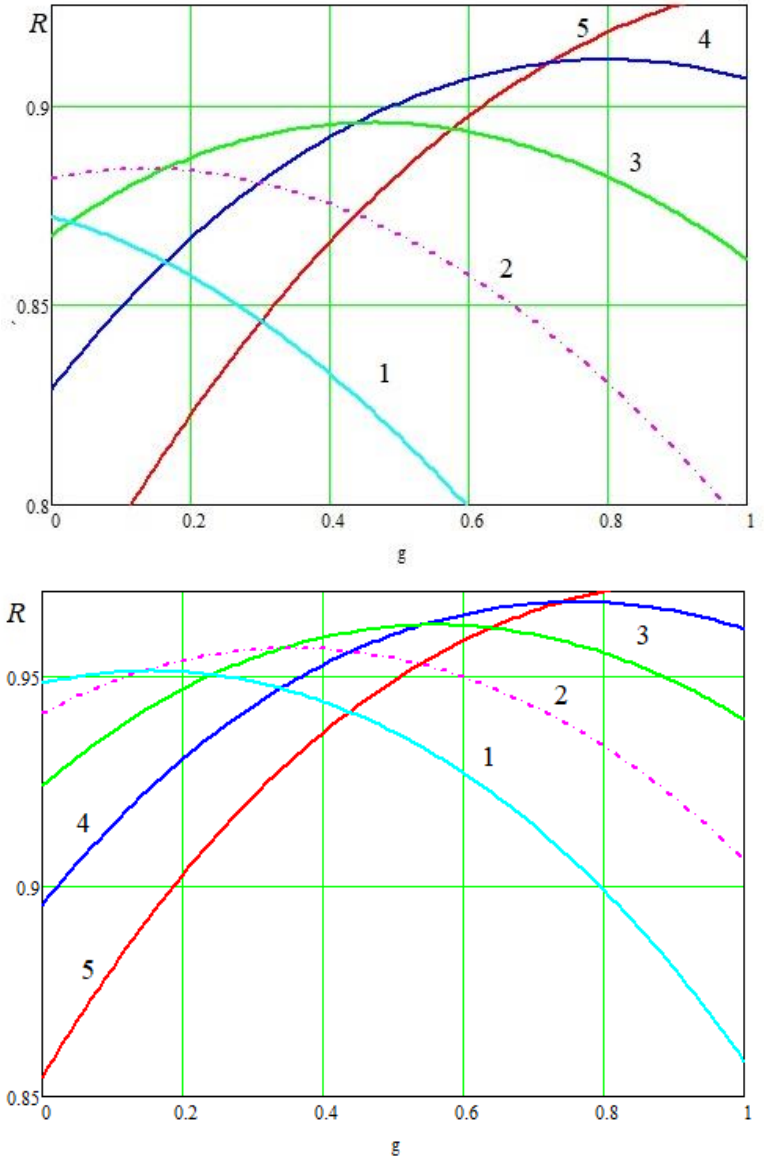


Figure 1. Probabilities of timely execution of requests of the first and second R threads (simultaneously)

3. Probability of timely servicing of a heterogeneous flow during redistribution with replication of requests

Let us consider the systems when two streams of requests are allocated with replication of requests of the second stream when the replicas are directed for execution to different clusters. Requests from the first thread are redistributed to the second cluster with probability g.

The probability of timely execution of the requests of the first thread is determined by the formula (1) at

$$\Lambda_1 = \Lambda(g+\beta)/n,$$

$$\Lambda_2 = \Lambda\left[(1-g)+\beta\right]/m.$$

The probability of timely execution of replicas of the second stream queries executed in the first and second cluster. defined as

$$P_{21} = \delta_1 \left(1 - \Lambda_1 v e^{\left(\Lambda_1 - \frac{1}{v}\right)(t_2 - D)} \right),$$

$$P_{22} = \delta_2 \left(1 - \Lambda_2 v e^{\left(\Lambda_2 - \frac{1}{v}\right) t_2} \right).$$

The probability of timely execution of at least one of the replica requests of the second thread is calculated as

$$P_2 = 1 - (1 - P_{21})(1 - P_{22}) \ . \qquad (3)$$

The probability of timely execution of requests of both threads is determined by formula (2) when calculating P_2 by formula (3).

The results of assessing the probability of timely execution of requests from the first and second R flows with possible replication of requests from the second flow are shown in Fig. 2, where curves 1-5 correspond to the redistribution of requests without replication at d = 0.1; 0.3; 0.5; 0.7; 0.9, and curve 6 represents the variant with replication of requests of the second stream when they are directed to different clusters. The calculations were performed at v = 0.1 s and the intensity of the first and second flow of requests Λ = 70 1/s (β = 1), at t_1 = 0.6 s, t_2 = 0.2 s and D = 0.01 s with the redistribution of the first flow between clusters.

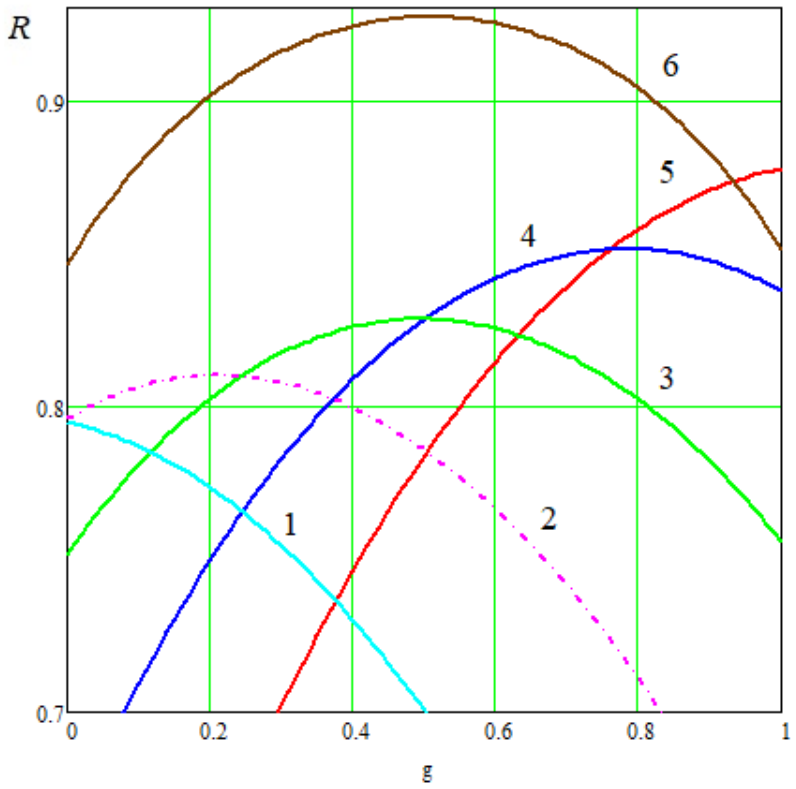


Figure 2. Probabilities of timely execution of requests of the first and second R threads with possible replication of requests of the second thread

4. System readiness for timely execution of requests

The readiness of the system is determined by the probability of finding it at the moment of receipt of the request in one of the operable states, which provide the necessary quality of performing the required tasks. In computing systems, one of the conditions for stable functioning (operability) is the ability to service requests in a stationary mode. To compensate for the negative impact of the accumulation of failures on the performance of a multi-cluster system, the redistribution of requests between clusters in order to ensure a stationary mode of service during the consolidation of resources of all clusters allows.

The probability of operability of a multicluster system, providing for redistribution of requests between clusters through the network, is defined as

$$P = (1 - R_L)R_1 + R_L R_2, \tag{4}$$

where R_L is the probability of the network operability R_1 and R_2 are the probability of the operability of the cluster system with and without redistribution of requests through the network.

Redistribution is possible provided that at least one node is operational in each cluster, which is used to redistribute (transmit and receive) requests through the network.

$$R_1 = \left[\sum_{i=1}^{n} \delta_i C_n^i p^i (1-p)^{n-i}\right]\left[\sum_{j=1}^{n_2} \delta_j C_m^i p^i (1-p)^{m-j}\right], \tag{5}$$

$$R_2 = \left[\sum_{i=1}^{n}\sum_{j=1}^{m} \delta_{ij} C_n^i C_m^i p^{i+j} (1-p)^{n+m-i-j}\right], \tag{6}$$

where p is the probability of a node's operability (for recoverable systems, the availability factor of a node), Λ, Λβ is the intensity of the flow of requests to one and the second cluster, v is the execution time of a request in the cluster server; m and n are the number of nodes (servers) in clusters, $\delta_i, \delta_j, \delta_{ij}$ the conditions for the combination of efficient nodes of both clusters to ensure the probability of timely servicing of requests with probabilities not less than the maximum permissible value p_{01} and p_{02}, respectively, for the first and second streams, for which the maximum allowable time for servicing requests is specified as t_1 and t_2.

The condition for fulfilling the requirements of timeliness of service for requests of the first and second streams δ_i, δ_j without their redistribution is given as

$$\delta_i = \begin{cases} 1, \ if \left(1 - \frac{\Lambda v}{i}\exp(-t_1(\frac{1}{v} - \frac{\Lambda}{i}))\right) \geq p_{01}, \\ 0, \ if \left(1 - \frac{\Lambda v}{i}\exp(-t(\frac{1}{v} - \frac{\Lambda}{i}))\right) < p_{01}, \end{cases}$$

$$\delta_j = \begin{cases} 1, \ if \left(1 - \frac{\Lambda v\beta}{j}\exp(-t_2(\frac{1}{v} - \frac{\Lambda\beta}{j}))\right) \geq p_{02}, \\ 0, \ if \left(1 - \frac{\Lambda v\beta}{j}\exp(-t_2(\frac{1}{v} - \frac{\Lambda\beta}{j}))\right) < p_{02}. \end{cases}$$

The condition of timely service of requests with their possible redistribution between clusters is specified as:

$$\delta_{ij} = 1, \text{ if } \left(b_{ij} \geq p_{01}\right) \wedge \left(a_{ij} \geq p_{02}\right), \text{ else } \delta_{ij} = 0,$$

moreover, the probabilities that the delays in queues do not exceed the limit of the admissible time t_1, t_2 for requests of the first and second flows (formed by the first and second clusters) are given as:

$$b_{ij} = g_{ij}\left(1 - \frac{\Lambda v\left(g_{ij} + \beta\left(1 - d_{ij}\right)\right)}{i}\exp(-t_1(\frac{1}{v} - \frac{\Lambda\left(g_{ij} + \beta\left(1 - d_{ij}\right)\right)}{i}))\right) +$$

$$+\left(1 - g_{ij}\right)\left(1 - \frac{\Lambda v\left(\left(1 - g_{ij}\right) + \beta d_{ij}\right)}{j}\exp(-\left(t_1 - D\right)(\frac{1}{v} - \frac{\Lambda\left(\left(1 - g_{ij}\right) + \beta d_{ij}\right)}{j}))\right),$$

$$a_{ij} = (1-d_{ij})\left(1-\frac{\Lambda v\left(g_{ij}+\beta(1-d_{ij})\right)}{i}\exp(-(t_2-D)(\frac{1}{v}-\frac{\Lambda\left(g_{ij}+\beta(1-d_{ij})\right)}{i}))\right)+$$

$$+d_{ij}\left(1-\frac{\Lambda v\left((1-g_{ij})+\beta d_{ij}\right)}{j}\exp(-(t_2)(\frac{1}{v}-\frac{\Lambda\left((1-g_{ij})+\beta d_{ij}\right)}{j}))\right),$$

where g_{ij} and d_{ij} are the proportions of requests from the first and second flows generated and sent for servicing to the first and second clusters, respectively, without their redistribution through the network, D is the average delay in the transmission of a request through the network. The shares of requests of the first and second threads can be adaptively set depending on the accumulation of failures in the first and second cluster. The shares of redistributed requests g and d can be set for the initial state, without adaptation to the accumulation of failures, which reduces the potential effect of distribution, but allows one to reduce the time spent on displaying the states of the remaining clusters in each cluster and recalculating the shares g_{ij} and d_{ij}.

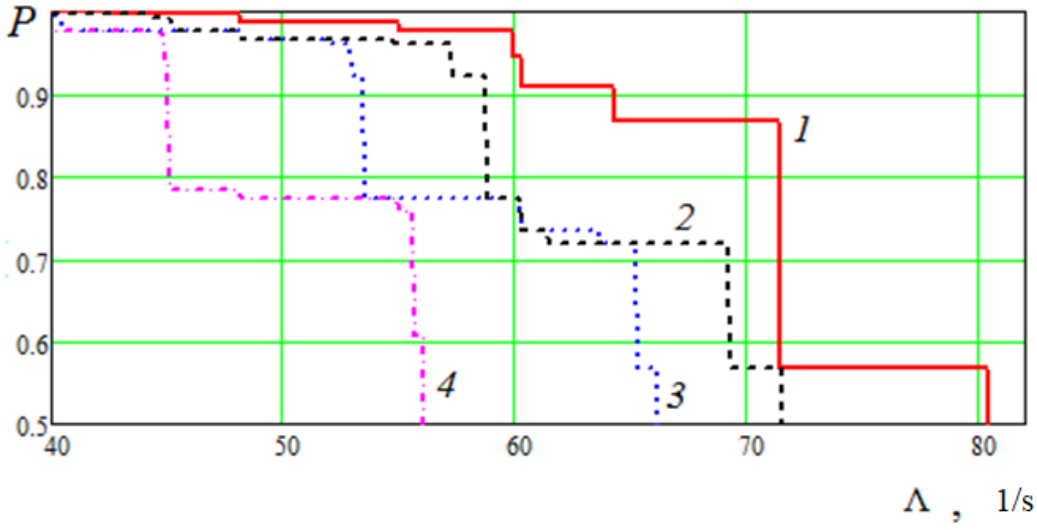
The dependence of the probability of the system operability with the specified shares (g, d) on their intensity, taking into account the probabilistically specified requirement of timeliness of service, is shown in Figure 3 a), and on the share of the flow of requests g - in Figure 3 b) (for the specified values of d) In fig. 4 and at t = 0.01 s curves 1-4 correspond to the cases when the shares (g, d) are equal to: (0.5; 0.5); (0.8; 0.8); (0.1; 0.1); (0.95; 0.1). In fig. 4 b) curves 1 -3 correspond to the share of requests d = 0.5; 0.1; 0.9, formed and executed in the second cluster. The graphs show the impact on the level of system reliability, the proportion of requests redistributed between clusters and the existence of their optimal values, which allow to achieve the maximum system reliability as a result of the redistribution of flows between clusters.

The presented graphs show the efficiency of redistribution of requests between clusters, as a result of which it is possible to increase the reliability of the system, taking into account the requirement of timely service of requests.

The effect is achieved as a result of the fact that, depending on the traffic intensity and waiting time limits for different types of requests, the shares of redistributed requests are set, at which the maximum functional reliability of the system is achieved. It should be noted that if the threads are not stationary, adaptive reallocation of requests is required, taking into account traffic monitoring.

Within the framework of the currently developed model-oriented design of infocommunication systems, the presented models and approaches to increasing the efficiency of multicluster systems can be used to justify the choice and optimization of design solutions for the construction of distributed fault-tolerant real-time systems, including included in cyber-physical systems [27-30].

a)



b)

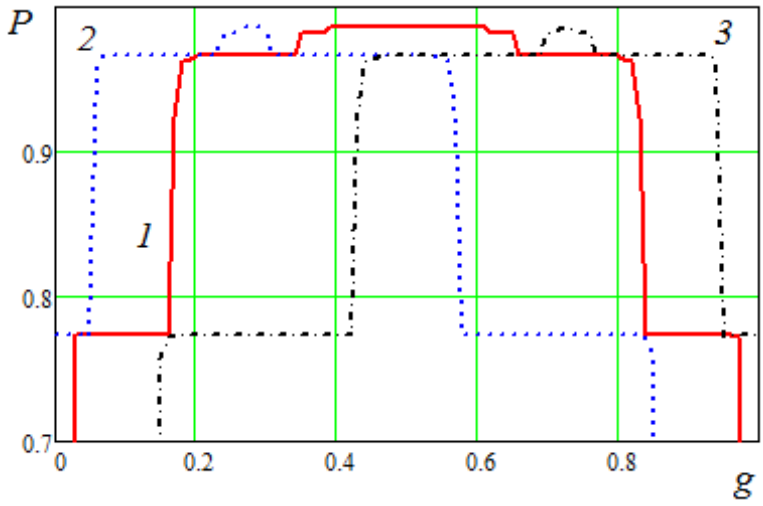


Figure 3. Influence of the intensity of requests and the proportion of their redistribution through the network on the readiness of the system for timely execution of requests

5. Conclusion

An analytical model is proposed and the efficiency of increasing the availability and timely servicing of a flow that is non-uniform in criticality to latency requests is shown, based on the redistribution of requests between clusters through the network with possible replication of the most delay-critical requests.

The influence on the level of reliability of a multicluster computer system, the proportions of requests redistributed between clusters, is established and the existence of their optimal values is shown, allowing to achieve the maximum readiness of a multicluster system for timely execution of a non-uniform flow of requests.

6. References

[1] Sorin, D. Fault Tolerant Computer Architecture. Morgan & Claypool 2009. 103 p.

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

[3] Jo, C., Cho, Y., Egger, B.: A machine learning approach to live migration modeling. In: Proceedings of the 2017 Symposium on Cloud Computing, vol. 17, pp. 351–364. SoCC (2017).

[4] Zakoldaev, D.A., Korobeynikov, A.G., Shukalov, A.V., Zharinov, I.O. Cyber and Physical Systems Technology Classification for Production Activity of the Industry 4.0 Smart Factory (2019) *IOP Conference Series: Materials Science and Engineering*, 582 (1), art. no. 012007. https://iopscience.iop.org/journal/1757-899X doi: 10.1088/1757-899X/582/1/012007.

[5] Astakhova, T., Verzun, N., Kolbanev, M., Shamin, A. A model for estimatingenergy 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).

[6] Zakoldaev, D.A., Korobeynikov, A.G., Shukalov, A.V., Zharinov, I.O. Workstations Industry 4.0 for instrument manufacturing //IOP Conf. Series: Materials Science and Engineering665 (2019) 012015IOP Publishingdoi:10.1088/1757-899X/665/1/012015.

[7] Poymanova, E.D., Tatarnikova, T. M. 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.

[8] Ya, S.B., Tatarnikova, T.M., Poymanova, E.D. Organization of multi-level data storage/ In Informatsionno-Upravliaiushchie Sistemy. Volume 2019, Issue 2, 2019, Pages 68-75 DOI: 10.31799/1684-8853-2019-2-68-75.
[9] Jin, H, Li, D, Wu, S, Shi, X, Pan, X 2009 Live virtual machine migration with adaptive memory compression *Proc. IEEE International Conference on Cluster Computing (CLUSTER '09*). New Orleans, USA, 2009. Art. 5289170. doi: 10.1109/CLUSTR.2009.5289170.
[10] Sahni, S, Varma, V 2012 A hybrid approach to live migration of virtual machines *Proc. IEEE Int. Conf. on Cloud Computing for Emerging Markets (CCEM 2012)* Bengalore India pp 12–16 doi: 10.1109/CCEM.2012.6354587.
[11] Machida, Masahiro Kawato and Y. Maeno, "Redundant virtual machine placement for fault-tolerant consolidated server clusters," *2010 IEEE Network Operations and Management Symposium - NOMS 2010*, Osaka, 2010, pp. 32-39, doi: 10.1109/NOMS.2010.5488431.
[12] Seontae, Kim, Young-ri Choi Constraint-aware VM placement in heterogeneous computing clusters Cluster Computing 23(SI) · March 2020 71-85.
[13] Yang, C., Liu, J., Hsu, C. *et al.* On improvement of cloud virtual machine availability with virtualization fault tolerance mechanism. *J Supercomput* 69, 1103–1122 (2014).
[14] Jo, C., Cho, Y., Egger, B.: A machine learning approach to live migration modeling. In: Proceedings of the 2017 Symposium on Cloud Computing, vol. 17, pp. 351–364. SoCC (2017)
[15] Keller, G., Lutfiyya, H.: Dynamic management of applications with constraints in virtualized data centres. In: Proceedings of IFIP/IEEE International Symposium on Integrated Network Management (IM) (2015).
[16] Wang, Yong Bin, et al. "Markov Process-Based Availability Analysis of Rendering Cluster Systems." Advanced Materials Research, vol. 225–226, Trans Tech Publications, Ltd., Apr. 2011, pp. 1024–1027. Crossref, doi: 10.4028/www.scientific.net/amr. 225-226.1024.
[17] Bogatyrev, V.A., Bogatyrev, S.V., Golubev, I.Yu_Optimization and the process of task distribution between computer system clusters. Automatic Control and Computer Sciences/ 2012. 46(3), c. 103-111.
[18] Bogatyrev, A.V., Bogatyrev, V.A., Bogatyrev, S.V. Multipath Redundant Transmission with Packet Segmentation (2019) *2019 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2019*, art. no. 8840643. doi: 10.1109/WECONF.2019.8840643.
[19] Bogatyrev, S.V., Bogatyrev, V.A., Bogatyrev, A.V. Redundant maintenance of a non-uniform query stream by a sequence of nodes that are grouped together in groups 2020 Wave Electronics and its Application in Information and Telecommunication Systems, WECONF 2020 9131463 DOI: 10.1109/WECONF48837.2020.9131463.
[20] Arustamov, S.A., Bogatyrev, V.A., Polyakov, V.I. Back up data transmission in real-time duplicated computer systems Advances in Intelligent Systems and Computing, 2016, 451, pp. 103–109.
[21] Merindol, P. Improving Load Balancing with Multipath Routing / P. Merindol, J. Pansiot, S. Cateloin // Proc. of the 17-th International Conference on Computer Communications and Networks, IEEE ICCCN 2008. – 2008. – P. 54-61.
[22] Rajeev, V., Muthukrishnan C.R. Reliable backup routing in fault tolerant real-time networks. Proceedings. Ninth IEEE International Conference on Networks, ICON 2001.
[23] Mescheryakov, S., Shchemelinin, D., Efimov, V. Adaptive Control of Cloud Computing Resources in the Internet Telecommunication Multiservice System. In: 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT-2014, pp. 287-293, St. Petersburg, Russia (2014).
[24] Ardulov, Y., Shchemelinin, D., Mescheryakov, S. Monitoring and Remediation of Cloud Services Based on 4R Approach. In: 41st International IT Capacity and Performance Conference, CMG-2015, San Antonio, USA (2015).
[25] Kleinrock, L. Queueing Systems: Volume I. Theory. New York: Wiley Interscience.1975 p. 417. ISBN 978-0471491101.
[26] Kleinrock, L. Queueing Systems: Volume II. Computer Applications. New York:Wiley Interscience. 1976 p. 576. ISBN 978-0471491118.

[27] Ji, H.; Park, S.; Yeo, J.; Kim, Y.; Lee, J.; Shim, B. Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects. *IEEE Wirel. Commun.* 2018, *25*, 124–130.
[28] Sachs, J.; Wikström, G.; Dudda, T.; Baldemair, R.; Kittichokechai, K. 5G Radio Network Design for Ultra-Reliable Low-Latency Communication. *IEEE Netw.* 2018, *32*, 24–31.
[29] Bogatyrev, V.A., Bogatyrev, S.V., Derkach, A.N. Timeliness of the Reserved Maintenance by Duplicated Computers of Heterogeneous Delay-Critical Stream. CEUR Workshop Proceedings. 2019. Vol. 2522. pp. 26-36.
[30] Bennis, M.; Debbah, M.; Poor, H.V. Ultrareliable and Low-Latency Wireless Communication: Tail, Risk and Scale. *Proc. IEEE* 2018, *106*, 1834–1853.