Replication of Requests when Dividing Cluster Nodes Between Threads of Different Criticality to Delays in Queues

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

For distributed real-time computer systems, the possibilities of increasing the probability of timely execution of requests of a heterogeneous stream as a result of replication of requests that are critical to waiting in queues and / or dividing cluster nodes into groups allocated for servicing requests of separate threads are investigated. The number of nodes allocated to service flows of different criticality of requests to wait in queues is determined based on the ratio of the allowable waiting time for critical and other requests, as well as on the ratio of the intensity of these requests. The efficiency of the redundant service of a real-time request is determined by the probability of executing at least one of the generated request copies within the maximum allowable time. For an inhomogeneous total flow of requests, the service efficiency is estimated by the probability of timely servicing of requests of all flows that differ in restrictions on the maximum allowable de-lays in the queues of cluster nodes. The options for servicing requests are analyzed, taking into account possible combinations of replication of requests and dividing the cluster into groups. On the basis of the proposed analytical models, it is shown that the reservation of the most critical to delays in the queues of re-quests can significantly increase the probability of timely servicing not only these requests, but also the total flow as a whole. It is shown that varying the number of cluster nodes designed to serve critical requests makes it possible to increase the probability of timely execution of requests from a heterogeneous flow. The expediency of setting and solving the optimization problem of determining the multiplicity of reservation of requests and dividing the cluster into groups in or-der to maximize the probability of timely execution of requests of a heterogeneous flow is shown.

Keywords

Cluster, Redundant Service, Criticality of Requests to Latency, Heterogeneous Flow, Real Time

1. Introduction

High demands on fault tolerance, availability, error-freeness, continuity and timeliness of real-time computing processes are imposed on the distributed computer sys-tems and networks that are intensively developing at present [1–3]. Reliability requirements are especially stricter for cyber-physical systems [4-9] real time.

High reliability and fault tolerance of distributed computer systems are based on redundancy, resource consolidation (including when combining them into clusters [10-13]), reconfiguration, controlled degradation, adaptive redistribution of request flows [14] and virtualization. [15, 16].

For real-time systems, the possibilities of ensuring high availability, fault tolerance of the structure and computational processes are limited by strict requirements for admissible service delays, and in some cases for ensuring the continuity of the computational process. The continuity of the computing

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

process can be maintained when it is redundantly executed on different equipment with switching to the reserve process after failures are detected.

To reduce the average network delays in the interaction of computer nodes through the network, transport coding allows [17, 18], in which the message is fragmented with the transmission of encoded fragments along different routes, as a result of which, if some fragments (frames) are lost or transmitted errors, the entire message can be restored without repetitions of programs.

To increase the availability of networks and reduce the time of their reconfiguration allows multi-path routing, in which the main and several backup routes (paths) are pre-formed. In the event of node failures, the main path is switched to one of the backup paths, which makes it possible to accelerate the reconfiguration and, in some cases, ensure the required timeliness of real-time computations.

The reliability and timeliness of queries execution in redundant computer systems (server clusters) can be increased as a result of redundant servicing of queries copies with the issuance of the first obtained result [19-20]. Reserved service of a real-time request is successful (timely) if at least one of its replicas is executed without errors without exceeding the maximum permissible waiting time in the queues of the nodes executing it [19-20].

The direction of redundant service requests is a development of the concepts of multipath routing [21-22], multicast transmissions [7], broadcast service [27, 28] and dynamic distribution of requests [14]. The solutions proposed in the article are related to the currently developed concept of Ultrareliable and Low-Latency Wireless Communication [1, 2, 26].

In the case of redundant servicing of a heterogeneous flow, the multiplicity of re-quests reservation can be set depending on their criticality to the permissible waiting time, while additional mechanisms may be involved to ensure the timely servicing of all flows. As such mechanisms, together with the replication of critical requests, the following can be applied: traffic prioritization, load balancing of cluster nodes, allocation of groups of nodes for exclusive or priority servicing of the most critical requests.

The purpose of the work is to study the possibilities of increasing the probability of timely execution of requests of a heterogeneous flow as a result of allocating a group of cluster nodes to serve the most waiting-critical requests, with the possible reservation of critical requests. The number of nodes allocated to serve waiting-critical re-quests in queues is determined based on the ratio of the allowable waiting time for critical and other requests, as well as the ratio of the intensity of these requests.

The efficiency of redundant servicing of a real-time request is determined by the probability of executing at least one of the generated request copies within the maximum allowable time. For a non-uniform total flow of requests, we estimate the ser-vice efficiency by the probability of timely servicing of requests of all flows that differ in restrictions on the maximum allowable delays in the queues of cluster nodes.

2. Options for splitting cluster nodes between threads with possible replication of critical requests

Consider a computational cluster that unites n identical computer nodes (servers), in each of which a queue of requests is organized. We will assume that the following are known: the average query execution time v, the total intensity of the inhomogeneous total input stream of queries Λ, uniting several streams that differ in the admissible waiting time of queries in the queues of cluster nodes.

A request arriving in the cluster can be distributed for servicing to any computer node, modeled by an M/M/1 queuing system with an infinite queue [27, 28]. Each request can be copied with the direction of copies (replicas) for backup execution in the queue of different cluster nodes.

Let us analyze the efficiency of servicing a heterogeneous flow of requests for var-ious combinations of options for booking requests and dividing cluster nodes between threads. In this case, we restrict ourselves to the case when there are two streams for which the maximum allowable waiting time in queues should not exceed t1 and t_2 ($t_1 \leq t_2$), with a possible division of cluster nodes into two groups, including n_1 and (n-n_1) nodes, the first of which is allocated for servicing waiting-critical requests, and the second - for other requests.

Consider the following options for servicing queries in a cluster, taking into account possible combinations of query replication and dividing the cluster into groups:

- The cluster is divided into groups and the queries are not replicated (base case B11).
- The cluster is not divided into groups, replication of critical requests is implement-ed (option B12).
- The cluster is divided into groups, replication of queries is not performed (option B21).
- The division of the cluster into groups and replication of queries are implemented (option B22).

For variant B22, when duplicating requests, it is possible to drill down, in which the options are highlighted:

- replicas are formed only for critical requests and are executed in a group of n_1 nodes allocated for them (option B221),
- replicas are formed only for critical requests, one replica is performed in the first group of n_1 nodes, and the other in the second group of n-n_1 nodes (option B222)

It is also possible to granularize options to replicate less pending requests.

3. Serving a heterogeneous stream without splitting cluster nodes between streams

Consider a cluster system of n parallel (redundant) nodes with a non-uniform flow of requests with the allocation of z types (flows) of requests according to the admissible waiting time in the queues t_1, t_2,…, t_z . The shares of flows of different types of requests are g_1, g_2, … , g_z, , respectively, and their intensities are Λg_1 , Λg_2 , … , Λg_z, and

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

We define the efficiency of servicing a non-uniform flow of requests as the probability that waiting in the queue of requests of all types does not exceed the maximum permissible time t_i for each of the

$$P_c = \prod_{i=1}^{z} P_i. \qquad (1)$$

The metric allows to take into account the influence of individual flows on the final probability of timeliness of servicing the total flow

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

corresponding to the mathematical expectation of the probability that any request for a heterogeneous flow will be completed in a timely manner (taking into account the waiting time allowed for it).

The probability that the delay in the queue of the unreserved request of the *i*-th thread is less than the maximum allowable time t_i with the average execution time of requests of all *z* types equal to v is calculated as

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

For redundant service of requests without dividing the cluster resources (option B12), the probability that the delay in the queue of at least one of the k_i copies of the request of the *i* -th flow does not exceed the admissible time t_i, we calculate as

$$P_i = 1 - \left[\frac{\Lambda_0 v}{n} e^{\left(\frac{\Lambda_0}{n} - \frac{1}{v}\right) t_i} \right]^{k_i} ,$$

where

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

The dependence of the probability of timely execution of requests of the first and total flows on their intensity without dividing cluster nodes into groups is shown in Fig. 1. Curves 1-4 correspond to the probabilities of timely servicing of the first stream of requests with a multiplicity of their reservation k_i = 1-4 (reservation of requests of the second stream is not implemented). Curves 5-8 correspond to the probabilities of timely execution of requests of the total non-uniform flow. The calculation was carried out for n = 5 nodes, g = 0.5, the average query execution time is v = 4 10-4 s, when the maximum allowable waiting time for requests of the first and second flows is equal to t_1 = 2v and t_2 = 5v. From the presented graphs it can be seen that with an increase in the reservation rate of the first flow requests (more critical to the waiting time), especially with a low intensity of the input inhomogeneous flow, it is advisable to increase the reservation rate of the first flow requests to certain limits. It should be noted that an increase in the frequency of reservation of requests of the first stream leads to an increase in the total system load and to a decrease in the probability of timely servicing of requests of the second stream, which negatively affects the probability of timely servicing Pc of the total stream of requests.

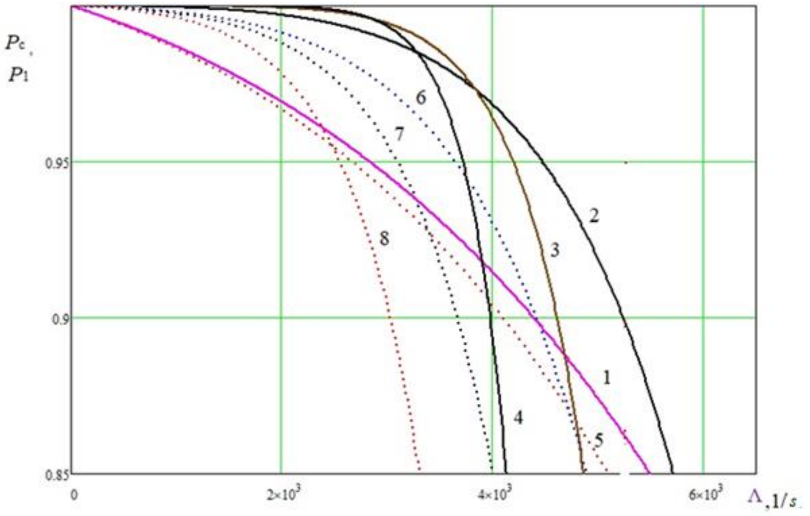

Figure 1: Dependence of the probability of timely execution of requests on the intensity of the total input stream without dividing the cluster nodes into groups

Calculations show that the expediency of reservation of requests depends not only on the intensity of the flow of requests, but also to a large extent on the permissible waiting time.

Fig. 2 shows the dependence of the probability of timely execution of requests of the total flow on the intensity of the input flow with different multiplicity of redundancy and criticality to the permissible waiting time t_1 of the requests of the first flow. Curves 1-4 represent cases when, when duplicating requests of the first thread (k_1=2), the maximum allowable waiting time is t_1=v, 2v, 3v, 4v. Curves 5-7 correspond to the case of non-reservation of requests of the first stream (k_1=1), when t_1 = v, 2v, 3v, and the maximum allowable waiting time for requests of the second stream is t_2 = 5v. The performed calculations confirm the possibility of increasing the probability of timely execution of all requests of a heterogeneous flow with redundant servicing of requests that are critical to waiting in queues.

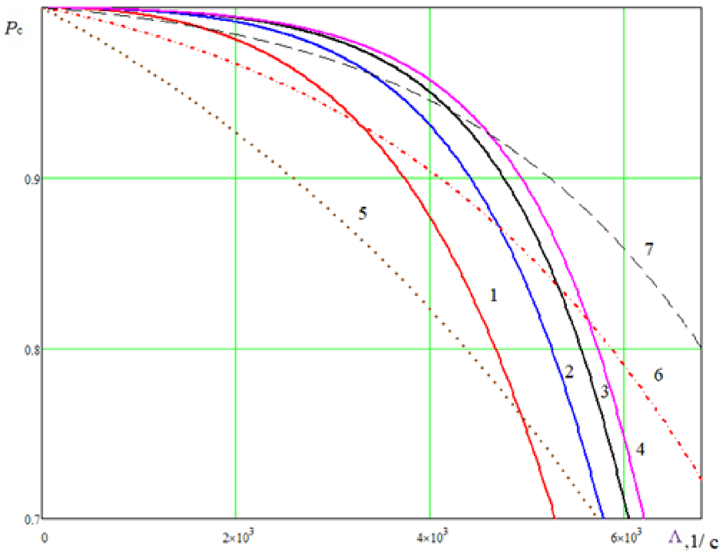

Figure 2: Dependence of the probability of timely execution of requests of the total flow on the intensity of the input flow with different criticality of the requests of the first flow to the admissible waiting time t1 in the case of duplicated and non-redundant service

4. Serving with split cluster nodes between threads without replicating requests

To study the effect of the separation of cluster nodes on the timeliness of computations, consider the case when the combined stream includes two streams with admissible waiting times in queues t_1 and t_2 ($t_1 \leq t_2$), while the probabilities of arriving requests from the first stream *g*, and the second 1-*g*.

Consider the case of dividing the cluster nodes into two groups, including n_1 and n-n_1 nodes, the first of which is allocated to serve waiting-critical requests, and the second is allocated to other requests. In this case, the probability of timely execution of the requests of the first thread without their replication will be:

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

and the second

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

The probability of timely servicing of the total inhomogeneous flow, that is, the probability that the delay in waiting in the queue of requests for both the first and the second flows does not exceed the maximum allowable time for each of them, is defined as

$$P_c = P_1 P_2 .$$

The dependence of the probability of timely servicing of a non-uniform flow on the intensity of requests is shown in Fig. 3, in which curve 1 corresponds to the variant of a cluster that is not divided into groups, and curves 2 to 6 to cases with the allocation of n_1 = 7, 6, 5, 4, 3 nodes into the first group. The dependence of the probability of timely servicing of a non-uniform flow on the number of nodes included in the group for servicing delay-critical requests is shown in Fig. 4, in which curves 1-5 correspond to the intensity of the total flow of requests Λ=150, 200, 300, 400, 500 s^{-1}. The calculation was carried out with the total number of nodes in the cluster n = 12 pcs., The share of the first stream of requests g = 0.1, and the admissible waiting times for requests of the first and second streams t_1 = 0.01 s and t_2 = 0.1 s. The presented dependencies show the importance of the impact on the probability of timely execution of all requests of a heterogeneous flow of dividing the cluster nodes into groups allocated to serve critical and other requests. Moreover, with a low total traffic intensity, the highest probability of timely servicing of all flows is achieved with an increase in the number of nodes allocated to serve the most critical requests. Calculations have shown that as the proportion of waiting-critical requests decreases, the effect of an increase in the number of nodes allocated for their service increases. The given graphs confirm the existence of an optimal number of nodes allocated to serve the most critical requests to waiting in queues.

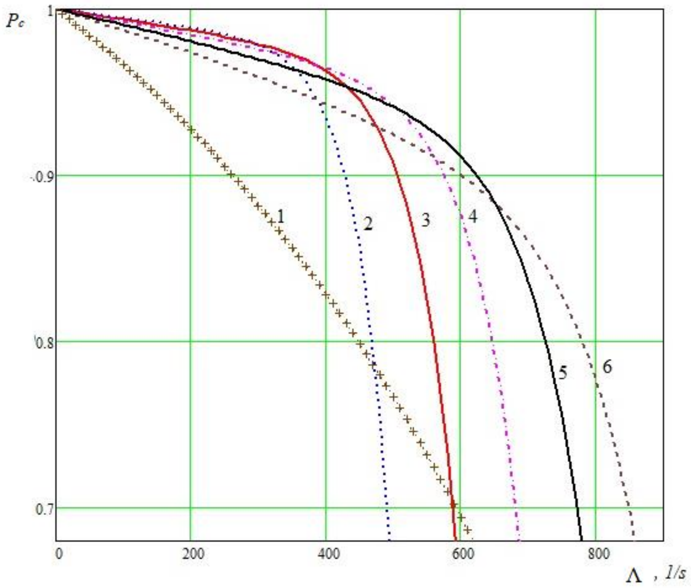

Figure 3: The dependence of the probability of timely servicing of a non-uniform flow on the intensity of the total flow when dividing the cluster nodes into groups

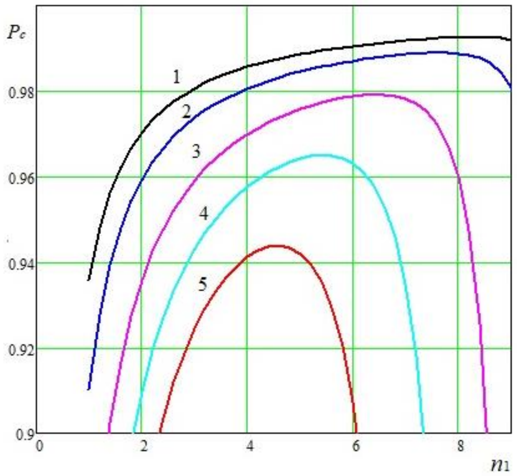

Figure 4: Dependence of the probability of timely servicing of a non-uniform flow on the number of nodes included in the group to serve the most delay-critical requests

5. Redundant service with division of cluster nodes between threads

Let us analyze the effect of replication of the most waiting-critical requests on the timeliness of servicing the total non-uniform flow when dividing the cluster nodes of a group. For the case under consideration, when requests from the first thread arrive, two of its replicas are created, one of which is sent for execution to the first group of n_1 nodes, and the second to a group of n-n_1 nodes, which also receives requests from the second thread (non-critical to waiting for requests). Replication of requests from the second thread is not performed.

The probability of not exceeding the permissible waiting time t_1 for waiting-critical queries in the first and second groups of cluster nodes, respectively, is found as

$$P_{11} = 1 - \frac{\Lambda g v}{n_1} e^{\left(\frac{\Lambda g}{n_1} - \frac{1}{v}\right)t_1},$$

$$P_{12} = 1 - \frac{\Lambda v}{n - n_1} e^{\left(\frac{\Lambda}{n-n_1} - \frac{1}{v}\right)t_1},$$

where v is the average execution time of requests from the first and second threads.

The probability of not exceeding the permissible timeout t1 for at least one of the waiting-critical replicas of a query executed in the first or second group of cluster nodes. is as

$$P_1 = 1 - (1 - P_{11})(1 - P_{12}) . \qquad (2)$$

The probability of not exceeding the admissible waiting time t2 for a request for the second thread running in the second group of cluster nodes is calculated as

$$P_2 = 1 - \frac{\Lambda v}{n - n_1} e^{\left(\frac{\Lambda}{n-n_1} - \frac{1}{v}\right)t_2} . \qquad (3)$$

The probability that the waiting time in the queue of requests of the first and second flows is less than the allowable limit for each of them is determined by formula (1) when calculating P_1 and P_2 by formulas (2) and (3).

The dependence of the probability of timely service on the intensity of the total flow when two groups of cluster nodes are selected and the replication (duplication) of requests of the first flow is shown in Fig. 5 by curves 1-5, respectively, for n_1 = 1-5 nodes. Curve 6 corresponds to the variant of duplicating requests of the first stream without dividing the cluster into groups. Curve 7 reflects the base case without replicating queries and without dividing the cluster into groups. The dependence of the probability of timely servicing of the total inhomogeneous flow on the number of nodes included in the first group is shown in Fig. 6, in which, when the requests of the first flow are duplicated, curves 1-5 correspond to the intensity of the total flow of requests Λ = 300, 400, 500, 600, 700, 800 s^{-1}.

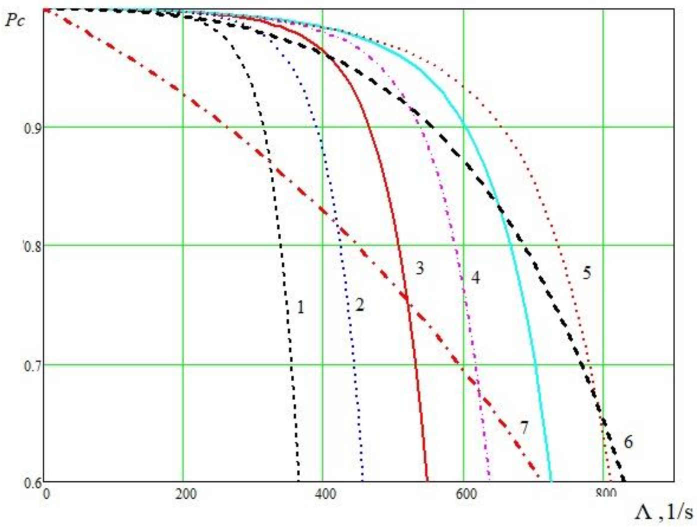

Figure 5: Probability of timely service depending on the intensity of the total flow when two groups of cluster nodes are allocated and the requests of the first flow are duplicated

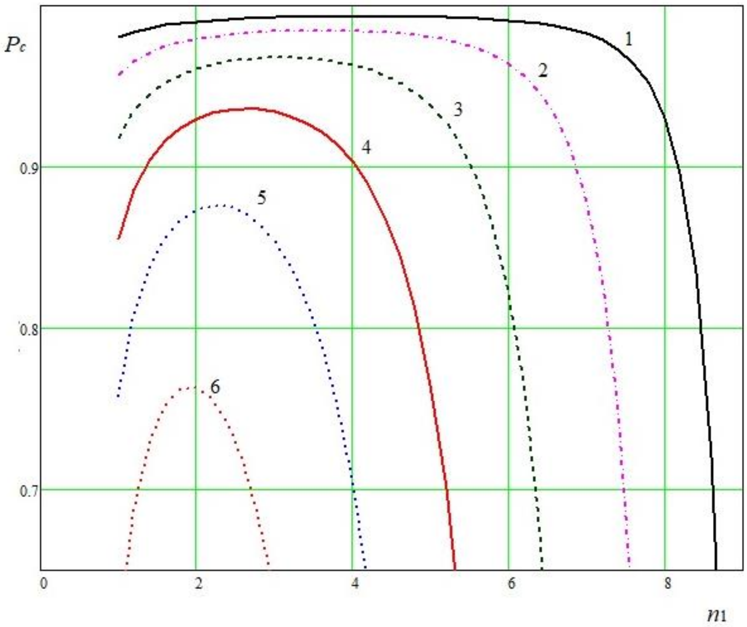

Figure 6: Probability of timely service depending on the number of nodes included in the first group when duplicating requests from the first stream

The presented dependencies allow us to conclude that the reservation of the requests most critical to delays in the queues can significantly increase the probability of timely 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 probability of timely execution of both requests critical to delays in queues and the entire total flow. The expediency of setting and solving the optimization problem of determining the multiplicity of reservation of requests and dividing the cluster into groups in order to maximize the probability of timely execution of requests of a heterogeneous flow is shown.

6. Conclusion

For computer systems of cluster architecture operating in real time, an analytical model is proposed and the effectiveness of redundant service options with the possible allocation of cluster nodes to solve the most critical requests to waiting in queues is determined.

The influence of the multiplicity of redundant service on the probability of timely execution of a non-uniform flow of requests, taking into account sequential redundant execution in nodes of all levels of the system, is analyzed.

It is shown that there is an area of efficiency for redundant servicing of a heterogeneous flow of requests and the division of cluster nodes into groups intended for servicing different flows.

For flows with different criticality to the waiting time in queues, the optimal multiplicity of reservation of requests and / or the number of cluster nodes allocated for their servicing is determined.

7. References

[1] 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.

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

[3] 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.

[4] Zakoldaev, D.A., Korobeynikov, A.G., Shukalov, A.V., Zharinov I.O., Zharinov O.O. 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.

[5] Astakhova T. N., Verzun N. A., Kasatkin V. V., Kolbanev M. O., Shamin, A. A. Sensor network connectivity models. Informatsionno-upravliaiushchie sistemy [Information and Control Systems], 2019, no. 5, pp. 38–50 (In Russian). doi:10.31799/16848853-2019-5-38-50.

[6] Sovetov, B, Tatarnikova, T, Cehanovsky, V. Detection system for threats of the presence of 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.

[7] Zakoldaev, D.A., Korobeynikov, A.G., Shukalov, A.V., Zharinov, I.O. Digital forms of describing Industry 4.0 objects//IOP Conference Series: Materials Science and Engineering, 2019, Vol. 656, No. 1, pp. 012057.

[8] Tatarnikova, T.M., Dzubenko, I.N. IoT system for detecting dangerous substances by smell// Informatsionno-Upravliaiushchie Sistemy. 2018. V. 93, No 2. P. 84-90. DOI 10.15217/issn1684-8853.2018.2.84.

[9] Astakhova, T., Shamin, A., Verzun, N., Kolbanev, M. A. Model for estimating energy consumption seen when nodes of ubiquitous sensor networks communicate information to each other. CEUR Workshop Proceedings MICSECS 2018 - Proceedings of the 10th Majorov International Conference on Software Engineering and Computer Systems. apr. 2019. http://ceur-ws.org/Vol-2344/paper5.pdf.

[10] Machida, F., Kawato, M., Maeno, Y: Redundant virtual machine placement for fault-tolerant consolidated server clusters. In: IEEE Network Operations and Management Symposium, pp. 32–39. IEEE Press, Osaka (2010), doi: 10.1109/NOMS.2010.5488431.020 71-85.

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

[12] Bogatyrev, V.A. Fault Tolerance of Clusters Configurations with Direct Connection of Storage Devices // Automatic Control and Computer Sciences - 2011, Vol. 45, No. 6, pp. 330-337.

[13] Bogatyrev, V.A. Exchange of duplicated computing complexes in fault-tolerant systems // Automatic Control and Computer Sciences - 2011, Vol. 45, No. 5, pp. 268–276.

[14] Bogatyrev, V.A. Protocols for dynamic distribution of requests through a bus with variable logic ring for reception authority transfer (1999) Automatic Control and Computer Sciences, 33 (1), pp. 57-63.

[15] Sahni, S., Varma, V. 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.

[16] Jin, H, Li, D., Wu, S., Shi, X., Pan, X. 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.

[17] Krouk, E., Semenov, S. Application of Coding at the Network Transport Level to Decrease the Message Delay // Proc. of 3rd Intern. Symp. on Communication Systems Networks and Digital Signal Processing. Staffordshire University, UK, 2002. pp. 109—112.

[18] Kabatiansky, G., Krouk, E., Semenov, S. Error Correcting Coding and Security for Data Networks. Analysis of the Superchannel Concept. Wiley, 2005. 288.

[19] Bogatyrev, A.V., Bogatyrev, V.A., Bogatyrev, S.V. Multipath Redundant Transmission with Packet Segmentation // Wave Electronics and its Application in Information and Telecommunication Systems (WECONF 2019) - 2019, pp. 8840643.

[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] Prasenjit, C., Tuhina, S., Indrajit, B. Fault-tolerant multipath routing scheme for energy efficient wireless sensor networksInternational Journal of Wireless & Mobile Networks (IJWMN) Vol. 5, No.2,April 2013 pp 33-45.

[23] Samuylov, A., Moltchanov, D., Kovalchukov, R., Koucheryavy, Y., . Characterizing Resource Allocation Trade-Offs in 5G NR Serving Multicast and Unicast Traffic. IEEE Transactions on Wireless Communications 2020 v19(5),9003488, с. 3421-3434.

[24] Lee, M.H., Dudin, A.N., Klimeno,k V.I. The SM/V/N queueing system with broadcasting service // Math. Probl. in Engineer. 2006. V. 2006. Article ID 98171. 18 p.

[25] Dudin, A.N., Sun, B.: A multiserver MAP/PH/N system with controlled broadcasting by unreliable servers. Automatic Control and Computer Sciences, No. 5, pp. 32–44 (2009).

[26] Bennis, M.; Debbah, M.; Poor, H.V. Ultrareliable and Low-Latency Wireless Communication: Tail, Risk and Scale. Proc. IEEE 2018, 106, 1834–1853.

[27] Kleinrock, L. Queueing Systems: Volume I – Theory. New York: Wiley Interscience. 1975 p. 417. ISBN 978-0471491101.

[28] Kleinrock, L. Queueing Systems: Volume II – Computer Applications. New York: Wiley Interscience. 1976 p. 576. ISBN 978-0471491118.