Redundant multi-path service of a flow heterogeneous in delay criticality with defined node passage paths

V.A. Bogatyrev¹, A.V. Bogatyrev² and S.V. Bogatyrev³

¹ Doctor of Technical Sciences, ITMO University, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, Russia
² Candidate of Technical Sciences, Chief technical officer, NEO Saint Petersburg Competence Center, Saint Petersburg, Russia
³ Chief Information officer, NEO Saint Petersburg Competence Center, Saint Petersburg, Russia

E-mail: vladimir.bogatyrev@gmail.com

Abstract. An analytical model is proposed for estimating the probability of timely redundant servicing of a heterogeneous flow, which implies the creation of replicas of requests with a multiplicity depending on the total allowable waiting time in the sequence of nodes involved in servicing the request. It is shown that the reservation of queries critical to the admissible waiting time allows increasing the probability of timely execution of queries and the intensity of the profit received from servicing a non-uniform flow.

Introduction
High reliability, fault tolerance, and availability of structurally redundant information systems and networks [1-4] can be supported on the basis of multipath routing [5,6], in which the primary and backup routes (paths) are pre-assigned, which allows reconfiguring rather quickly after failures of the primary nodes system with activation of the redundant path. For info-communication systems operating in real-time, including those being a part of cyber-physical systems [7-9], in some cases, to maintain functional reliability, it is required to ensure the continuity of the computing process and the timeliness of servicing requests in case of random failures and malicious influences [10-12].

The criterion for multipath service effectiveness in real-time is the probability of timely execution of requests, taking into account the passage through all nodes sequentially involved in the computational process.

The probability of timely and error-free servicing in cluster systems and multipath transmission systems under certain conditions can be increased as a result of redundant servicing, in which replicas of requests are created where for each of them the path of nodes sequentially involved in the computational process (data transfer) is specified.

In the systems under research, the condition for the timeliness of multi-path redundant service is the passage of at least one replica of all nodes that make up the service path in a time less than a specified maximum permissible value.

For a heterogeneous flow in terms of criticality to the waiting time of requests, the efficiency criterion is the probability of timely execution of all types of requests [7-9].
If the flow of requests is not homogeneous in terms of the maximum allowable waiting time, this article explores the potential for increasing the probability of timely execution of requests of all types as a result of managing the reservation rate depending on their criticality to queue delays.

The complexity of assessing the probability of the timeliness of multipath redundant service is due to the need to consider the accumulation of waiting time in the nodes that make up the paths of the sequential passage of the query replicas.

The peculiarity of the proposed estimate of the desired probability is the calculation of the probability of not exceeding the residual admissible waiting time while the replica passes the next node included in its service route, taking into account the average delays in the queues of the previously passed nodes. The proposed article is devoted to the construction of an analytical model of the probability of the timeliness of redundant servicing of a heterogeneous flow of requests and is aimed at justifying the multiplicity of reservation of requests of various criticality to service delays in order to ensure the timely servicing of all requests of a heterogeneous flow.

1. A redundant service model for a heterogeneous redundant flow

Systems with inhomogeneous input flow in terms of criticality to waiting for requests when they pass sequentially through \( m \) functional nodes will be considered. We will assume that the \( i \)-th node is reserved with multiplicity \( n_i \). The number of replicas to be created (redundancy rate) of queries will be set depending on the maximum allowable waiting time. For each replica, a route (path) of sequential passage of \( m \) nodes involved in its maintenance is specified.

When one of the \( k_j \) replicas of the \( j \)-stream passes through the \( i \)-th node, the probability of not exceeding the maximum permissible time \( t_j \), taking into account the average delays in the queues of the previously passed nodes, can be calculated as:

\[
P_j = 1 - v_i \Lambda_i e^{-\left(\Lambda_i - \frac{1}{N} \sum_{g \neq j} k_g \alpha_g \Lambda g \right) t_j - \sum_{g \neq j} w_g},
\]

where:

\[
\Lambda_i = \sum_{j=1}^{N} \frac{\alpha_j \Lambda k_j}{n_i},
\]

\[
w_g = \frac{\left(1 - \Lambda g v_g \right)}{1 - \Lambda g v_g},
\]

\[
\Lambda g = \sum_{j=1}^{N} \frac{\alpha_j \Lambda k_j}{n_g},
\]

\( v_i \) is the average query execution time by the \( i \)-th node, taking into account all \( N \) types of queries, \( \Lambda \) is the intensity of the total flow of queries (without replication), \( \alpha_j \) is the share of \( j \)-th queries, \( t_j \) is the maximum allowable waiting time for \( j \)-th queries; \( w_g \) is the average waiting time at the \( g \)-th node included in the execution path [13] of the request.

The probability of not exceeding the total allowable time \( t_j \) when traversing the path of one specific replica of a query of the \( j \)-th type and at least one of the \( k_j \) replicas of this type, is defined, respectively, as:

\[
P_j = \prod_{i=1}^{m} P_{j_i},
\]

\[
R_j = 1 - (1 - P_j)^{k_j}.
\]

Compliance with the requirement of timely service for each type of requests of a heterogeneous total flow reflects the criterion:

\[
R = \prod_{j=1}^{N} R_j.
\]
However, the presented criterion does not consider the influence of the shares (intensity) of individual flows on the final probability of timely servicing of the total flow. This influence is taken into account by the metric

\[ A = \sum_{j=1}^{N} \alpha j R_j, \]

corresponding to the mathematical expectation that any non-uniform flow request will be completed in a timely manner.

Combining performance indicators of various types of requests in a single criterion allows taking into account their impact on the overall efficiency of servicing the entire total flow, including the risks associated with the delay in servicing different types of requests. The metric of the total flow efficiency should be determined based on the features of the applied problems solved by the system.

The profit from the provision of information services can be chosen as a generalized indicator of the efficiency of the system functioning. So, if the profit from the timely servicing of the request of the \( j \)-th flow is \( c_j \), and the penalties for the delay in servicing are \( s_j \), then the mathematical expectation of the profit from fulfilling one request of the total flow will be

\[ C = \sum_{j=1}^{N} \alpha j \left( R c_j - (1 - R_j) s_j \right). \]

Profit per unit of time from the provision of information services (profit intensity) can be defined as

\[ D = \Lambda \sum_{j=1}^{N} \alpha j \left( R c_j - (1 - R_j) s_j \right). \]

2. Results of calculating the probability of timely service and the intensity of profit from servicing requests of a heterogeneous flow.

We need to determine the effectiveness of the considered design solutions for the organization of redundant multipath transmissions.

In the calculations, we will assume that there is a heterogeneous flow that includes two types of requests \((N = 2)\), for the first of which the maximum allowable total waiting time in all \( m \) nodes involved in the service process is \( t_1 = 0.1 \) s, and for the second - \( t_2 = 0.5 \) s. The average service time of requests of different types at all nodes will be assumed to be the same and equal to \( v = 0.1 \) s. Reservation of requests of the second type is not assumed.

The dependences of the probabilities \( R \) of timely servicing of requests of the first and second types on the intensity of their arrival \( \Lambda \) is shown in Fig. 1, and on the multiplicity of reservation of requests of the first type \( a \) - Fig. 2. In Fig. 1, the probabilities of timely execution of requests of the first and second types \( R \) correspond to curves 1–3 with the multiplicity of reservation of requests of the first type \( k = 1, 2, 5 \) and their share \( \alpha = 0.1 \). In Fig. 2, for the intensity of the total flow \( \Lambda = 5 \) 1/s, curves 1-4 correspond to the probabilities \( R \) of timely servicing of requests of two types with \( a = 0.2, 0.3, 0.4, 0.5 \). The calculation was carried out with sequential passage of all query replicas through \( m = 5 \) nodes, with \( n = n_i = 10 \).

From the presented dependencies, we can conclude that the reservation of the most critical queries to the allowable waiting time increases the probability of timely execution of all types of queries. At the same time, there is an optimal multiplicity of reservation of requests, depending on their criticality to the permissible waiting time, which provides the maximum probability \( R \) of timely execution of all types of requests of a heterogeneous flow. Moreover, the effect of reserving latency-critical requests is most clearly manifested in the case of ensuring a high probability of timely execution of requests of a heterogeneous flow, which is important for practice.

The dependence of the profit \( D \) obtained by servicing the total flow of requests per unit of time on the intensity of the total input flow of requests \( \Lambda \) is shown in Fig. 3, and on the multiplicity of reservation of requests of the first stream in Fig. 4. The calculations are performed when the requests of the second stream are not reserved, and \( t_1 = 0.1 \) s, and \( t_2 = 0.5 \) s.
In Fig. 3, curves 1-3 correspond to the multiplicity of reservation of requests of the first stream $k = 1, 2, 5$. The calculation is performed at $\alpha = 0.1$, when the requests of the first stream are duplicated ($k = 2$).

In Fig. 4, curves 1 - 4 correspond to the proportions of the first flow $\alpha = 0.2, 0.3, 0.4, 0.5$. The calculation results presented in Fig. 4a and 4b are performed, respectively, at the intensity of the total flow $\Lambda = 5 \, \text{1/s}$ and $\Lambda = 12 \, \text{1/s}$. The given dependences confirm the efficiency of reservation of the most critical requests of a heterogeneous flow.

Fig. 1. Dependence of the probability of timely servicing of requests of different types on the intensity of the input flow of requests.

Fig. 2. Dependence of the probability of timely servicing of requests of different types on the multiplicity of reservation of requests of the first type.
Fig. 3. Dependence of the profit $D$ obtained when servicing the total flow of requests per unit of time on the intensity of the total input flow of requests.

Fig. 4. Dependence of the profit $D$ obtained when servicing the total flow of requests per unit of time on the multiplicity of reservation of requests of the first type.
The proposed models can be used for the optimal design of real-time computer systems, including control and communication systems operating [14-16] as part of cyber-physical systems [17-19], developed within the framework of the concept of building highly reliable systems with low transmission delays (Ultrareliable and Low-Latency Wireless Communication) [20-23].

Conclusion
An analytical model is proposed for assessing the probability of timely redundant servicing of a non-uniform flow by a sequence of redundant system nodes, which assumes replication of requests with a frequency depending on their criticality to the allowable request waiting time.

It is shown that the reservation of the most critical queries to the admissible waiting time allows increasing the probability of timely execution of queries and the intensity of the profit received from servicing a non-uniform flow.

The expediency of optimization is shown for the multiplicity of reservation requests depending on their criticality to the permissible waiting time.

References
[1] Seontae Kim, Young-ri Choi Constraint-aware VM placement in heterogeneous computing clusters Cluster Computing 23(SI) · March 2020 71-85
[2] 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).
[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] 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)
[5] Prasenjit Chanak, Tuhina Samanta, Indrajit Banerjee Fault-tolerant multipath routing scheme for energy efficient wireless sensor networks, International Journal of Wireless & Mobile Networks (IJWMN) Vol. 5, No.2, April 2013 p 33-45
[6] Rajeev V., Muthukrishnan C.R.. Reliable backup routing in fault tolerant real-time networks. Proceedings. Ninth IEEE International Conference on Networks, ICON 2001.
[7] 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, WECNF 2019, art. no. 8840643. doi: 10.1109/WECNF. 2019.8840643
[8] 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, WECNF 2020 9131463 DOI: 10.1109/WECNF48837.2020.9131463
[9] Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Timely Redundant Service of Requests by a Sequence of Cluster. CEUR Workshop Proceedings. 2020. Vol. 2590. pp. 1-12.
[10] Yan, J., Zhang, M., Fu, Z. An intralogistics-oriented Cyber-Physical System for workshop in the context of Industry 4.0 (2019) Procedia Manufacturing, 35, pp. 1178-1183. http://www.journals.elsevier.com/procedia-manufacturing doi:10.1016/j.promfg.2019.06.074
[11] 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
[12] Bogatyrev, V., Derkach, A. Evaluation of a cyber-physical computing system with migration of virtual machines during continuous computing. Computers 2020 9(2),42 doi:10.3390/computers9020042/. www.mdpi.com/journal/computers
[13] Kleinrock, L. Queueing Systems: Volume I – Theory. New York: Wiley Interscience. 1975 p. 417. ISBN 978-0471491101

[14] Sovetov, B.Y., Tatarnikova, T.M., Cehanovsky, V.V. Detection system for threats of the presence of hazardous substance in the environment (2019) Proceedings of 2019 22nd International Conference on Soft Computing and Measurements, SCM 2019, art. no. 8903771, pp 121-124

[15] Astakhova, T.N., Verzun, N.A., Kasatkin, V.V., Kolbanev, M.O., Shamin, A.A. Sensor network connectivity models (2019) Informatsionno-Upravliaiushchie Sistemy, (5), pp. 38-50. http://www.i-us.ru/index.php/ius/article/view/4566/2609 doi: 10.31799/1684-8853-2019-5-38-50

[16] Ya, S.B., Tatarnikova, T.M., Poymanova, E.D. Organization of multi-level data storage (2019) Informatsionno-Upravliaiushchie Sistemy, 2019 (2), pp. 68-75. https://www.i-us.ru/jour/article/view/469/399 doi: 10.31799/1684-8853-2019-2-68-75

[17] 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

[18] 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

[19] Zakoldaev, D.A., Korobeynikov, A.G., Shukalov, A.V., Zharinov, I.O. Workstations Industry 4.0 for instrument manufacturing (2019) IOP Conference Series: Materials Science and Engineering, 665 (1), art. no. 012015. https://iopscience.iop.org/journal/1757-899X/ doi: 10.1088/1757-899X/665/1/012015

[20] Murtaza Ahmed Siddiqi1, Heejung Yu and Jingon Joung 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

[21] 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.

[22] 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.

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