Intercomplex exchange alternatives in duplicated computing complex clusters

Stanislav Turbin
Department of Information Systems and Technologies
St. Petersburg State Forestry University
Saint-Petersburg, Russia
st.turbin@gmail.com

Vladimir Bogatyrev
Department of Computation Technologies
ITMO University
St. Petersburg, Russia
vladimir.bogatyrev@gmail.com

***Abstract*— Increased reliability of distributed systems can be achieved as a result of redundancy nodes implemented as duplicated computer complexes, which combined into clusters, and the efficiency of these clusters is largely depended on the intra and intercomplex data exchange organization. The options of intercomplex data exchange are considered from the point of view of a redundant channel, where the implementation of duplicated computing requires data transfer to the memory of the two computers of the addressed duplicated complex. The objective of this research is the development of the simulation models supporting the selection of design solutions for the intra- and intercomplex data exchange organization in the clusters of computer complexes aimed at the high reliability and performance increase in duplicated computing service. Based on the simulation models implemented on the AnyLogic platform, the efficiency areas of the considered intercomplex exchange options are established.**

Keywords*— *reliability, duplicated computing, redundancy, intercomplex exchange.

I. Introduction

Fault tolerance, reliability and dependability of distributed computer systems are achieved through redundancy of key computer and communication nodes and their interconnections [1-5]. When solving critical tasks, in addition to using structural redundancy of computer nodes, reliability may be improved by duplicating the computational process [6, 7]. In this case, the computers may be united into duplicated computing complexes (DCC), where resource consolidation is maintained while they are united into clusters [8-10]. The reliability, fault tolerance and productivity of the duplicated computations in DCC clusters are determined to a great extent by setting up intracomplex and intercomplex data exchange [7-11], including using multipath data transmission [12-15] and channel aggregation [16-19].

Selection of alternatives for constructing computer systems for domestic use requires a system-engineering basis of analytical and simulation models [19-20].

Publications [7-11] studied the impact of exchange on the reliability and delay in servicing duplicated computations in DCC clusters, based on analytical models that propose an exponential distribution of random quantities affecting the e effectiveness of the studied processes.

The objective of this work is to develop simulation models to support the selection of design solutions to set up intracomplex and intercomplex exchange in duplicated computer complex clusters in order to ensure high reliability and lower servicing costs when setting up duplicated computations during failures, malfunctions and computational errors.

II. OBJECT OF RESEARCH

The object of the research is a group of n duplicated computing complexes (DCC), united into a cluster [10,11]. The system structure is shown in Fig. 1.

Each of the DCC semi-complexes contains a processor, memory module (M), network adapters (NA) and intracomplex connection adapter (S). The connection between the DCC cluster is made via mainline redundancy. When important tasks are performed, computations are duplicated in two semi-complexes. During intercomplex exchange, information must be sent to the memory modules of two machines (semi complexes) of the addressed complex (node).

We will review the alternatives [20] of exchange between the DCC through a redundancy communication medium, when it is necessary to enter transmitted data into the memory modules of two semi-complexes of the addressed computing node.

Alternative 1 (Fig. 2): the first and second semi-complexes of the i-th DCC transmit the packets respectively through the switch *Sw1* and the switch *Sw2*. As a result of the exchange, data of the i-th complex are entered into memory module M of both semi-complexes of the j-th complex [10]:

$$NA_{1i} \rightarrow Sw_1 \rightarrow NA_{1j} \rightarrow M_1;$$

$$NA_{2i} \rightarrow Sw_2 \rightarrow NA_{2j} \rightarrow M_2;$$

Alternative 1 is implemented if the two main lines are operating and the computer complexes are interacting. Modifications of the first exchange alternative are possible based on multiple access.

Alternative 1.1: Both semi-complexes of the i-th DCC present requests for access respectively to the first switch *Sw1* and the second switch *Sw2*, after the presentation of which the transmission is made through them. As a result, the delivery of packet copies may not be simultaneous.

Alternative 1.2: Multiple access is implemented only to one switch, while the second is made available together with the first, after which transmission occurs. In this case, delivery of the packets to the addressee will be simultaneous. Alternative 2 (Fig. 3): one of the semi-complexes of the i-th complex sends data over one of the switches to the memory of the semi-complex of the j-th DCC; they are sent from the memory module of the semi-complex that received the data to the memory module of the other semi-complex through the intra-complex connection adapter:

Fig. 1. Duplicated computing complex cluster

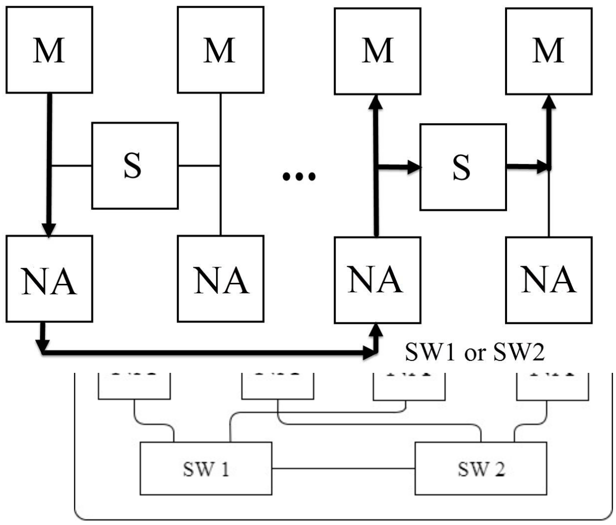
$$NA_{1i} \rightarrow Sw_1 \rightarrow NA_{1j} \rightarrow M_1;$$

$$M_{1j} \rightarrow A_j \rightarrow M_{2j}.$$

Modification of alternative 2 is possible, in which the request from the DCC is presented simultaneously to different switches. After access is given to one of the switch, the request for access to the other switch is removed at the hardware level.

III. ANALYTICAL MODEL FOR ASSESSING THE AVERAGE INTERCOMLEX EXCHANGE TIME

We present the exchange through each main line by the simplest mass servicing system of type M/M/1 [5]. For the cluster structure in Fig. 1, in the initial condition (without fzailures), the average stay time of requests for intercomplex exchange (from formation of the request to data transmission to the memory modules of both semi-complexes of the addressed DCC) for alternative 1.1 and alternative 1.2 is calculated as:



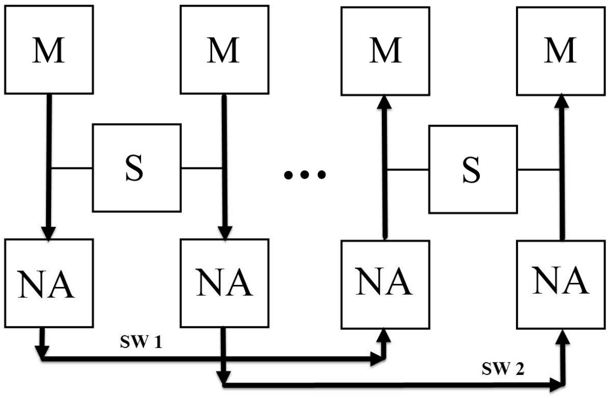


Fig. 3. Alternatives for setting up exchange, second exchange alternative

$$T_{11} = \frac{v}{1-\lambda v} + d, \tag{1}$$

$$T_{12} = \frac{v}{1-\lambda v}, \tag{2}$$

and for alternative 2 as:

$$T_2 = \frac{v}{1-0,5\lambda v} + \alpha v. \tag{3}$$

Here: λ - intensity of requests for intercomplex exchange;

ν - average transmission time through the main line, coefficient;

α - establishes a correlation of transmission speed through the main line via network adapters and via multicomplex exchange adapter A;

d - determines the difference in time for providing

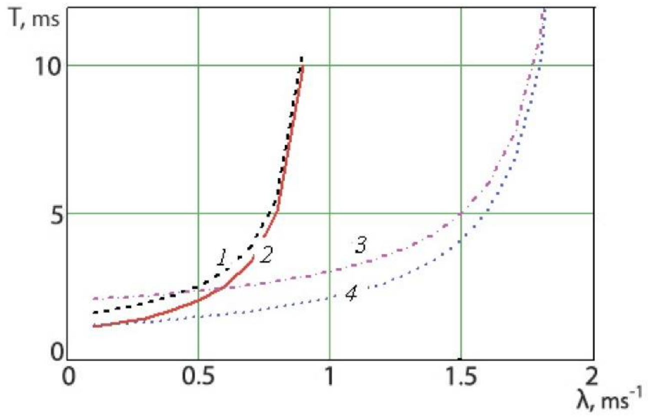


Fig. 4. Average stay time of requests for intercomplex exchange

access to different main lines from one DCC (due to the identical time of the process of multiple access to different main lines).

In the determined methods of multiple access, in particular, in the marker method (if the logic ring of authority transmission coincides for both main lines and only duplicated transmission occurs through the min line) d is close to zero. For random access methods, the value d may be substantial [20].

The dependences of the average stay time of requests for intercomplex exchange on their formation intensity with ν=1 ms are given in Fig. 4. Curves 1 and 2 correspond to setting up exchange by alternative 1 (with d=0 ms) and alternative 2 (with d=0.5ν ms), while curves 3-4 correspond to alternative 2 with α =1 and α =0.1.

These calculations show the distinct advantage of setting up exchange by alternative 2, and as the request intensity increases λ, this effectiveness rises. Additionally, alternative 2 permits an increase in the maximum request intensity maintained by the system under conditions of stability. Comparison of curves 3 and 4 shows the expediency of using high-speed exchange units for intracomplex connection. If request intensity is low, alternative 2 may prove less effective, but it should be noted here that for practical purposes it is more important to reduce the stay time of the requests if their intensity is high. The existence of an area for effective use of these alternatives of intercomplex exchange has thus been demonstrated. For more detail research taking into account failures of errors of transmissions and a possibility of loss of data simulation models are built.

IV. CONSTRUCTION OF SIMULATION MODEL

It is believed that all requests entering the system are critically important, therefore fulllment of the requests is duplicated in the semicomplexes. In order to duplicate the computations, the data to be transmitted between the complexes are entered into the memory modules of each of the two semi-complexes of the addressed node. The parameters of each system component assign:

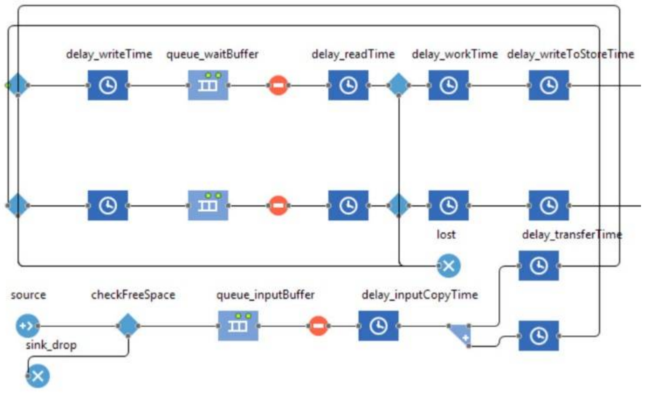


Fig. 5. Simulation model of DCC functioning

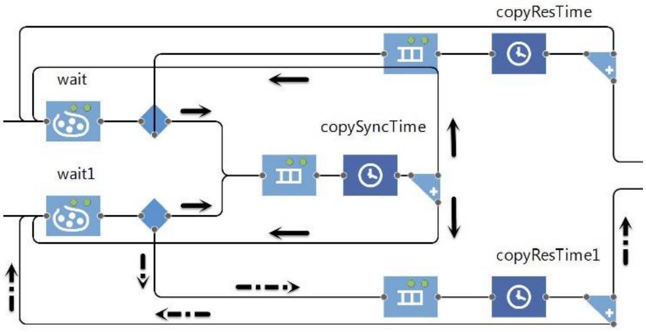


Fig. 6. Model of comparison and redundant copying

- the buffer size for incoming requests;
- the time for copying, sending, processing and comparing requests;
- the sizes of the transmittable packets, bit error probability;
- verification period of duplicated computations;
- time for generation and delivery of controlling signals;
- intensity of request receipt;
- intensity of inter-machine exchange.

The simulation model for functioning of the duplicated computing complex is shown in Fig. 5

The controller verifies whether there is space in the input buffer and places the incoming request in the processing sequence, otherwise a disabling block is activated to halt the flow of messages. The controller copies the received requests and places the copies into the semi-complex working The processor processes the incoming requests from the working memory and places the results in the external storage.

The *Source* unit is the packet generator source. The packet input frequency is regulated by the intensity parameter (number of packets generated randomly during a certain time interval).

The *Queue* units are used as simulators of memory spaces and buffers. The *Wait* units operate similarly to the *Queue*, but may provide access to any request in the sequence, and not only those at its end and beginning. The *Delay* units make delays in the DCC. The *Sink* units are the end points in the packet route.

The input buffer is implemented in turn (*inputBuffer*), by the selection and variable unit *bufferSpace* (the value equal to the buffer space is initially stored). If a request enters in turn or is made from it, in the *bufferSpace* the buffer space is correspondingly decreased or increased. If the size of the next incoming packet exceeds the free space, the packet will be dropped by the selection operator. Copying is simulated by the *Split* unit and by several delays. The *copyTime* unit determines the request duplication time, while the *transferTime* unit determines its transfer time to the DCC halves. The *waitBuffer* unit arranges the request line. The wait unit stores the processed requests and prepares packets for data transmission through the network. The *writeTime*, *readTime* and *writeToStoreTime* units read, delay for reading and write to memory. After the *readTime* unit, the selection operator is engaged to check the packet loss, the probability of which is adjusted in the parameters. The *workTime* unit assigns the request processing time. The gateways stop the packets during certain processes (e.g., during failure or during synchronization).

The model of comparison and redundant copying is provided in Fig. 6. The unit *copySyncTime* determines time delay on copying. The wait units realize queue on processing with a possibility of access to any member. The unit copyResTime will organize backup. Pointers would explain the direction movement of requests in system.

V. ASSESSMENT OF INTERCOMPLEX EXCHANGE EFFECTIVENESS

The simulation experiments established the mean time for stay of the requests in the system and their likely delivery to the addressed complex from the alternative of arranging intercomplex exchange. The simulation modelling results are given in Fig. 7. Its shows the dependence of the likely packet delivery in the assigned time on the intensity of the input ow with varying data transmission speed.

Fig. 8 shows the dependence of the mean time for stay of a request in the system on the intensity of request influx and the likely bit errors in channel b, assuming b =0.0001, 0.00001. It is apparent from the quires that with assigned parameters, transmission with internal space ensures the best indicators for mean time of stay of the requests in the system and their likely timely delivery.

When the system functioning systems change (e.g., likelihood of bit errors, low intensity) and failures accumulate, the proposed simulation modules are applicable for a comparative analysis of the reviewed exchange alternatives and substantiation of their selection, including, adaptive during system functioning.

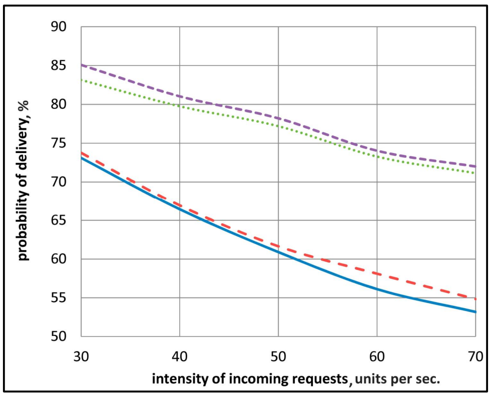


Fig. 7. The dependence packet delivery in the assigned time

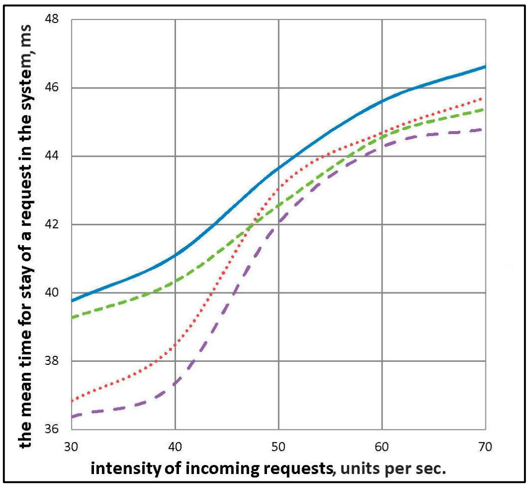


Fig. 8. Mean time of stay of requests in the system

VI. EFFECT OF INTERCOMPLEX EXCHANGE VARIANTS ON THE RELIABILITY OF THE SYSTEM

Effect of intercomplex exchange variants on the reliability of the system. Let us compare the reliability of DCC clusters with the redundancy main lines in the implementation of duplicated computations with various options for organizing of intercomplex exchange. For duplicated computations with intercomplex exchange, there must be the delivery of the transmitted data to the memory modules of both half-complexes. When implementing a duplicated computational process in the i-th and j-th complexes, the data storage and data processing facilities in the two half-complexes of both interacting complexes need to be there and the serviceability of equipment used in the implementation of the exchange depends on its version, thus, the reliability of the Pc system, realizing the duplicated computational process in two complexes, is defined as

$$P_c = p_b^4 P_0 \tag{4}$$

where p_b is the probability of an operational state (readiness) of means of storage and processing of one half-complex.

P_0 is the readiness of the means of exchange, involved in the implementation of intercomplex exchange, depending on its variant.

If there are no recovery queues

$$p_b = \mu_b / (\mu_b + \lambda_b), \tag{5}$$

where λb, μb - intensity of failure and recovery.

When exchanging on the variant 1, all four network adapters of the i-th and j-th complex are required to be fixed, thus

$$P_0 = p_c^4, \tag{6}$$

where p$_c$ - readiness of network adapter.

$$p_c = \mu_c / (\mu_c + \lambda_c), \tag{7}$$

where λ_c, μ_c - the intensity of failures and recoveries of the network adapters.

In the case of intercomplex exchange, variant 2 requires the operability of a pair of network adapters of interacting complexes, connected to at least one trunk, in addition, the receiving side needs the operability of the intercom adapter (the probability of which is pa) and, therefore, in a one-way exchange:

$$P_0 = p_a[1-(1-p_c^2)^2], \tag{8}$$

in a bilateral exchange:

$$P_0 = p_a^2[1-(1-p_c^2)^2], \tag{9}$$

where

$$p_a = \mu_a / (\mu_a + \lambda_a), \tag{10}$$

λ_a, μ_a - intensity of failures and recovery of adapters of intracomplex communication.

Let us consider a combined variant of intercomplex exchange, in which:

- if adapters of intracomplex communication of both interacting complexes are functioning, communication is realized according to the variant 2,
- if the adapter of intracomplex communication only in one complex is functioning, then one of the directions is exchanged, according to the variant 1, and the other - according to the variant 2,
- if the adapters of intracomplex communication are not working in both interacting half-complexes, then the exchange is carried out according to the variant 1.

The reliability of intercomplex exchange on the described combined variant for a one-way exchange is defined as:

$$P_0 = p_a[1-(1-p_c^2)^2] + (1-p_a)p_c^4, \tag{11}$$

Herewith p_a - the probability of the operability of the intercom adapters in the complex of the data receiver.

In a two-way exchange, the reliability of a system of two complexes that realize duplicated computations is defined as

$$P_0 = p_a^2[1-(1-p_c^2)^2] + (1-p_a^2)p_c^4, \tag{12}$$

Herewith p_a^2 - the probability of the availability of intercom adapters in both interacting complexes

The above estimates of reliability are obtained on the assumption of the ideal control; in which we assume that all failures are detected instantaneously without the influence of control on a possible reduction in the reliability of the system. Models can be used in case of development of research in the direction of analysis the influence of control on the reliability of clusters, duplicated complexes [20]. At the same time, models can be taken as a basis that take into account the effect of the variability of organization of control of duplicated complexes on the probabilities of operable, inoperative and unsafe conditions of complex operation in conditions of undetected failures, including when duplicated calculations are performed with periodic and in particular optimal initialization of test control.

VII. CONCLUSIONS

Alternatives were analyzed for setting up intercomplex exchange between duplicated complexes, united into a cluster via duplicated main lines.

A simulation model has been proposed that was implemented on the AnyLogic platform and analyzes the effectiveness, and depending on the functioning conditions and accumulation of failures, also substantiates the selection (including adaptive) of alternatives for intercomplex exchange in duplicated computer complex clusters.

The influence of the organization of intercomplex exchange on the reliability of the system in the implementation of duplicated calculations is considered.

Areas of the effective use of these alternatives of intercomplex exchange were depended.

It was demonstrated that if the main line load is great, exchange is effective for data transmission to the addressed complex via one of the main lines with subsequent intracomplex transmission between the memory modules of its semi-complexes.

REFERENCES

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

[2] Koren I. Fault tolerant systems. Morgan Kaufmann publications, visit our San Francisco, 2009.

[3] Aysan H. Fault-tolerance strategies and probabilistic guarantees for real-time systems Mälardalen University, 2012.

[4] Aliev T. The synthesis of service discipline in systems with limits // Communications in Computer and Information Science, IET. vol. 601, pp. 151-156.

[5] Vishnevskii V.M. Teoreticheskie osnovy proektirovaniya (Theoretical Foundations of Design). Moscow: Tekhnosfera, 2003.

[6] Dudin, A. N, Sun B. A multiserver MAP/PH/N system with controlled broadcasting by unreliable servers // Automatic Control and Computer Sciences 2000, vol 43, №. 5, pp.32-44.

[7] Bogatyrev V. A. and Bogatyrev A. V. Functional reliability of a realtime redundant computational process in cluster architecture systems // Automatic Control and Computer Sciences. 2015, vol. 49, №. 1. pp. 46-56.

[8] Bogatyrev V. A. and Bogatyrev S. V. and Golubev I. Yu. Optimization and the process of task distribution between computer system clusters // Automatic Control and Computer Sciences. 2012, vol. 46, №. 3. pp. 103-111.

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

[10] Bogatyrev V.A. Exchange of Duplicated Computing Complexes in Fault tolerant Systems // Automatic Control and Computer Sciences, 2011, vol. 45, № 5, pp. 268-276.

[11] 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, IET – 2016, vol. 451, pp. 103-109.

[12] Parshutina S.A , Bogatyrev, V.A Redundant Distribution of Requests Through the Network by Transferring Them Over Multiple Paths //Communications in Computer and Information Science, IET - 2016 vol. 601, pp. 199-207.

[13] Bogatyrev V.A., Parshutina, S.A. E-ciency of Redundant Multipath Transmission of Requests Through the Network to Destination Servers // Communications in Computer and Information Science, IET – 2016, vol. 678, pp. 290-301.

[14] Poptcova N.A , Bogatyrev V.A., Parshutina S.A., Bogatyrev A.V. Ef-ciency of Redundant Service with Destruction of Expired and Irrelevant Request Copies in Real-Time Clusters // Communications in Computer and Information Science, IET – 2016, vol. 678, pp. 337-348.

[15] Kolomoitcev V.S., Bogatyrev V.A. The fault-tolerant structure of multilevel secure access to the resources of the public network // Communications in Computer and Information Science, IET – 2016, vol. 678, pp. 302-313.

[16] Bogatyrev V.A. An interval signal method of dynamic interrupt handling with load balancing // Automatic Control and Computer Sciences, 2000, vol. 34, pp. 51-57.

[17] Bogatyrev V.A. Protocols for dynamic distribution of requests through a bus with variable logic ring for reception authority transfer // Automatic Control and Computer Sciences, 1999, vol. 33, № 4, pp. 57-63.

[18] Bogatyrev V.A. Increasing the fault tolerance of a multi-trunk channel by means of inter-trunk packet forwarding // Automatic Control and Computer Sciences, 1999, vol. 33, № 2, pp. 70-76.

[19] Bogatyrev V.A. On interconnection control in redundancy of local network buses with limited availability // Engineering Simulation, 2016, vol. 16, № 4, pp. 463-469.

[20] Bogatyrev V.A., Vinokurova M.S. Control and Safety of Operation of Duplicated Computer Systems // Communications in Computer and Information Science - 2017, vol.700, pp. 331-344.