DOI: https://doi.org/10.30837/csitic52021232786

Double-Circuit Organization of a Distributed Computer System for the Effective Control of ITS Functioning

Vadym Mukhin[1]

Yaroslav Kornaga[1]

Valery Zavgorodnii[2]

Maksym Basaliy[1]

Oleg Mukhin[1]

[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institut", 37 Peremohy Ave, Kyiv UA-03056, Ukraine, e-mail: v_mukhin@i.ua

[2]State University of Infrastructure and Technology, 9 Kyrylivska st, Kyiv,UA-04071, Ukraine, e-mail: zavgorodnii@i.ua

Abstract. *The ensuring of the continuous control and software updates in the control nodes of a distributed data processing system requires the development of a network structure in such a way that a group of control nodes supports the functioning of its subnetwork and, if necessary, may transfer the data processing to the another group of control nodes. This structure consists of two circuits: a control node circuit and a data circuit.*

Keywords: *double-circuit system, computer system, network-centric.*

I. Introduction

Let's describe the functioning of this two-circuit system. As shown on Figure 1, a computer system consists of two circuits and three subnets. Each subnet has a control node in the control node circuit. This structure for the distributed data processing ensures the control of the entire system and based on the network-centric control. To ensure the effective control, it is necessary to determine the number of nodes, where the controlling software will be installed, that is, to solve the optimization problem: to find the optimal number of control nodes depending on the processing time of data packets. So, it is necessary to determine the dependence of the time that is spent for the passing from the one control node of a subnet to the another node of the another subnet from the number of control nodes in the circuit of control nodes and the time that is spent for the packet to be delivered to the database nodes, depending on the number of nodes in the subnet.

II. Double-Circuit Structure System

In order to determine the first dependence we have to choose the algorithm that will find the shortest path from the one node to another, which takes into account the weight of the edge (the rate of data transmission via the channel) and the rate of data processing at the nodes, this is Dijkstra's modified algorithm. Due to the fact that the packet delivery time is influenced not only by the date transmission time via the channels, but also by the data processing time in the nodes, we will assume that the data processing time in the nodes can be neglected.

Due to the fact that Dijkstra's algorithm allow to find the shortest path with the minimum number of edge weights, but in this case it is required to find the path along which packets are transmitted in the fastest way from the one node to another, it is necessary to modify this algorithm and to calculate not the shortest path according to the minimum the weight of the edges, but the longest path in terms of the maximum weight of the edges, i.e. to find the minimum data transfer time, depending on the baud rate.

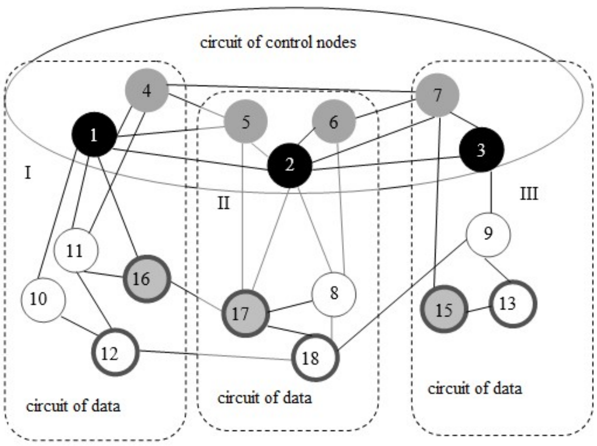

Figure 1. Double-circuit structure for the requests processing in a distributed data processing system

Let us define the first dependence, that is, the time of packet delivery to the corresponding control node of the first circuit of the distributed data processing system. It will depend on two factors: the time T_{pd} that is spent for the message to be delivered to the corresponding node of the control circuit and the time T_{pp} that is spent on the control node processes the packet with the request. The first time is determined using the modified Dijkstra's algorithm as:

$$T_{pd} = \sum_{i=1}^{m} \frac{S}{v\left(a_{i-1}, a_i\right)}, \qquad (1)$$

Where S is the packet size, $v(\mathrm{i}, \mathrm{j})$ is the rate of packet transmission via the communication channel between the nodes i and j of the distributed system, $a_1, a_1, \ldots, a_m$ is the set of numbers of the nodes of the distributed system via which the packet passes, m is the number of nodes via which the packet passes with the minimum data processing time.

The second time T_{pp} is determined only by the performance of the node, so we will consider it as constant. Hence, the time that is spent for the packet to transfer from the initial node to the corresponding control node is:

$$T_{com} = T_{pd} + T_{pp}. \qquad (2)$$

and after substitution T_{pd} :

$$T_{com}(m) = \sum_{i=1}^{m} \frac{S}{v(a_{i-1}, a_i)} + T_{pp}. \qquad (3)$$

It can be seen from the formula (3) that the more nodes the packet passes via the corresponding control node, the longer will be the delivery and processing time of the packet to the control node.

Let us assume that the number of channels that need to be traversed to a given control node is equal to half of the complete graph:

$$m = \frac{M*(M-1)}{4}. \qquad (4)$$

Then, respectively, substituting the formula into T_{com} , we will get the dependence on the number of control nodes:

$$T_{com}(M) = \sum_{i=1}^{\frac{M*(M-1)}{4}} \frac{S}{v(a_{i-1}, a_i)} + T_{pp}. \qquad (5)$$

Let's define the second dependence, that is, the delivery time of the package from the first circuit (the circuit of the control node) to the database node as:

$$T_{proc}(k) = \sum_{i=1}^{k} \frac{S}{V_i} + \frac{S}{V_q}, \qquad (6)$$

where S is the packet size, V_i is the transmission rate of the i communication channel in the data circuit of the distributed system, V_q is the rate of packet processing in the node of the heterogeneous database of the distributed system, and k is the number of nodes on the way from the control node to the node with the database.

The number k of nodes depends on the number of control nodes M , and the total number of nodes N , namely are:

$$k = \frac{N-M}{M}. \qquad (7)$$

This formula allows to determine the average number of nodes on the way from the control node to the node with the database, so we substitute it into the formula T_{proc} , namely:

$$T_{proc}(M) = \sum_{i=1}^{\frac{N-M}{M}} \frac{S}{V_i} + \frac{S}{V_q}. \qquad (8)$$

For the general dependence, we take the sum of processing times depending on the number of control nodes:

$$T_{all}(M) = T_{com}(M) + T_{proc}(M). \qquad (9)$$

The proposed approach allows to define the optimal number of control nodes depending on the processing time of packets and the total number of nodes.

REFERENCES

[1] Buhyl B., Huskov P., Lavriv O., Bak R., Luntovskyy A. Maximization of Service Flows Rates as a Solution of Network Capacity Allocation Problem. Internet Things Eng. 2018. Appl., vol. 3, no. 1. P. 110 – 121.

[2] Heidarian Z., Movahedinia N., Moghim N., Mahdinia P. Intrusion Detection Based on Normal Traffic Specifications. International Journal of Computer Network and Information Security. 2015. Vol.7. No.9. P.32 – 38. DOI: 10.5815/ijcnis.2015.09.04.

[3] Her Q., Malenfant J., Zhang Z., Vilk Y., Young J., Tabano D., Hamilton J., Johnson R., Raebel M., Boudreau D., Toh S. Distributed Regression Analysis Application in Large Distributed Data Networks: Analysis of Precision and Operational Performance JMIR Med Inform. 2020. V. 8(6):

[4] Huang Y., Bessis N., Norrington P., Kuonen P., Hirsbrunner B. Exploring decentralized dynamic scheduling for grids and clouds using the community-aware scheduling algorithm. Future Generation Computer Systems. 2013. Vol. 29. Issue 1. P.402 – 415.

[5] Popov S. G., Fridman V. S. Review of methods for dynamic distribution of data in distributed database management systems. SPbSPU Journal. Computer Science. Telecommunication and Control Systems, 2018. V. 11, Issue 4. P. 82 – 107.

[6] Zheng K., Gu D., Fang F. et al. Data storage optimization strategy in distributed column-oriented database by considering spatial adjacency. Cluster Comput 20, 2017. P. 2833 – 2844.