Method for detecting data synchronization errors in distributed information systems

Belov A.V., Slastnikov S.A.

National Research University Higher School of Economics

The article offers a generalized mathematical model of a distributed information systems (DIS). The main objects of the constructed model are:
1. Oriented connected graph ***Γ (N, G, B***) of an arbitrary topology with no cycles.

2. The set of nodes ***N (I, V, C***), each of which is characterized by the set of ***I*** - information objects that are used in this node, the set of ***V*** - version numbers of these objects in the node and ***C*** - the events of the change of the version numbers of each object.

3. The set of "green" arcs ***G (i, L, S, λ)*** that correspond to the channels for transfer of object changes, where ***i*** is the object whose changes are transmitted, ***L*** is the channel latency (i.e., the time required to transmit the change block minus the time required to transmit each individual change), ***S -*** the time of transmission of one change, ***λ*** is a random variable that determines the time that elapses since the first new version number of the object i appeared after the previous sending, until all the new versions begin to be transmitted along the green arc. That is, this random variable determines the time during which new versions that are included in one change block accumulate.

4. The set of “blue” arcs ***B (I, S, λ)*** that correspond to the channels of current information transmission, where ***I*** is the set of objects participating in the data transfer, ***S*** is the data transmission time, ***λ*** is the random event flow of the new transmitted data packet .

Subject to random event streams and transmission delays assigned to green arcs, new version numbers of objects appear in the model nodes and are copied to all nodes associated with the node of the original

appearance of the new version over time. At the same time on the blue arcs there are forwarding of the current data, at times, determined by the streams of random events. All the random event streams in the model and the random variables are considered independent of each other.
Based on this model, the problem was formulated to detect situations when a data package sent via the blue channel arrives at the receiving node earlier than the receiving node on the green arcs will deliver all versions of the object that were in the sending node at the time of sending. As a result, the message will not be correctly processed by the receiving party. This situation was called “The situation of late update”.

The following is an analytical solution to the problem of calculating the probability of a delayed update situation, in the event that object changes are transmitted immediately after they occur. Let the flow of occurrence of changes and sending of messages be the Poisson point process, and all nodes that are connected by blue arcs are also connected to the blue green arcs through which all objects that participate in the generation of current data transmitted through the blue arcs are updated. The graph of ***G*** is not It contains cycles of green arcs that update the same objects.

Under these conditions, the probability of the DIS running smoothly during the lifetime of T_0 is calculated by formula (1)

(1)
Where C is the information object, i is the index of the graph node, is the total flux density of the update events of the object C in node i, taking into account the updates coming from other nodes, j are the indices of the blue arcs originating from the node i nodes, - the density of the flow of events of the transmission of current data along the j-th arc. - the so-called vulnerability interval after updating the object C in node i during which data can be sent to node j that will cause the situation of a delayed update; finally, this number is known Greater than the number of updates to object C Data on the j-th arc - the so-called “vulnerability interval” after updating the object C in the node i during which data can be sent to the node j that will cause the situation of the late update. Finally, this

number is certainly greater than the number of updates of the object C, which can occur at node i. that is often known to the DIS designer at the designing phase of the system.

Then an analytical solution of the problem is carried out for the case when the delay between the occurrence of a new change and the sending of a block of changes along the green arc is determined by a random variable with exponential distribution while all other conditions are preserved. The solution obtained is substantially more complicated than the above, and its practical application will require approximating the exact solution by numerical methods. Hence, it is concluded that the simulation method can be more productive in solving the problem of calculating the probability of the uninterrupted operation of a distributed information systems in the general case, rather than searching for a general analytical solution.

Computer simulation was conducted in Matlab and confirmed the adequacy and efficiency of the proposed method.