ISSN 2079-3316 PROGRAM SYSTEMS: THEORY AND APPLICATIONS vol. **13**, No 4(55), pp. 27–37

Research Article HARDWARE AND SOFTWARE FOR SUPERCOMPUTERS

UDC 004.272.45

10.25209/2079-3316-2022-13-4-27-37;

Synchronous execution of group operations in distributed supercomputer components and computer clusters

Gennady Georgievich **Stetsyura**

ICS V. A. Trapeznikov of RAS

(*learn more about the author on p. 37*)

Abstract. This paper proposes decentralized processes for synchronizing the actions of a distributed group of active components (objects) in supercomputers and computer clusters, allowing them to move to specified states or influence the external environment synchronously. The object action depends on the current state of the object and the external environment. The actions should start with the minimum delay after the possibility of their execution is detected. Synchronization is performed by exchanging optical signals over wireless communication channels through an optical signal repeater, combining one group of objects or sequences of groups of objects (layers). Accurate distance measurement performs the compensation of possible changes in distances between objects. Group operations accelerate synchronize and simultaneously receive data from a group of distributed objects. Data processing occurs during their transfer, without increasing the time. The operation time does not depend on the quantity of data processed by the operation. A group operation is performed in a repeater containing no computational means.

(*Аннотация по-русски на с.* **??**)

Key words and phrases: supercomputers, group operations, decentralized control, multilayer synchronization of object actions, distributed in-network computing

2020 *Mathematics Subject Classification*: 65Y05; 68Q10

For citation: Stetsyura G. G. *Synchronous execution of group operations in distributed supercomputer components and computer clusters* // Program Systems: Theory and Applications, 2022, **13**:4(55), pp. 27–37. http://psta.psiras.ru/read/psta2022_4_27-37.pdf

Introduction

The article deals with the following problem. There is a group of distributed components (objects) of a supercomputer (SC) or computing clusters. Interaction between objects is performed via wireless optical communication channels. Objects of the group unexpectedly detect the occurrence of an event which requires their joint transition to a new state as quickly as possible. Objects must inform the whole group as quickly as possible about the occurrence of this event and organize the simultaneous transition of the group to the required state.

The requirement for a fast response to an unpredictable event eliminates the often used but slow to implement transition time sign with a clock. Instead, the exchange of signals or messages only makes objects aware of the occurrence of an event that requires their coordinated action.

Transition operations must be performed in a decentralized manner. This is the requirement for obtaining a high transition speed, since the participation of the control center is accompanied by additional time consumption.

Transition comprises many operations. Synchronization of actions of objects, formation of commands by a group of objects, controlling actions of objects, exchange of information on a condition of objects, probably rearrangement of structure of communications between objects and a number of other operations are required. All or part of these operations can be perormed or prepared in advance, before the required moment of state change.

The author did not get the specified results using known methods and the author developed new ways and operations that perform them. Some of them have been published earlier and adapted to the proposed problem, and I propose some of them in this article for the first time. The main ones are group operations and a decentralized method of synchronizing object actions that require organizing the structure of links between objects in order to execute the method given below.

A group operation is an elementary operation, which simultaneously receives and processes data from a group of objects of the supercomputer. Objects perform group operations when they receive one such operation — a group command, which is sent to all objects by one or more of them. The group operation transfers data by objects, without increasing this time.

The processing time does not depend on the amount of data simultaneously processed by the operation.

The group operations in the article are performed mainly in the network's means that connects the objects, in its simple signal repeater, which does not contain any logical elements. The repeater also performs an important function central to the article. It replaces groups of arbitrarily in space sources and receivers of signals by one source and receiver. Only with this replacement did it become possible to perform group operations involving arbitrarily located relative to each other sources and receivers of messages.

This solution is characterized because here the network facilities, which do not contain computers, simultaneously carry out message transmission, control the transmission process and perform distributed computations. The lack of software processing has also increased the speed of group operations.

The paper discusses single-layer and multilayer group operations. In a single-layer operation, all receiver objects form a single group that receives a group command common to it. In a multilayer group operation, objects are divided into ordered layers.

In the first layer, the layer receivers simultaneously execute sources commands. In the second and subsequent layers, receivers of the previous layer become sources of data and commands for receivers of the next layer. In each such pair of layers, simultaneous execution of commands by receivers is ensured. Several multilayer operations can exist in the system simultaneously, composition of objects in layers and operations is formed in dynamics.

It was noted above that the acceleration of responses to emerging events is most important for operation in hard real time. The first complex computer systems, which the press refers to as supercomputers, appeared in the 1970s and were intended to solve exactly this tasks. These were the RADCAP associative systems, their successors, the STARAN system, and the later MPP. STARAN, for example, regulates aircraft traffic at Kennedy Airport in New York City. Now, in most publications, supercomputers are used in tasks that are less demanding in terms of speed of response to events. Recently, however, it has become possible to use compact supercomputers for such tasks. Practicall desktop supercomputers of Nvidia DGX series, with performance exceeding 1 pflops, can serve as an example.

Similar supercomputers should appear for mobile systems, as the complexity of tasks they solve increases simultaneously with the increase in the speed of response to unforeseen events.

This article focused primarily on such systems. In the article, the network facilities perform not only message transmission but also manage the use of the network and perform calculations in it.

But such combinations in the network of many functions have already been. In 1997, an article [?1] proposed the use of network computer facilities for traffic management, then there appeared to suggest the use of such facilities for network configuration management [?2], In-Network Computing direction appeared [3,4,5], in which several types of computational operations are also performed by network computers.

In contrast to these works, in order to increase the system's reactivity, this paper do not use computers in the network. They have been replaced by non-computing repeaters.

The solutions proposed below accurately determine the time interval required for the transmitted signal to travel to the distance between system components. Such solutions require fast and accurate ways to correct for possible changes in distances between system objects. Small changes in distances, such as those resulting from ambient temperature changes, must be considered for fast-acting computer systems. The basis for such measurements is an industrial network protocol PTP correcting the object clock readings [?6, ?7, ?8] and a protocol created in Project White Rabbit for precision physics experiments [?9, ?10]. Currently developed a standard that combines these two methods [?8]. The listed methods use a leading object (leader), which performs the necessary measurements. In PTP, the leader has an accurate clock and step by step synchronizes the clock readings of slave objects, measuring the distance of objects from the leader. In [?9, ?10], the error in time measurements lies in the sub picosecond range. These results are used unchanged in the article. But the leader replaced by arbitrary objects.

All solutions of this article are oriented to distances between SC objects of tens of meters or less. uLimiting the distances is necessary to get high reactivity of the distributed system. The results of the paper are distributed inside as follows. In section 1, the structure of objects performing group operations is discussed. In section 2, methods for synchronizing receivers of group commands are given. In **??**, synchronization of command sources

Figure 1. Simple network with a single repeater

(a) current version (b) earlier version

Figure 2. two repeaters containing a retroreflector

is described. In **??**, peculiarities of group interaction of objects are considered and a brief description of group commands not included in the article is given. Sections describe a complete set of tools necessary for synchronization of objects' actions.

1. Means for group operations perform

1.1. Structure of connections between objects

The means of performing multilayer group operations in SC is developing the structure of performing single-layer group operations [?11] proposed by the author earlier. Its main fragment is shown in **??**.

The figure shows objects of two types: objects Oi objects – (group So) and receivers of signals (group Re) and object RS – repeater of optical signals of Oi. To receive signals from RS objects Re, send continuous optical signals of frequencies $^*!f_1$ and $^*!f_0$. A source So sends to an RS signals f_1 and f_0, carrying messages. An RS modulates with message signals the continuous signals returned to the receiver. Objects can contain a switch to select the necessary RS.

As shown in Figure $2a$, the RS object is an optical retroreflector (Rf), which returns the signal coming onto the RS to its source. The RS contains photodetectors (Pd) of f_1 and f_0 signals and filters-modulators (Md) of $^*f_1$ and $^*f_0$ signals. The filters are typically closed, and the $^*!f_1$ and $^*!f_0$ signals entering the RS do not return to their sources. When a signal f_1 or f_0 arrives at the photodetectors Pd, the frequency-filter $^*!f_1$ or $^*!f_0$ opens, respectively. All receivers sending $^*!f_1$ and $^*!f_0$ signals will record the arrival of f_1 and f_0 signals in the RS from the source objects. These actions determine the scheme of message exchange signals Ms between objects O_i. This messaging structure elaborates the method developed earlier in [?12] to receive data from the source at the expense of the energy of the data-requesting device. It is shown in Figure $2b$

FIGURE 3.

FIGURE 4. Network with Repeater Group

The receiver Re data sends a continuous optical signal to the data source retroreflector So. The latter returns the receiver their signal, modulating it with the contents of the requested data. Here there is no separate RS object. The source So has a modulator Md of incoming polling signals from Re. There are no photodetectors. The message signals (Ms) coming into So form the Re return signals in Md. This method was developed for marine applications. It is also proposed to use it for communication between satellites [?13]. For communication systems, this is enough, but the SC requires additions Figure 2a and switching of links between many RS.

A particular synchronization of message exchange, discussed in other sections of this article, will also be required. The switching system using the node of **??** is shown in **??** Consider the switching process in the single-layer structure of **??**.

Here there are groups of objects O_i and RS_k. The value of k can start from 1 and exceed the number n of O_i objects. Each O_i must have the means to send to any, defined by the current requirements of the task being performed, object RS_k signals f_1, f_0, $*!f_1$, $*!f_0$, and receive from any RS_k signals $*!f_1$, $*!f_0$, modulated by signals f_1, f_0 of other O_i.

Since only RS_k is the only intermediary between message source and receiver, which does not introduce delay into the transmission process, switching is straightforward. The source and receivers of the message are connected in the above way to the RS_k known to them in advance, and a general message transmission to the entire group of receivers is performed. At the same time, using other RS_k will perform the transfer of different sources. In **??**, one of the RS_k group repeaters (repeater number 4) will be needed as $^*RS_k$ to communicate between the groups of objects in **??**. In order to transmit the message simultaneously to all O_i, it is reasonable to introduce a broadcast RS, which sends the message received by it to all objects via a non-directional optical or radio channel.

The organization of the multilayer operation is illustrated in **??**. Here theO_i objects are divided into groups assigned to different layers. Each such group is assigned a group RS_k. Within this group, the repeater of the

${}^*RS_k$ layer is distinguished. It differs from the rest RS_k layer only in the way it is used in layer interaction.

The object ${}^*RS_k$ is accessed simultaneously by O_i of two layers - the current and the preceding. As indicated in the introduction, the O_i objects of the preceding layer work with ${}^*RS_k$ as sources and the current layer'sO_i as receivers. Switching links between objects within a layer needs to be done more often and faster thanO_i links with ${}^*RS_k$. Therefore, these types of switching can be performed by technical means that differ significantly in speed.

Note to **??**. The linkage structure of **??**, when the SC is included in a control system with mobile objects, prohibits the latter from leaving the field of signal reception by the retroreflector. This limitation is eliminated in the article [?14]. Let us note the basic capabilities of **??** structures.

(*1*) A group of objects may use a separate channel for each paired connection (switch mode) or a common channel for connecting sources to a single receiver (bus mode). Both types of structures may exist simultaneously. The structure of links is changed with high speed by sending a joint command to objects without reducing data processing speed.

(*2*) Between sources and receivers, there is a single intermediary RS, which does not reduce the speed of interaction between objects, and allows accurate synchronization of transmitted data when they come to the RS. Alignment of arrivals to RS of the same name bits of messages of a group of sources serves as a basis for all group control, and computational operations applied further in the paper.

(*3*) The structure allows replacing the failed communication channel quickly. It is enough to have one redundant RS to replace any failed RS. In addition, as a rule, channels have a reserve of bandwidth, and the load of the failed channel may be redistributed between them.

1.2. Means of measuring distances between objects

Here we summarize the basic principles of measuring the time intervals required for an optical signal to cover the distance between two devices in the PTP method given in the introduction. Two objects, master and slave, interact, as shown in Figure 5*a*. Master sends to slave in moment $t1$ the signal of the beginning of synchronization and its clock. This information at the moment $t2$ comes to slave. The slave corrects their

(a) RS_k (b) O_j

Figure 5. Time measurement in PTP and a variant of its use in the article

clock on the master clock. At the moment $t3$, the slave sends to master a response signal and his clock. Master at moment $t4$ receives this reply, determines the time of signal transfer between master and slave $= (t4 - t3)/2$ and reports the value to slave the corrected clock reading. The slave corrects the clock. Without several details, this is the basis of time correction in PTP.

The version of measurement used in the article without using the clock is shown in Figure 5b. The signal transfer time between the object and the RS signal repeater is measured. For optical signals, a passive retroreflector can be used as RS. An arbitrary objectO_i has a timer, turns it on, and at time $t1$ sends a signal to RS. The signal arrives at RS at time $t2$. With the delay of RS at time $t3$, the signal will be sent to RS, and at time $t4$ objectO_i determines the signal transfer time betweenO_i and RS as $T_{OiRS} = ((t4 - t1) - (t3\breve{\ }t2))/2$. The value of T_{OiRS} is small and easily controlled.

WR has achieved a more accurate way of measuring time intervals. Simple means achieved accuracy in the femtosecond range [?10]. WR uses a digital version of the beat method.

In the article, this method is most helpful in synchronizing objects performing high-speed cooperative distributed computations. Objects measure distances one by one in PTP and WR, so the repeater RS should not change the input signals.

2. Synchronization of actions of command receivers

This section describes the synchronization of distributed objects of the SC to perform required actions simultaneously or with additional time delays that can be set for each object. In subsection 2.1, a single-layer synchronization of command receivers sent by a single source is considered. In **??**, we present multilayer synchronization with a single command source.

2.1. Single-layer synchronization of command receivers

Let there be a group of command sources of which only one, arbitrarily located O_i, sends a command to object receivers via RS_k. Let the command receivers know their distance from RS_k and the signal transfer time between these objects. In order for the group to act simultaneously, each member of the group, upon receiving the command, must delay their action in order to equalize their distance from RS_k compared to the distances of other objects.

To compensate, object O_i performs a delay $O_i = T - T_i$ after receiving a command. Here T is the time interval, not less than the signal transfer time between RS_k and the farthest from RS_k receiver, and T_i is time interval of signal transfer between RS_k and O_i. It is easy to check; with this O_i, the actions of the receivers' commands will be executed simultaneously. If we add delay ai to O_i for O_i, then ai will shift the action of O_i relative to other receivers. This action completes the one-layer synchronization.

The execution of synchronous actions becomes considerably more complicated when there are changes in the dynamics of T_i times. There are two main reasons for such changes changes in the distances between RS_k and O_i due to external influences, e.g., temperature changes. If the duration of signals transmitted between RS_k andO_i exceeds T, the changes can be neglected. Nevertheless, in the SC, when working with signals of picosecond duration, changes in Ti must be taken into account. Let us outline the necessary actions for this.

Since it will be necessary to have several independently acting groups of objects simultaneously, let us require that the correction of Ti changes does not require a particular central device (leader). AllO_i groups will initiate the correction by performing the following actions.

The receiver sends RS_k a synchro signal S of duration not less than T if it has not detected before that moment the presence in RS_k of signal S sent by other sources. The superposition of signals S creates a joint signal S of variable duration. The moment of its termination $^*S$ serves as a synchronization start signal. Let the receivers have ordinal numbers. The receiver with the lowest number measures the distance from RS_k. Then the other receivers measure it one by one. The next receiver starts the measurement after the moment of measurement completion by the previous receiver, known to the objects.

When all the objects in the group have been completed, the measurement process is repeated. It is reasonable to perform measurements following the solutions [?6, ?7, ?8, ?9, ?10] given in the introduction. These solutions are based on the measurement of signal transfer time between two objects, which is used to synchronize the objects' clocks. From these solutions, only measuring the mutual distance of objects is taken.

After performing the time measurements, the synchronization is completed as described at the beginning of **??**. Supplement. Let us use signals with frequencies different from those of the message signals to measure distances. Also, assume that the modulator in the RS passes the measuring signals in any of its states. In this case, measurements are accelerated because all objects can measure distances simultaneously. This result follows the retroreflector property to return the signal passed through the directional channel to its source.

In **??**, a time interval T is allocated to each object to measure the distance from RS_k. A more complex measurement method, which allows performing measurements for all receivers in time T, described in the paper [?15], is possible.

Object in this section corrects the distance from RS only after all objects have completed the distance measurement. The new method for eliminating the accumulated error uses the channel allocated only for it. The action of the method is as follows. Let the objects first perform the above described alternate measurement of distances to the RS. After that, the objects send a signal S. Receipt by an object of the signal $^*S$ serves as the beginning of accelerated synchronization. On receiving a signal $^*S$ with a delay O_i, each object O_i transmits the scale to the RS.

The scale is a sequence of binary positions. Each bit of the scale is assigned to one of the O_i participating in the synchronization. The object O_i places in its scale position with the order number i the signal St. This signal has a shorter duration than the allocated for the transmission of the scale position. It is placed in the center of the discharge. The objects scales with the alignment of the same-numbered discharges go to RS and are returned to the objects. When the distances between O_i and RS change, the St signal shifts within the discharge, allowing all objects to account for the change in distance from RS during the scale transfer time.

Comment. Let signals with frequencies different from the frequencies of message signals be used to measure distances. Let also an additional retroreflector without a modulator be used in RS (or the modulator always passes the measurement signals). In this case, measurements are accelerated, since all objects can measure distances at the same time. This result follows from the property of a retroreflector to return a signal that has passed through a directional channel to its source. An acceleration has been obtained, without which each object will need to allocate a time interval T to measure the distance to RS_k.

Received	01.07.2022;
approved after reviewing	18.09.2022;
accepted for publication	19.09.2022.

Recommended by *д.ф.-м.н. С. М. Абрамов*

Information about the author:

Gennady Georgievich Stetsyura

About the author

iD 0000-0003-4606-4424

e-mail: gstetsura@mail.ru

The author declare no conflicts of interests.