The algorithms for parallel information processing in many-stage commutation systems for high performance computing systems and communication systems

I.O. Barabanov, E.A. Barabanova, N.S. Maltseva, O.V. Kudryvtseva, Yu.A. Lezhnina

Astrakhan State Technical University, street Tatishcheva 16, Astrakhan, Russia

igorussia@list.ru, elizavetaalexb@yandex.ru, maltsevans@mail.ru, kudryavtzevaov@mail.ru, lejninau@mail.ru

Abstract. *The problem of increasing the throughput of data transmission networks and the output of multiprocessor systems, consisting of many-stage commutation systems, intended for a large number of the inputs, is solved. Two parallel processing algorithms are offered. One of them is parallel search and the other is parallel identification. Traffic detentions are much less in the many-stage system with parallel processing than in the many-stage systems, using the consistent search of communication channels.*

Keywords

an information processing, the communication system, a parallel searching, a parallel identification, the algorithm, a capacity, a simulation

1 Introduction

Traditionally the producers of the commutation equipment use in the function of the commutating sphere such topologies as the bus-line, the shared memory and the matrix switch. Clusters and networks require high-performance and reliable systems that the topology of the bus-line doesn't provide. Shared-memory systems are historically the first type of commutation systems (CS). Their main disadvantage is size limitation. The largest systems can connect about 2000 processors. The main advantage of the matrix switch is the simplicity of the construction, and the disadvantage is the limitation to the number of the inputs, because the complexity of this system increases sharply during the increase of the number of the inputs.[1] One of the popular commutation solutions for the high-performance computing systems is InfiniBand. But such system is complex in design. [2]

The way-out of this situation is the use of three or more cascade schemes, which allow to obtain the increase of the system throughput at less cost. That is why the one-stage structures are used in case of small number of the inputs, the many-stage structures are used in case of great number of the inputs. This rule is applied for the commutation systems with the parallel adjustment [3], which are built on the principle of the iteration build-up of the cascades.

The commutation systems with the parallel adjustment are intended for the use in high-performance computing systems and high-speed networks of data transmitting, where the systems with a small number of the inputs as well as the multi-port switchers are needed. Thereby, the elaboration of the algorithms of the parallel data processing in many-stage CS is essential and urgent.

The subject of the research is the schemes and the algorithms of the work of the many-stage CS with the parallel adjustment. The purpose of the research is to elaborate the universal algorithm of the work of the many-stage CS with the parallel adjustment for the high-performance computing systems and the communication systems.

2 The algorithm of the parallel identification of the data channels

The many-stage CS allow to use a set of the crosspoints for the formation of several junction paths through the circuit diagram. The block diagram of the many-stage CS with the parallel identification of communication channels is shown in Fig. 1. The many-stage CS with the parallel identification of communication channels has an odd number of the stages: input, intermediate and output. Inputs and outputs of the CS are divided into the subgroups of three types. The first subgroup contains p inputs and p outputs, the second subgroup contains x inputs and x outputs and the third subgroup contains x inputs and p outputs. The inputs of each subgroup are served by the separate orthogonal circuit diagram – the bank of swithes (they are indicated as *1 ... X, 1 ... R, 1 ... X* in fig.1). The banks of switches of all stages have commutation spots. It is necessary to adhere Pole's criterion to provide full availability and non-blocking of the many-stage switch network with the parallel identification of the data chanels. The main point of this criterion is that the number of the outputs of each bank of switches of the input stage, and, hence, the number of the bank of switches in an intermediate stage must be not less than *R = 2p-1*, where p is the number of the inputs in the bank of switches of the input stage.

For example, if the CS has 2048 inputs and 2048 outputs, and must consist of five stages, the input and the output stages of the system will have 128 blocks with dimensions of 16x64 and 64x16 correspondingly. The second and the fourth stages will consist of 64 banks of switches with dimensions of 128x128. The intermediate stage of the system consists of 256 banks of switches with dimensions of 128x128.

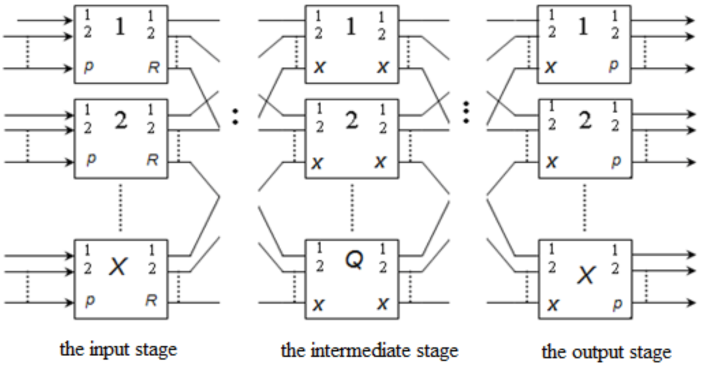

Fig. 1. The block diagram of the many-stage commutation system with the parallel identification of the vacant data channels

The controllers are connected to the outputs of CS (c), they are all connected by the basic bus-line, designed for the exchange of information, and are connected to the control device, uniform for the whole system (the control device) (Fig. 2).

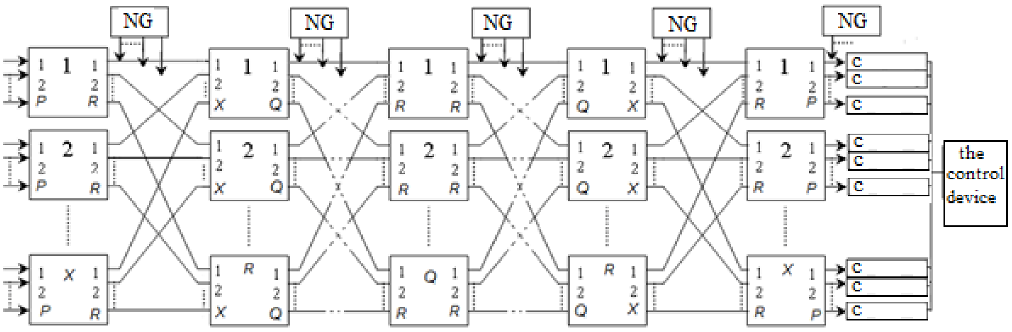

Fig. 2. The block diagram of the many-stage commutation system with the parallel identification of the vacant data channels

The control device provides functioning of the switch network with the parallel identification of the data channels. The automated device with the unconditional logic can be chosen as a control device.

The commutation system refers to the systems with the separate poles, although it has the both-way communications links. It is connected with the peculiarities of system functioning in the process of establishing the connections and the information transfer. In the process of identifying the open ways and establishing the tuning information can circulate in the direction from the inputs to the outputs and vice versa (but into each stroke only in one direction- it is determined by the algorithm of the system work. [4] After the establishment of the necessary connections in the CS, the information can move only in one direction - from the information inputs of the system to its outputs.

Name Generator (NG) is a device that allows at some time to supply to the appropriate input of the commutation spot the name of the output line of the bank of switches, to which this commutation spot is connected .

During the realization of the method of the parallel identification of the vacant data channels the search for the parallel identifiers of the vacant data channels takes place in the external devices in regard to the switch network. The avr-microcontrollers can be chosen as such devices.

3 The algorithm of the parallel search

The many-stage CS with the parallel search are built on the basis of the three-stage CS. The three-stage CS (fig. 4) contains z banks of switches 1.1,1.2,...,1.Z, forming the output stage of y banks of switches, y banks of switches 2.1,2.2, ..., 2.Y, forming the intermediate stage, x banks of switches 3.1, 3.2,.,., 3.X, forming the input stage, nxs information inputs of the system (U.x.n), mxz information outputs of the system (V.z.m), communication lines (C.x.y.) between the blocks 3.X and 2.Y of input and intermediate stages, connecting the outputs of these blocks 3.X of the input stage with the inputs of these blocks 2.Y of the intermediate stage, and the communication lines (D.y.z) between the blocks 2.Y and 1.Z of the intermediate and the output stages, connecting the corresponding data outputs of these blocks 2.Y to the inputs of these blocks 1.Z. [5]

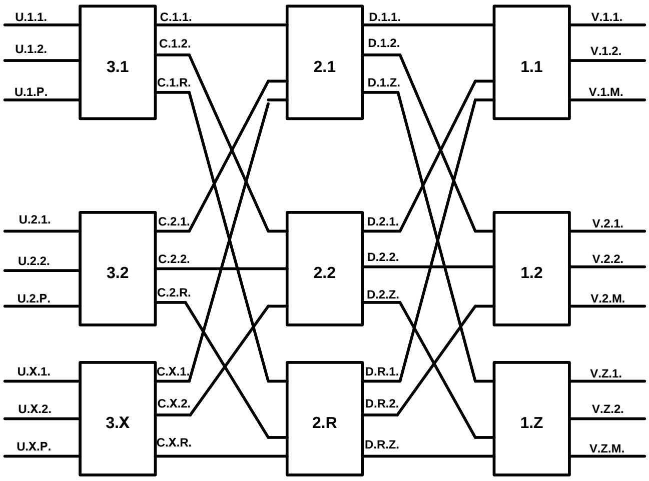

Fig.4. Three-stage CS with the parallel search of the data channels

Figure 5 shows a block diagram of a five-stage CS with a parallel search, it is built of the banks of switches of two types: 32 × 64 for the input stage , 64 × 32 – for the output stage and 64 × 64 for the intermediate stage.[4] The algorithm of CS work is a calculation of the numbers of the inputs into the banks of switches of all the stages of CS and consists of several steps:

1) The array of data, containing the information about the CS state, is formed. In this case the five-stage CS is conventionally considered to be three-stage. The array does not contain the spots with the same values of the parameter b and the different values of the parameter p;

2) Upon receipt of an application for a new connection the array is inspected to the presence of the elements with the same value of b in the same row. If this element is found, it is recopied to the lower row and it happens so until there is a row that doesn’t contains such an element. This eliminates the possibility of blocking in the CS.

3) The same actions are carried out for the central stage of the CS. The array of data for three intermediate stages is filled. The number of the banks of switches of the output stage is laid off horizontally (in this case it is Z). The number of the banks of switches of the intermediate stage R is laid off vertically. Input and output data are considered to be the results, obtained at the previous step.

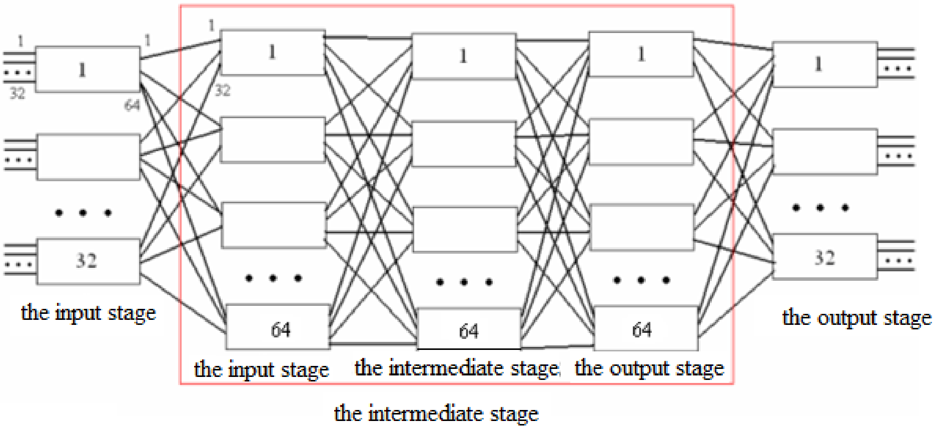

Fig. 5. The block diagram of the five-stage CS

Each adjustment stroke is carried out during two semi-strokes. Searching for the communication channels through the blocks of the intermediate stages to the blocks of the input stage is performed during the first semi-stroke. Searching for the communication channels to the certain inputs in the blocks of the input stage and the formation of the intermediate stage of the connections, branching in the blocks, is performed during the second semi-stroke. The maximum number of the strokes, during which all the considered CS must be tuned, is 64.
The difference of the CS with the parallel search is that the search of the vacant communication channels takes place inside the switch network, and due to this the CS structure is complicated.

4 Simulation modeling

Simulation model is a computer programme, written in Visual basic for Application, which reproduces the algorithm of CS work step by step [6]. The programme of the modeling of the structure and the algorithm of the proposed CS work includes a number of modules. In accordance with its functional purpose the above-mentioned modules can be joined into three basic blocks of the programme:
the block of CS structure;
the block of call flows;
the block of commutation.
The modeling of CS structure is produced in the module Set up model Graf. The initial conditions for CS construction are the number of commutational blocks (CB) of the input cascade x, the number of CB of the output cascade z, the number of the inputs into one CB of the input cascade n, the number of the inputs into one CB of the output cascade m.
As the structure of the processed CS is symmetrical, x=z and n=m. The number of the intermediate cascades y is calculated according to the proportion of y=2m-l, which provides the absence of blocking of the communications links in the intermediate cascade. During the process of the programme execution there appears the need to add the buffer stores to the CS structure. For this purpose the module Set up model Graf has a call to the function Byfer.
The process of the generation of the ingoing flux of calls (the commutation commands) is carried out in the Modeling_Sream module. The uniform law of random distribution is used for getting the random-number sequence.
The modeling of the values of random variable X with the regularly spaced distribution on the section [0,1] is available in Visual Basic for Application with the help of the function Rnd(). The random-number generator, addressing to which takes place in the following cases, is provided in the programme:
during the assignment of the CS engaged outputs in percentage terms from 0 to 100%;
during the assignment of the inputs and the outputs of CS (the commutation programme), the installation of the connection with which is necessary.
It is possible to assign the number of the phases (steps) of the process of the installation the connections.
The initial commutation programme is a variety of the pairs, the first element of which was the output storage location of CS, the second one was the input storage location, with which the given output will be connected.
The continual pairs are absent in the commutation programme, generated for the certain step of the process of the connections establishment, and it excludes the probability of addressing to one and the same CS input.
The modeling of the commutation process is carried out with the help of the commutation block of the programme. The commutation block consists of the following modules: Commutation, Prom_k_Tackt, Byfer, Otrab.
The modeling of the process of CS functioning is divided into the group of m experiments (parts), where the equal number of n experiments is held in each part. It is necessary to choose the number of the testings in each experiment so that the measured statistical characteristics of the analyzed probabilistic observations would be presentable enough.
The modeling allows to obtain the dependences in standard time units (strokes) and the units of measurement of amount of information (spots). When necessary it is always possible to equate 1 model unit of time (1 stroke) to the concise unit of time (N of nanoseconds or M of picoseconds and so on), and 1 model of the unit of measurement of amount of information (1spot) to the consise unit of measurement of amount of information (N bits, M kbits and so on).
The experiments were held with the CS, having 256 inputs and outputs. It means that CS has $x = 16$ CB of the input cascade, n=16 inputs into one CB of the input cascade, y= 31 CB of the intermediate cascade (calculated after Pol's estimation, in accordance with which $y \le 2n\text{-}1 = 31$), z = 16 CB of the output cascade, m=16 inputs from the CB of the output cascade.
The mean time of the process of establishing the connections (adjustment) was first calculated, it was equal to $t_a = 11$ (strokes) for the maximum number of the engaged inputs of 100%. The mean time of the process of establishing the connections is the time interval, during which the commutation commands to the CS outputs from the buffer store are given.
Then the experiments at various values of the engaged outputs in % were held: 10%; 20%; 30%; 40%; 50%; 60%; 70 %; 50%; 90%; 100 %. The different packet length of useful information was defined in each experiment: : $l_p = 15$ spots; l_p =30 spots; l_p = 45 spots; l_p= 60 spots. It is assumed for convenience of holding the next analysis that 1 spot is passed during 1 stroke.

N=6 experiments are held in each experiment. The experimental points are defined for each experiment: numbers of the processed claims, numbers of the submitted claims and the maximum buffer size.

The waiting probability for the calls, held in obeyance (p_w), in each testing of the certain experiment is calculated according to the formula:

$$p_w = \frac{N_\mathrm{sb} - N_\mathrm{pr}}{N_\mathrm{sb}}, \tag{1}$$

where

N_sb is the number of the submitted commutation commands;

N_pr is the number of the processed commutation commands

The mean time of the packet expectation in turn (T_pe) in each testing of the certain experiment is calculated according to the formula:

$$P_\mathrm{pe} = N_\mathrm{bs} \cdot (t_\mathrm{a} + t_\mathrm{t}), \tag{2}$$

where

N_bs is the maximum buffer size (which is defined experimentally);

t_t is the time of transmitting the information, because 1 spot = 1 stroke is taken $t_\mathrm{пер} = l_p$;

t_a is the adjustment time.

We define the experimental point of the throughput of the device (C_cs) according to the formula:

$$C_\mathrm{cs} = \frac{N_\mathrm{p} \cdot l_p}{k \cdot t_\mathrm{a}}, \tag{3}$$

where

N_p is the number of the processed commutation commands;

l_p is the time of transmitting the information (the packet length);

t_a is the time adjustment

k is the number of adjustment phases

It is possible to compare the following characteristics of CS with the help of simulation modeling on the basis of the calculated parameters:throughput and the deference in the input buffers of the devices.

The programme provides the comparison of CS characteristics with the consistent adjustment, CS, working in non-reccurent commutation mode (Banyan's commutators, sorting the schemes), and CS, using the elaborated algorithm of the parallel search, according to which the process of establishing the connections takes place parallel to the packet transfer.

The first diagram reflects the results of CS work, using the algorithm of the parallel search of the communications links, according to which the process of establishing the connections is combined with the transmission of packes at l_p = 65 spots.

The second diagram reflects the results of CS work, using the algorithm of the parallel search of the communications links, according to which the process of establishing the connections is combined with the transmission of packes at l_p = 60 spots.

The third diagram reflects the results of CS work, using the algorithm of the parallel search of the communications links, according to which the process of establishing the connections is combined with the transmission of packes at l_p = 15 spots.

The fourth diagram reflects the results of CS work, using the non-reccurent commutation mode. The fifth diagram reflects the results of CS work, using the principles of the consistent connection. The inaccuracy of modeling in 5% is reflected in the diagram.

Having analyzed the data received, we can draw the following conclusion: the throughput of the elaborated CS with the parallel search of the communications links with the maximum number of the engaged outputs (in %) is 2,5 times larger than the throughput of CS with the consistent method of establishing the connections and 2 times larger than the throughput of CS, using the non-reccurent commutation mode under one and the same initial conditions.

The parallel mode of information processing is indicated with 1, the non-reccurent mode processing is indicated with 2, the consistent mode is indicated with 3 in fig.7.

Analyzing the diagrams of the dependence of the throughput of the elaborated CS from the workload (N) with the different packet length and comparing them with the rest we can draw a conclusion: as the diagram for l_p =15 spots moves closer to the diagram, characterizing the work mode of CS with the non-reccurent commutation, and the diagram for l_p = 65 moves closer to the diagram for l_p = 60, the field of the effective application of the elaborated CS is defined with the following correlation: $11 < l_\mathrm{p} \leq 65$, t_a = 11 strokes.

The inaccuracy of modeling in 5% is reflected in the diagram.

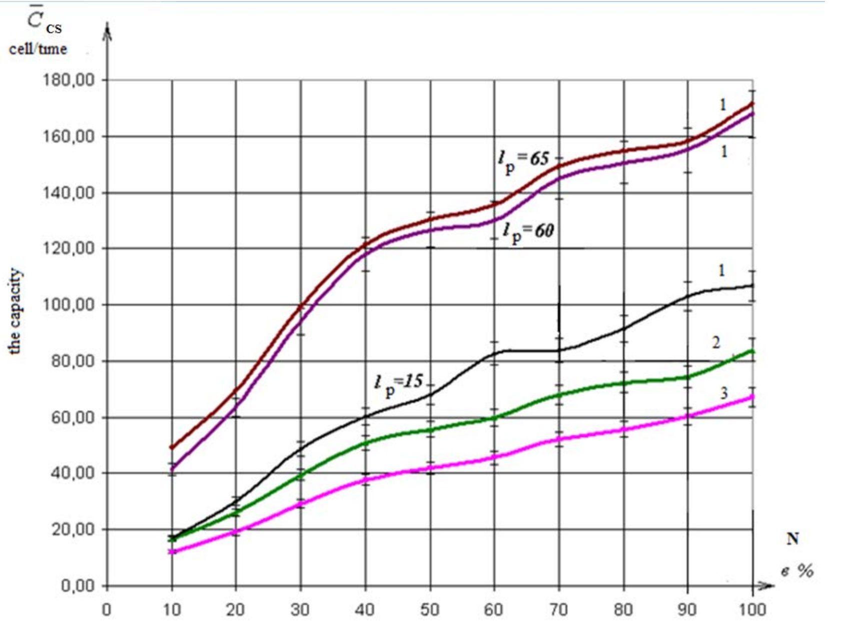

Fig.7. The dependence of the packet waiting probability in turn on the packet length with the number of the engaged outputs 80%.

Therefore, CS with the parallel search of the communications links allows to operate the submitted packets faster, which is a positive characteristic of the system in case of transmitting the traffic, sensitive to the deferences.

5 Conclusion

The algorithms of the search of the free communications links for the many-stage CS with the parallel adjustment, built on the principles of the iteration build-up of the cascades, are elaborated as part of the study.

Such a principle of the build-up allows to increase the number of CS inputs and to decrease the number of the commutation spots simultaneously. Two algorithms were considered: the algorithm of the parallel identification and the algorithm of the parallel search. The CS schemes are offered for each algorithm.

According to the results of the simulation modeling we can draw a conclusion that the use of the parallel algorithms for multistage CS of data processing for multistage CS increases the output of multiprocessor computer systems and the throughput of data-transmission network two times.

References

[1] F. Song, S. Tomov, and J. Dongarra. Enabling and scaling matrix computations on heterogeneous multi-core and multi-gpu systems. In Proceedings of the 26th ACM international conference on Supercomputing, 2012, pages 365–376.

[2] Xiaoshuang Xia, Yi Liu, Yunbin Wang, Tengfei Mu. Infiniband-Based Multi-path Mesh/Torus Interconnection Network for Massively Parallel Systems. FCST '09 Proceedings of the 2009 Fourth International Conference on Frontier of Computer Science and Technology, 2009, pages 52-58.

[3] Kutuzov D, The structure and modeling results of the parallel spatial switching system. Computer Science / Networking and Internet Architecture/ IEEE International Siberian Conference on Control and Communications (SIBCON-2007). Proceedings. Tomsk, April 20-21, 2007. pp. 86-88.

[4] Barabanova E.A. Maltseva N.S. The algorithms of processing of parallel commutation systems // // Vestnik. Astrachan. State. tehn. Univ. Ser.: Management, Computer Engineering and Computer Science. - 2011. № 1.-Astrakhan Astrakhan State Technical University Publishing House. - pp. 150-156.

[5] A patent for an invention. 2359313 Russian Federation, the IPC G06F 7/00, a three-stage switching system / core VV, Barabanova EA, Maltsev NS (RU). - № 2007107780/09; reg.01.03.2007, publ. 20.06.2009 Bull. Number 17.

[6] 2008611841 Certificate of official registration of computer programs FIPS Russia. The simulation program structure and algorithm of switching systems / EA Barabanova N. Maltseva: ASTU holder. - № 2008611147 appl. 20.03.2008, publ. from 14.04.2008.