Е. С. Тихонов

DATA MINING И ПРОБЛЕМА ИСПОЛЬЗОВАНИЯ «ГРЯЗНЫХ ДАННЫХ»

В современной жизни интеллектуальный анализ данных получил широкое признание как мощный и универсальный инструмент анализа данных: не только в информационных технологиях, но и во многих других отраслях, прежде всего в клинической медицине, социологии, физике. Вычислительный процесс анализа больших объемов данных имеет целью извлечение полезной информации. В этой статье будут рассмотрены методы борьбы с грязными данными, которые значительно замедляют поиск ценных данных.

Ключевые слова: получение данных (Data Mining); аналитическая обработка в реальном времени (OLAP); качество данных; пропущенные значения; грязные данные.

Y. Tykhonov

DATA MINING AND USE PROBLEM «DIRTY DATA»

In modern life, Data Mining has been widely recognized as a powerful yet versatile analysis tools in various fields, not only in information technology but also clinical medicine, sociology, physics. Data calculated defined as a computational process of analyzing large amounts of data to extract useful information. This article will discuss methods of dealing with dirty data, because they significantly slow the search for valuable data.

Keywords: Data Mining; analytical processing in real time (OLAP); data quality; missing values; dirty data.

УДК 621.325.5:621.382.049.77

M. KOSOVETS,
L. TOVSTENKO,
Quantor scientific and production enterprise
Institute of cybernetics of V. Glushkov NAS of Ukraine

THE CONCEPT OF CREATION OF THE MODERN CLOUD COMPUTING ON THE BASIS THE DISTRIBUTED MULTIPROCESSOR OF REAL TIME

The solution of the organization of cloud computing on the basis of architecture of the multiprocessor of real time distributed in space is proposed, using perspective modules of processing and telecommunications. The attempt of convergence of system of telecommunications 5G, the Internet of things (IoT) and the multiprocessor distributed in space is made. This combining allows realizing all types of cloud computing, exchange and information representation at the level of perception by the person.

Keywords: cloud computing; multiprocessor of real time; convergence.

INTRODUCTION

Development of computer technique and telecommunications are intensified researches with development of computation on remote computers. Their use led to a concept of the virtual computing environment, and soon generated a successful brand — *cloud computing*. Now the mankind endure a boom on their creation and implementation.

Urgent there is a research of the technique of creation of cloud computing based on the information processing multiprocessor distributed in space in real time with the wireless front-side bus driver of exchange and open, flexible architecture allowing to increase a computing resource for today.

Each type of cloudy services and method of deployment provides the level of monitoring, flexibility and controllability. The «infrastructure as service» model includes in itself basic elements for creation of cloudy IT structure. In this model we get access to network resources, the virtual calculators and databases. The user has all advantages which provide clouds virtually. In case of desire the user can independently create clouds and provide paid access to their resources. It will provide the maximum protection against information leakage, low cost of support of service and complete freedom of creativity. SPC «Quantor» organized cloud computing in Italy with control from Ukraine. It is some kind of outsourcing of cloud.

Creation of cloud requires knowledge of parameters of information flows, redistribution of tasks between specialized coprocessors accelerators and the central multiprocessor system, technical characteristics of the processing modules, features of algorithmic support and program service.

The architecture of the calculator must maintain high parallelism and a tunable configuration that allows to solve problems of mathematical physics, digital filtering, image processing, electrodynamics and others (tasks of processing of multivariate signals). The special part is assigned to the communication environment, allowing to establish flexibly connection between processors.

1. DEVELOPMENT OF ARCHITECTURE OF THE SPATIAL MULTIPROCESSOR OF REAL-TIME SCALE

The analysis of algorithms of the standard tasks using cloud computing specifies their heterogeneity. Thereof it is possible to speak about use of the multiprocessor with architecture of MCMD as basic [1].

According to the classification diagram of Flynn of MCMD system consists of a set of processors which independently execute various commands over different data, to say asynchronous system with decentralized control. At the same time the multiprocessor can be considered or as poorly related system consisting of clusters each of which contains unequal processing nodes or as strongly related system from the identical processing nodes communicating with each other through space of messages.

There are multiprocessor clusters from functionally and technologically uniform elementary processors with a local memory. Besides is a part of each cluster also controlling and communication processors. The first of them is the administrator of a cluster, keeps account of employment, fixes a status of process and distributes jobs. The second — controls media access of communication and provides provision to a cluster of communication services, coding/decoding and the conflict resolution in case of implementation of multiple access. All clusters are connected among themselves to the help of the microwave bus of information exchange realizing high message transmission rate, multiple access to the communication environment. The system is easily reconfigured taking into account specifics of the task. Optimum dimensionality of a cluster is defined from requirements of effective productivity. Results of the researches conducted by SPC «Quantor» allow to claim that productivity of cloud can't be evaluated without task parameters.

Options of possible implementation of the microwave bus are connected to use of methods of orthogonal division: spatial, the frequency, phase, temporal and code or their combinations as owing to mutual orthogonally they can be combined in arbitrary combinations.

One of methods of increase in system performance is creation of accelerators. The accelerators included in the trunk 5G must support the expensive exchange protocol and have the universal potential for the solution of a wide class of tasks. The accelerators included in the trunk IoT shall support standard cheap exchange protocols.

The reasons explained in the present section, lead to the flowchart of the standard cloudy calculator. It has the following technical characteristics: central service module of the base station, modules of the executive processors, processor administrator, system microwave trunk, the communication processor of the high-speed trunk, coprocessors and mobile stations.

We use real-time operating system of the distributed type in time partitioned mode. Each processor of system works independently and performs all functions relating to it: supervisor, executive, program of start, diagnostics, testing, process control of reconfiguration and runtime allocation of tasks, support of parallel computing.

As instrumental system it is applicable simulation modeling. The main feature of systems of preparation of programs includes a warranty of correctness of execution of programs, productivity assessment, the sizes of queues to shared locations, the maximum wait time of requests for service in queue on a system input, the required number of processors. Besides it serves as good addition to their mathematical models.

The program model of structure of system describes set of communications of separate modules of multiprocessor system among themselves, algorithms of interaction of modules, time response characteristics of these interactions. The program model of an operating system repeats the main functions of the modelled multiprocessor system, but at the same time are written in an algorithmic base of work benches.

The main objects of the considered system are models of elements of multiprocessor system: a terminal subsystem, working processors, the controlling processors, the downlink processors.

In case of simulation of a terminal subsystem all instruction set of the microprocessor of mobile addition is reflected [2]. Just with the same level of detailing working managing directors and the downlink processors of the base station are simulated taking into account that they have elements of microwave devices which are subject to simulation. On them communication between the terminal and processing subsystems is carried out. As such characteristics can serve: total time of waiting of shared locations, an operating time of each processor according to application programs, total time of operation according to system programs, the maximum and minimum quantity of the processors occupied with execution of the useful operation, average queue to each shared location, the average time of waiting by the processor of access to a resource, summary time of the decision of the task.

Problem definition of system engineering of the multiprocessor is based on structural analysis of the given algorithm and data. Division of a set of statuses of objects of some class to which belongs the valid status of an object is the cornerstone of structural analysis of algorithms. Formalization of problems of implementation of algorithms is based on representation of an algorithm of operational model by

the oriented count which peaks identify with operations, and arcs — with communications in between. Informative interpretation of operations depends on the algorithm detail level.

The actual standard of means of interpretation of macro pipeline computation is the multitask operating system [3]. The object-oriented architecture of operating systems gives the chance to apply mathematical methods of a research of efficiency of options of hardware-software implementation of parallel algorithms. Creation of an optimal variant of implementers of the given algorithm is provides multiple arbitration of the system trunk and resident resources.

We reduce the basic principles of the structural organization of an operating system to the organization of calculating process with simultaneous functioning in a runtime environment of several processes, using the mechanism of switching of processors or the mechanism of coordination of use of resources the competing processes.

We build an operating system as the interpreter of operators and directives of a basis of macro pipeline computation. The reached effect consists in macro level string management of implementation of components of an algorithm by means of a data stream and events. The specialization of functions of an operating system consisting in orientation to the organization of interaction of components of an algorithm considerably simplifies a choice of strategy of distribution of resources of the calculator.

The spatial multiprocessor oriented on implementation of algorithms comes down to system of the loosely coupled and independent operations based on searching of the tier and parallel form of an algorithm [4]. The found tier and parallel form of an algorithm we describe the oriented graph which peaks are identified with operations, arcs — with communications and the relations in between, and we describe characteristics of components of an algorithm by means of the set of attributes and display setting a binding of attributes to peaks and arcs of a graph. The operational model of an algorithm will have the following appearance:

$$D = < G (Z, B), T, P >, \quad (1)$$

where G (Z, B) — *the tier and parallel form of submission of the graph of an algorithm determined by a set of* $Z = \{Z_j\}$ *of operations and by the relation B, setting information communications on a set Z operations;*

$T = \{T_j\}$ — *the set of attributes of peaks and arcs of graph G (Z, B);*

$P: T \rightarrow > G (Z, B)$ — *the display setting a binding of attributes to peaks and arcs of graph G (Z, B).*

Tasks of a choice of an optimal variant of implementation of the given algorithm have no single decision. If folding of a vectorial index of efficiency in scalar is admissible, then this task comes down to the optimization task of one-criterial type which general setting has the following appearance:

$$R = < D, V, O, H, K, U, u^ >, \quad (2)$$

where D — *the set of the problems connected to implementation of algorithms of the given class;*

V — *set of possible architectural and structural options of a computer;*

O — *a set of the restrictions superimposed on algorithm implementers;*

H — *display of a set of controlled characteristics of implementers of the given algorithm to a set of measure values of efficiency;*

K — *index of efficiency of implementers of the given algorithm;*

U — *a set admissible (satisfying to restrictions for algorithm implementers) architectural and structural options of the computer;*

u^ — *optimal architectural and structural variant of a computer.*

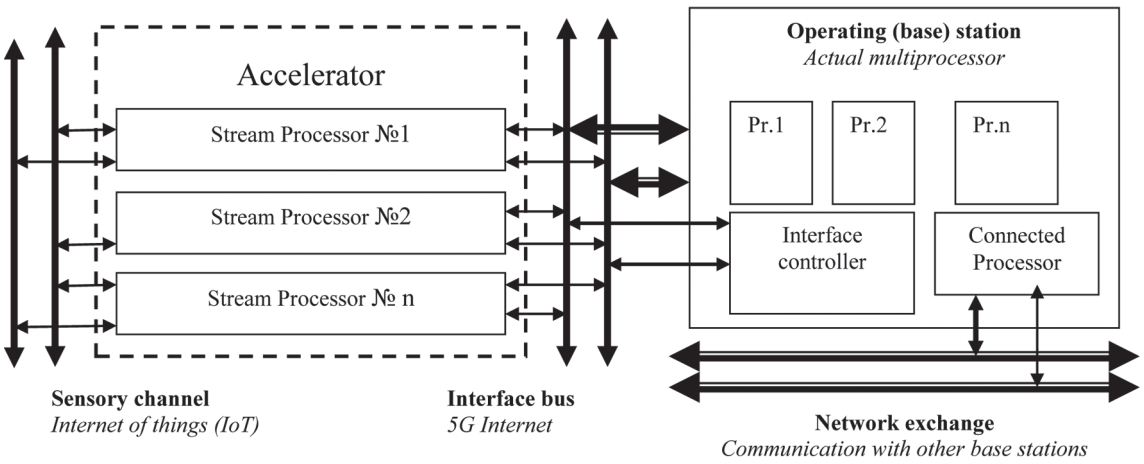
2. FORMATION OF THE CHANNEL OF INFORMATION EXCHANGE IN THE ENVIRONMENT OF CLOUD COMPUTING

Exchange between clusters on the asynchronous bus allows communicating at the same time between two clusters with different speed. The abstract model of the multiprocessor used at the level of architecture has an appearance of a dataflow graph. In the course of synthesis the abstract model undertakes and structural idea of a circuit by determination of necessary resources and parameters of implementation of behavior model is created. The purpose of architectural synthesis consists in formation from the abstract model of an optimum circuit. The model consists of two components: information channel and control.

Actually the multiprocessor is added by the accelerator oriented on signal processing executing operations in time and spatial domain. In this case actually the multiprocessor will represent the driving car performing functions of the central control, planning of resources and superfast conjugation of input-output to matrix processors. In turn the accelerator will effectively perform such operations as digital filtering, correlation and others.

Proceeding from algorithms of the application-oriented task we shall provide architecture of the calculator: actually multiprocessor, system of conjugation, exchange network, accelerator accordingly such figure.

The system of conjugation which is turning on the controller of conjugation and the bus of conjugation realizes functions of swapping of a multivariate array from an accelerator, and also controls processing of the input information arriving from a source of a signal and redistribution of a resource of an accelerator in the presence of a failure of one of them.



The architecture of a fault-tolerant real-time multiprocessor for cloud computing

At the system level principal components of the multiprocessor include actually multiprocessor; controller of conjugation to an accelerator; exchange networks; stream calculators. Algorithms will display structure of the multiprocessor.

The accelerator is interfaced to actually multiprocessor by means of the bus or direct access channel in memory and performs functions of loading, unloading and buffering of data of the stream processor, control of interruptions and data formatting.

At the architectural abstraction layer function of all system is described in terms of algorithms with simulation of behavior of the diagram. Accurate semantics and syntax of model is provided that leads to the coordinated and single-digit submission of the specification. The abstract model at the level of architecture generally begins as a dataflow graph which doesn't contain implementation parameters. In the course of synthesis we take this abstract model and create structural representation of computation by determination of necessary resources and parameters of implementation of behavior model. The purpose of architectural synthesis consists in formation from the abstract model of the optimum calculator.

The set of functionally oriented stream processors in problem of oriented accelerator, in the best way reflects a data structure of a multivariate signal. Each task is carried out on the stream processors intended for it and uses a network for facilitation of pipelining between processes.

Processes are distributed on a network the strongly connected of the processor modules intended for multistream processing. Queue of processes and side-by-side execution of interacting processes is formed. Each processor module saves the status — local variables of process and processor state at different stages of operation [5]. In case of implementation of exchange between processes it is necessary to consider the following points.

1. Exchange of information between processes is carried out only through a communication network.

2. The structure and functions of a network shall be the transparent from controlled processes.

3. It isn't mandatory to processes to provide information on between what processors other processes are distributed.

4. The network will be expanded and meet requirements of processes.

Complete decomposition of system on basic elements gives an opportunity to embody the principle of structured programming in programs of control of separate processes. The entity of decomposition (partition) is in revealing and using the main features of natural structure of system [6]:

1) it is desirable to use so-called natural decomposition;

2) a row of quasi-independent processes with rather feeble interaction is result of decomposition;

3) all correlations of controlled processes are carried out by message passing on a communication network;

4) it is undesirable that all information in system was available to all parts of system;

5) it is necessary to aim at independence of one controlled processes of surge characteristics of other controlled processes, critical on time;

6) it is desirable to appropriate the process to each separate task.

CONCLUSION

In this article the principles of the organization of cloud computing on the basis of architecture of the multiprocessor of real time distributed in space for the decision of heterogeneous tasks of signal processing are considered. Parameters of information flows are determined; redistribution of tasks between specialized coprocessors accelerators of signal processing and processor controls, testing's and diagnosing is made. Requirements to the system trunk which throughput stopped being the restraining factor restricting productivity of computation when using 5G are defined.

The considered organization of cloud computing widely is used in tasks of a location, radiometry, spectrum analysis, machine vision, image processing, processing of seismic and medical signals and many other tasks.

The basic algorithmic kernel of ICCore is realized on FPGA, DSP large scale integrated circuits. The functional devices for Wireless of networks (5G) were synthesized: PLL, BPSK-Modulator/Demodulator, FFT, Decoder Viterbi, LDPC-Decoder and Matched Filter.

LIST OF REFERENCES

1. ***Hockey, R.*** *MIMD computing in the USA – 1984 / R. Hockey // Parallel Computing.— 1985.— No. 2.— P. 119–136.*

2. ***Palagin, A. V.*** *About a choice of the simulating complex by development of computer system / A. V. Palagin, Yu. Yakovlev, K. A. Kirin // USIM.— 2002.— No. 1.*

3. ***Multiprocessor*** *systems and parallel computing. / Under the editorship of F.G.Enslou.— M.: Patterns, 1976.— 383 p.*

4. ***Kosovets, M. A.*** *About hardware of reliability augmentation of fault-tolerant microprocessor systems / M. A. Kosovets // Cybernetics and computing. — 1993.— Release No. 99. Difficult management systems.— P. 102–104.*

5. ***Kosovets, M. A.*** *Development of architecture of the high-performance multiprocessor of real time with the asynchronous bus of exchange for tasks of radiolocation / Radiolocation. Navigation. Communication. X International scientific and technical conference / M. A. Kosovets, L. N. Tovstenko // Voronezh, 2004.— Vol. 1.— P. 627–639.*

6. ***Kosovets, M. A.*** *Features of architecture and structure of a microcomputer with the varied ratio of productivity and reliability / M. A. Kosovets // Special electronics.— 1990.— Series 10. The issue I (26).— P. 23–28.*

Рецензент: доктор техн. наук, професор **А. І. Семенко**, Державний університет телекомунікацій, Київ.

М. А. Косовець, Л. М. Товстенко

КОНЦЕПЦІЯ ПОБУДОВИ СУЧАСНИХ ХМАРНИХ ОБЧИСЛЕНЬ НА ОСНОВІ РОЗПОДІЛЕНОГО МУЛЬТИПРОЦЕСОРА РЕАЛЬНОГО ЧАСУ

Запропоновано вирішення щодо організації хмарних обчислень на основі архітектури розподіленого в просторі мультипроцесора реального часу. Використано перспективні модулі обробки та телекомунікації. Здійснено спробу конвергенції системи телекомунікації 5G, інтернету речей (IoT) і розподіленого в просторі мультипроцесора. Таке об'єднання дозволяє реалізувати всі типи хмарних обчислень, обмін і подання інформації на рівні сприйняття, притаманного людині.

Ключові слова: хмарні обчислення; мультипроцесор реального часу; конвергенція.

Н. А. Косовец, Л. Н. Товстенко

КОНЦЕПЦИЯ ПОСТРОЕНИЯ СОВРЕМЕННЫХ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ НА ОСНОВЕ РАСПРЕДЕЛЕННОГО МУЛЬТИПРОЦЕССОРА РЕАЛЬНОГО ВРЕМЕНИ

Предложено решение по организации облачных вычислений на основе архитектуры распределенного в пространстве мультипроцессора реального времени. Использованы перспективные модули обработки и телекоммуникаций. Осуществлена попытка конвергенции системы телекоммуникаций 5G, интернета вещей (IoT) и распределенного в пространстве мультипроцессора. Указанное объединение позволяет реализовать все типы облачных вычислений, обмен и представление информации на уровне восприятия, свойственного человеку.

Ключевые слова: облачные вычисления; мультипроцессор реального времени; конвергенция.